

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

journal homepage: www.joiv.org/index.php/joiv



Karonese Sentiment Analysis: A New Dataset and Preliminary Result

Ichwanul Muslim Karo Karo^{a,b,*}, Mohd Farhan Md Fudzee^a, Shahreen Kasim^a, Azizul Azhar Ramli^a

^a Faculty of Computer Sciences and Information Technology, Universiti Tun Hussein Onn, Johor, 86400, Malaysia ^b Computer Science, Medan State of University, Medan, Indonesia

*Corresponding author: *ichwanul@unimed.ac.id*

Abstract—Amount social media active users are always increasing and come from various backgrounds. An active user habit in social media is to use their local or national language to express their thoughts, social conditions, socialize, ideas, perspectives, and publish their opinions. Karonese is a non-English language prevalent mostly in North Sumatra, Indonesia, with unique morphology and phonology. Sentiment analysis has been frequently used in the study of local or national languages to obtain an overview of the broader public opinion behind a particular topic. Good quality Karonese resources are needed to provide good Karonese sentiment analysis (KSA). Limitation resources become an obstacle in KSA research. This work provides Karonese Dataset from multi-domain social media. To complete the dataset for sentiment analysis, sentiment label annotated by Karonese transcribers, three kinds of experiments were applied: KSA using machine learning, KSA using machine learning with two variants of feature extraction methods. Machine learning algorithms include Logistic Regression, Naïve Bayes, Support Vector Machine and K-Nearest Neighbor. Feature extraction improves model performance in the range of 0.1 - 7.4 percent. Overall, TF-IDF as feature extraction on machine learning has a better contribution than BoW. The combination of the SVM algorithm with TF-IDF is the combination with the highest performance. The value of accuracy is 58.1 percent, precision is 58.5 percent, recall is 57.2, and F1 score is 57.84 percent.

Keywords- Karonese sentiment analysis; support vector machine; k-nearest neighbor; logistic regression; naïve bayes.

Manuscript received 7 Jan. 2022; revised 15 Mar. 2022; accepted 19 May 2022. Date of publication 31 Aug. 2022. International Journal on Informatics Visualization is licensed under a Creative Commons Attribution-Share Alike 4.0 International License.



I. INTRODUCTION

Social media accommodates various languages from active users in expression. Users can use their local or national language in writing statuses, tweets, comments, ideas, reviews, posts, perspectives, and more. This makes a new user habit on social media and makes them more comfortable on social media. The implication of active users' frequent expressions using the local language is the abundance of local language texts on social media.

Sentiment analysis (SA) is the most common text classification tool that analyses people's condition from a text; it could be emotion, opinion, a hot issue in their circle, or personality [1]. SA is helpful for decision-making since the information text type is abundant by active users and makes it knowledge or wisdom. It is widely used to categorize literature into good, negative, or neutral categories. Recently, SA has been used in a variety of local languages, such as Balinese [2], Sundanese [3], Minangkabau [4], Bambara [5], Marathi [6] and Afaan Oromoo [7] or nationality languages such as Bangla [8], Portuguese [9], Tagalog [10], Thailand [11].

However, SA on non-English text faces under-resourced [12] and framework limitation. In addition, corroborate the opinion [12], firstly, Dataset limitation [4], such as text and label of sentiment (positive, negative, or neutral) and corpus. Secondly, data preparation framework of non-English is not complete to solve the particular problem [13], such as. morphology and phonology [14], library tokenizing, stemming, lemmatization, or Stop-word removal [15]. Based on the previous work, non-English or local languages SA is more challenging than English SA.

Karonese is one of the most widely spoken indigenous languages in North Sumatera, Indonesia [16] It has a unique morphology and phonology [17, 18, 19]. A term on Karonese language potential has multiple pronunciations and spelling with same meaning, the example shown on Table 1.

TABLE I
EXAMPLE OF KARONESE TERM

English	Karonese	
Don't	\ola; \ula; \oula ; aula	
Corn	\jong; \jaung	
There is	\lit; \let	

Surely, that is a challenge and an opportunity to preprocess text to provide good quality resources on Karonese Sentiment Analysis (KSA). Machine learning is a popular method for analyzing opinions of non-English languages. Machine learning has been implemented in another national languages, such as Turkish [10], Arabic [1], Azerbaijani [11], Indonesian [20], and Chinese [21]. Meanwhile, machine learning also has been implemented to analyze local language sentiments, such as Sundanese [22], Bambara [5], Malayalam [24], and Afaan Oromoo [25]. According to previous studies, the machine learning approach is more commonly used for initial sentiment analysis research on local or non-English nationality languages.

This research is an extension of the research [23], with a more varied dataset domain, a number of machine learning algorithms tested, and feature extraction methods. Two kinds of feature extraction are used: bag-of-word (BoW) and Term Frequency–Inverse Document Frequency (TF-IDF). The remaining sections of this paper are organized as follows: Section 2 explains data methodology, whereas Section 3 describes the Machine learning techniques used on SA. Section 4 presents and discusses the results. The last section concludes the paper.

II. MATERIAL AND METHOD

This works was divided into two scopes, Data preparation and SA process (Fig. 1). Data preparation is discussed about crawling dataset, text pre-processing, annotated text, dan spillting dataset. Data preparation is the most important and time-consuming part of sentiment analysis [19]. Data collection is an important part of this work because there has never been Karonese SA in previous studies. In other words, the unavailability of the Karonese dataset for SA becomes the main focus of Karonese Sentiment Analysis (KSA). Meanwhile, the SA process talking about Karonese sentiment analysis using several machine learning algorithms (Logistic regression, Support vector machine, *K*-nearest neighbor and Naïve bayes) to identify positive, negative, and neutral classes.



Fig. 1 Research flowchart

A. Retrieved Data

The initial stage of providing Karonese Dataset is crawling text from social media. There are four famous social media platforms for Karonese people: Facebook, Twitter, YouTube, and Instagram. The Karo people most widely use the social media platform. The crawling process uses the python programming language and application programming interface (API) from every social media platform. The sample text obtained is attached in Table 2.

TABLE II Example of karonese text

Text	Sources
Sedak bas pusuh ku ja nge kuturikenbage pe nindu	Twitter
aku la beluh	
Usuren jungut-jungut asangken ngataken bujur.	Twitter
Banci kin la ayo e lampas ringut.	
ciger wari, las wari sipasna minem paula manis	Youtube
muat sen gundari sel kel serana,	Facebook
Reh kam kujumata kade kade	Facebook
NANDE KARO MELIAS NAKEEE	Facebook
Lalan si Man Ambekenken Asang si	Instagram
Mentanggerenken.	
Aku kalak mesera. Per Juma-Juma	Instagram

B. Text Pre-processing

Text pre-processing has a critical role in SA [19]. Text preprocessing applies natural language processing to transform unstructured data into structured data (NLP). The transformation process adjusts to meet requirements, such as analyzing sentiment, summarizing documents, document clustering, Part of Speech tagging, etc. Concisely, text preprocessing is altering text into index terms, and the objective is to generate a set of index terms that can represent a document. There are many processes on recent NLP [20]. However, several NLP processes do not apply to certain languages, such as library Stop-word removal for English conjunction and particular lemmatize for languages [21, 22]. The NLP processes that can be applied to Karonese include case folding, tokenization, symbolic removal, removing emoji, and removing URLs.

C. Annotated

To complete the Karonese, it can be used to analyze Karonese sentiment; a sentiment label must accompany the Karonese text. Sentiment labels that are commonly used are positive, negative, or neutral. In this work, the labeling process is annotated by four transcribers from Karonese figures, and determining the final label follows the previous research [23].

D. Bag-of-Word

A bag of words is a Natural Language Processing technique of text modeling as feature extraction. This method only counts the frequency of occurrence of words in the entire document [26]. It does not pay attention to word placement or subtle grammatical variances; it simply keeps track of term frequency. Bag of Words (BoW) is one of the simplest and most flexible methods of converting text data into vectors.

E. Term Frequency – Inverse Document Frequency (TF-IDF)

Term Frequency – Inverse Document Frequency (TF-IDF) is a statistical measure that evaluates how essential a term is to a document collection document. TF-IDF is a combination of 2 processes Term Frequency (TF) and Inverse Document Frequency (IDF) [27]. Term Frequency (TF) counts the frequency of occurrence of words in a document. Since the length of each document can be different, generally, the TF value is divided by the length of the document (the total number of words in the document). Inverse Document

Frequency (IDF) is a value to measure how important a term. IDF will assess terms based on how it appears in the entire document. The smaller the IDF value, the less important the word will be, and vice versa. Mathematically, the TF-IDF value for the term t in document d from the document set D is calculated using equation (1).

$$tf idf(t,d,D) = tf(t,d).idf(t,D)$$
(1)

F. Logistic Regression (LR).

LR is a transformation of linear regression using the sigmoid function [28]. The transformation process changes linear regression, originally to predict data classification. Logistic regression creates a model of the relationship between several variables, much like linear regression. When the variable being predicted is a probability on a binary range between 0 and 1, logistic regression is appropriate. In this study, the logistic equation used follows equation (2) and the LR algorithm guide based on other studies [23, 29].

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

G. Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of the effective machine learning algorithms to accommodate multiple variables and classes [30]. SVM is a learning classifier developed to train computers or machines efficiently with training data by applying generalization theory [31]. In addition, SVM attempts to minimize test data misclassification possibility caused by being invisible to the model or being drawn at random from a fixed but unknown probability distribution. The idea of the SVM algorithm is to create a hyperplane capable of separating datasets [30]. This study used a fundamental SVM algorithm like previous work [23, 29].

H. K-Nearest Neighbor (K-NN)

K-Nearest Neighbor (K-NN) classifier is another popular classification method. The algorithm approach uses a similarity function to identify a class of objects [32]. The fundamental idea of the algorithm is to calculate the similarity between K closest objects and group them into the highest similarity class. So, similarity functions significantly contribute to identifying the class of data. The most famous distance applied for data analysis is the Euclidean distance [32]. This study uses the K-NN algorithm guidance as has been provided by previous studies [23, 32].

I. Naïve Bayes (NB)

Naive Bayes classifier is a simple and powerful machine learning algorithm for predictive modeling [22] Naïve Bayes classifier applies Bayes theorem (equation 3) to find the highest probability value to classify test data in the most appropriate category. Naïve Bayes classifier will determine the likelihood that the input data belong to a particular class denoted as A, by examining the values (input data) of a given set of features or parameters, denoted by B in the equation. Naive Bayes assumes that each input variable is independent. It is a strong assumption and makes it become amazingly simple approach that frequently produces highly accurate and stable models with small sample sizes. This study uses the Naive Bayes algorithm guidelines as has been provided by previous studies [22, 20, 23, 29].

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$
 (3)

J. Evaluation

The output of text classification is a classification model. The model is tested against testing data and evaluated using several metrics. The most used evaluation metric is accuracy. However, the accuracy metric has a weakness against unbalancing classes. Thus, the study presents evaluation metrics of precision, recall, and F1 scores. All metrics was calculated from the confusion matrix (Table 3).

TABLE III	
CONFUSION MATRIX	

		Pro	edict
		Positive	Negative
Astual	Positive	True Positive (TP)	False Negative (FN)
Actual	Negative	False Positive (FP)	True Negative (TN)

Precision (P) is the percentage of correctly anticipated positive observations to all the positively predicted observations [32]. It can be calculated using equation (4).

$$P = \frac{TP}{TP + FP} .100\%$$
(4)

Recall is the proportion of correctly predicted positive observations to all available samples [32]. It can be calculated using equation (5).

$$Recall(R) = \frac{TP}{TP+FN} .100\%$$
(5)

F1 score is the weighted average of Precision and Recall as well as a technique for evaluating the model's effectiveness. It can be calculated using equation (6).

$$F1\,score = \frac{2PR}{P+R} \tag{6}$$

The proportion of cases that were correctly classified is called accuracy. It measures the proportion of accurately anticipated observations to all observations. It can be calculated using equation (7).

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} .100\%$$
(7)

III. RESULTS AND DISCUSSION

A. Driven Dataset



Fig. 2 The Phase of Obtaining Karonese Dataset

Karonese Dataset undergoes three phases (Fig. 2). In first stage, this work crawled Karonese text from Facebook, Twitter, Instagram, and YouTube. In second stage, the text is cleaned, corrected typos, case folding, and noise removal. The number of Karo language texts that successfully passed stages 1 and 2 was 1001 texts. As for the distribution of Karonese text from each social media, it is presented in Fig. 3 More than 60 percent of Karonese is obtained from Twitter. So, it makes Karonese tweets dominant than other domain.



Fig. 3 Distribution of text sources

Last stage, Karonese text was annotated by four transcribers into positive, negative, and neutral classes. It means that a Karonese text gets 4 labels from transcriber. The final class decision is determined by the purity method. Final class decision of Karonese text shown on Table 4. The Karonese dataset consists of 305 texts with positive labels, 351 texts with negative labels and the rest with neutral labels. Based on this number, the Karonese dataset does not experience an imbalance class.

TABLE IV
ANNOTATIONS

Source Text	Positive	Negative	Neutral	
Facebook	27	40	43	
Twitter	187	246	234	
YouTube	36	30	38	
Instagram	55	35	30	

Table 4 is the final description of the Karonese Dataset. Furthermore, the dataset is used to analyze the Karo language sentiment. To implemented Karonese Sentimen Analysis (KSA), training and test data were separated from the dataset. The training dataset is used to create the model, while the testing dataset is used to assess the model's performance. In this work, there are several compositions dataset for running algorithm. Table 5 is an experimental scenario based on the results of the split dataset. The goal is to get the best model

	TABLE V Experimental scenario						
Scenarios	Data composition (Training: Testing)						
Ι	80:20						
II	60:40						
III	50:50						
IV	40:60						

B. Experiment I

Experiment I is the experimental baseline of this study. This experiment does not involve the feature extraction process in the data pre-processing stage. In other words, the Karo language text does not go through a feature extraction process. Experiment I used four scenarios by following the guidelines of Table 5. This means that each algorithm produces four classification models. The results of the experiment I, can be seen in Table 6.

TABLE VI Experiment I result (%)

Saanarias	-	NB LR					SVM				K-NN					
Scenarios	Acc	Р	R	F1	Acc	Р	R	F1	Acc	Р	R	F1	Acc	Р	R	F1
Ι	47.8	50.1	49.7	49.9	50.03	49.7	49.9	49.8	50.07	48.15	47.2	47.7	49.92	50.1	51.02	50.56
II	44.3	47.6	42.9	45.13	48.19	50.4	50.9	50.65	50.7	49.8	48.7	49.2	49.23	50.6	51.1	50.85
III	35.5	35.1	32.2	33.6	40.6	45.3	43.03	44.2	43.4	45.7	39.9	42.6	45.1	47.3	49.8	48.52
IV	31.4	41.4	30.8	35.32	30.5	32.2	30.05	31.23	32.5	31.7	35.3	33.4	37.5	36.4	34.2	35.26

Based on Table 6, each algorithm's best model performance is 48-50 percent. It means that the classification model can only analyze the Karonese text's sentiment in half of the available text. The best model performance of each algorithm is obtained from the scenario I or II. In other words, to produce the best classification model, the recommended dataset composition is 80: 20 or 60: 40. Meanwhile, the composition of the 40:60 dataset has only 30 percent performance. So that the composition of this dataset becomes the bad choice of all algorithms in the provided classification model. It hypothesizes that a large amount of training data allows a better classification model performance than smaller training data. However, if the amount of training data is disproportionate, there will be model overfitting [33].



Fig. 4 Comparison of performance machine learning algorithms to KSA without feature extraction

Fig. 4 summarizes Table 6 to show the performance comparison between algorithms in experiment I. Based on the figure, the SVM algorithm has the best accuracy but the worst performance. This is because the precision, recall, and F1 values of the SVM algorithm are the lowest of the other three algorithms. The performance of the K-NN algorithm has the highest Precision, Recall, and F1 scores among the other three algorithms. Thus, the K-NN algorithm outperforms in

analyzing Karonese sentiment without involving feature extraction.

Furthermore, both SVM and K-NN algorithms are classification algorithms with a similarity approach. In other words, to identify the class of a text, these two algorithms calculate the closeness between texts with a similarity function. Then, this experiment show that the classifier based on similarity function is superior to the Naïve Bayes algorithm (probabilistic based) or LR (predictive based).

TABLE VII Experiment II result (%)

Samarias	NB LR						SVM				K-NN					
Scenarios	Acc	Р	R	F1	Acc	Р	R	F1	Acc	Р	R	F1	Acc	Р	R	F1
Ι	47.7	48.32	48.72	48.52	47.03	48.7	46.6	47.6	51.7	49.5	48.25	48.9	50.13	50.32	51.8	51.05
II	48.1	49.9	50.02	50	50.13	50.2	51.19	50.7	52.8	51.2	54.5	52.8	50.03	49.01	49.8	49.4
III	35.1	31.3	34.1	32.6	44.15	44.3	43.03	43.7	47.4	45.5	49.1	47.23	46.2	43.3	48.4	45.71
IV	22.4	21.3	21.2	21.2	29.28	31.3	30.1	30.7	42.3	45.2	45.4	45.3	38.1	33.4	40.1	36.44

C. Experiment II

Experiment II is KSA using a machine learning algorithm with BoW. BoW functions as feature extraction. It extracts features from Karonese text and converts the text into vectors. Experiment II is also run using four scenarios to find the best model for each algorithm. The scenario follows the guidelines of Table 5. The results of experiment II can be seen in Table 7.

Based on Table 7, the best model performance of each algorithm is in the range of 48-54 percent. That is, the resulting classification model can only analyze the sentiment of the Karonese text half of the entire text. Scenario II is a recommendation for the best dataset composition to produce the best performance from machine learning algorithms with BoW. In other words, to produce the best classification model, it is obtained, and the recommendation for the composition of the dataset is 60:40. While the dataset composition of 40:60 is a bad choice as the composition of the dataset for all algorithms. It is because the scenario only performs in the range of 30-40 percent. It hypothesizes that a large amount of training data allows producing a classification model with better performance compared to smaller training data. However, if the amount of training data is disproportionate, there will be model overfitting [33].

Fig. 5 is a summary of Table 7, to show the performance comparison between algorithms in experiment II. Based on this figure, the SVM algorithm's performance with BoW outperforms other algorithms. It can be seen from the value of accuracy, Precision, Recall, and F1 score is the highest (above 51 percent). The next best algorithm is the K-NN classifier, with the value of accuracy, precision, recall and F1 in the 50-51 percent range. The Naïve Bayes algorithm with BoW is a bad combination to KSA, because the performance is the lowest than the others.



Fig. 5 Comparison of performance machine learning algorithms with BoW to KSA

Furthermore, both SVM and K-NN algorithms are similarity-based classification algorithms. In other words, to identify the class of a text, these two algorithms calculate the closeness between texts with a similarity function. Then, it can be concluded that the classifier based on similarity function is superior to the Naïve Bayes algorithm (probabilistic based) or LR (predictive based).

Further analysis, this study noted that BoW produces many vectors of text value 0 because there are many terms in a sentence; there is no occurrence in other text. Then, It resulted in a sparse matrix. As a result, this work confirms that the weakness of nave Bayes is the handle sparse dataset [33, 23].

TABLE VIII Experiment III result (%)

									~ /							
S		Ι	NB			I	LR			S	VM			K	-NN	
Scenarios	Acc	Р	R	F1												
Ι	51.7	51.8	51.3	51.55	44.9	44.7	45.6	45,15	58.1	58.5	57.2	57.84	51.2	55.3	53.4	54.3
II	34.1	27.6	40.1	32.7	50.1	50.4	52.9	51.62	53.8	50.2	54.7	52.35	40.2	41.7	40.3	41.09
III	45.5	44.9	44.3	44,6	51.2	54.1	51.6	52.82	40.3	39.8	40.8	40.3	50.1	51.3	49.8	50.54
IV	21.4	15.4	33.8	21.16	41.5	42.2	40.1	41.12	31.8	30.7	38.8	34.28	38.5	37.4	38.8	38.09

D. Experiment III

Experiment III is KSA using a machine learning algorithm with TF-IDF and is the final experiment of this study. The TF-IDF method is feature extraction, extracting features from Karonese text and measuring essential terms from available documents. Similar to experiments I and II, this experiment is run using four scenarios to find the best model for each algorithm. The scenario follows the guidelines of Table 5. Table 8 shows the results of experiment III.

Based on Table 8, the best model performance of each algorithm is above 50 percent. It means the classification model has succeeded in identifying the sentiment of Karonese text in more than half of the available documents. In this experiment, we cannot conclude the best scenario in producing the best model classification, sometimes scenario I, II, or III. However, we can recommend avoiding scenario IV (dataset composition is 40:60) to produce model classification, because the composition dataset gives a low model performance.



Fig. 6 Comparison of performance machine learning algorithms with TF-IDF to KSA

Fig. 6 summarizes Table 8 to show the performance comparison between algorithms in experiment III. Based on the figure, the SVM algorithm's performance with TF-IDF outperforms other algorithms. It can be seen from the value of accuracy, Precision, Recall, and F1 score is the highest (above 57 percent).

E. Influences of Feature Extraction

This section presents an analysis of the influences of feature extraction on machine learning to KSA. The effect of feature extraction is reviewed based on the accuracy value and F1 score of model classification. This work used two feature extraction methods for machine learning: BoW and TF-IDF. In general, this section presents the model's performance without feature extraction (baseline) and the model's performance with feature extraction (with BoW or TF-IDF).

Based on Fig. 7, the KSA model's accuracy increases after feature extraction. The BoW feature extraction can increase the accuracy of KSA model 0.1-2.1 percent, and TF-IDF can increase the accuracy of the KSA model by 0.17 - 7.4 percent. So, it can be said that feature extraction positively contributes to machine learning to produce KSA models. Furthermore, TF-IDF provides a greater increase in accuracy than the BoW to produce KSA model. The SVM and TF-IDF algorithms are a combination of machine learning and feature extraction algorithms with the best accuracy than other combinations.



Fig. 7 Influences of feature extraction on machine learning to accuracy of KSA model.

It is not sufficient to only consider the effects of feature extraction from the perspective of metric accuracy, because metric accuracy has a weakness against unbalancing classes of datasets [9, 23]. So, to complete the analysis of the effect of feature extraction, this section also presents a comparison of the F1 scores of algorithms on each experiment.



Fig. 8 Influences of feature extraction on machine learning to F1 score of KSA model

The F1 score is a performance representation of a classification model. Based on Fig. 8, the performance of KSA model increases after using feature extraction. BoW feature extraction can increase the performance of the ASF model by 0.05 - 3.6 percent. TF-IDF can increase the performance of the KSA model 1.65 - 8.64 percent. So, it can be said that feature extraction positively contributes to the performance of machine learning to produce KSA model.

Furthermore, feature extraction (BoW or TF-IDF) significantly affects the SVM algorithm, as evidenced by the gradient value larger than other algorithms. TF-IDF provides greater performance improvement than BoW on machine learning algorithm to produce KSA model. The SVM and TF-IDF algorithms are a combination of machine learning and feature extraction algorithms with the best performance compared to other combinations.

F. Comparison analysis with Previous Study

This section presents a comparative analysis of Karonese sentiment analysis. Research by [23] has been provided tweets Karonese sentiment analysis using machine learning. Based on Table 9, the performance of this work is slightly better than in previous studies. However, with more datasets and feature extraction, it turns out that the performance on this work has not given satisfactory results. The performance value is still below 60 percent.

We assume that the text pre-processing in this research is still weak. There are many existing text pre-processing techniques that are not available on Karonese text, and this requires further handling of text pre-processing to provide more optimal results.

TABLE IIX

COMPARISON ANALYSIS OF KSA RESEARCH											
Related work	Dataset	Algorithms	Accuracy								
		SVM	53								
[22]	Tweet	LR	51								
[23]	Karonese	NB	45								
		KNN	51								
This study	Multi domain	SVM	58.1								
-	text from social	LR	51.2								
	media	NB	51.7								
		KNN	54.3								

IV. CONCLUSION

This work provides an experiment of KSA using a machine learning algorithm. There are three kinds of experiments, KSA using machine learning, KSA using machine learning with BoW, and KSA using machine learning with TF-IDF. Karonese text dataset crawled from multi-domain social media, that is, Facebook, Twitter, Instagram, and YouTube. Sentiment label of Karonese text annotated by Karonese transcriber. Machine learning algorithms include LR, NB, SVM, and K-NN. Feature extraction (BoW and TF-IDF) has improved model performance in the 0.1 - 7.4 percent range. Overall, TF-IDF as feature extraction on machine learning has a better contribution than BoW. The combination of the SVM algorithm with TF-IDF is the combination with the highest performance. The value of accuracy is 58.1 percent, precision is 58.5 percent, recall is 57.2, and F1 score is 57.84 percent. However, this result is unsatisfactory, considering that the resulting performance is still below 60 percent. So, it needs a more intense pre-processing text to solve morphology and phonology Karonese text.

ACKNOWLEDGMENT

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. In addition, Researchers thank the Karonese community leaders for the opportunity and their thoughts to annotate the Karonese Dataset that has been obtained from various social media.

References

- A. M. Alayba, V. Palade, M. England dan R. Iqbal, "Arabic language sentiment analysis on health services," dalam 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR), Nancy, France, 2017.
- [2] I. P. M. Wirayasa, I. M. A. Wirawan dan I. M. A. Pradnyana, "Algoritma Bastal: Adaptasi Algoritma Nazief & Adriani untuk Stemming Teks Bahasa Bali," Jurnal Nasional Pendidikan Teknik Informatika: JANAPATI, vol. 8, no. 1, pp. 60-69, 2019.
- [3] Y. Cahyono dan 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, 2019.

- [4] F. Koto dan I. Koto, "Towards Computational Linguistics in Minangkabau Language: Studies on Sentiment Analysis and Machine Translation," dalam The 34th Pacific Asia Conference on Language, Information and Computation, 2020.
- [5] M. Diallo, C. Fourati dan H. Haddad, "Bambara Language Dataset for Sentiment Analysis," dalam International Conference on Learning Representations (ICLR), 2021.
- [6] M. S. Divate, "Sentiment analysis of Marathi news using LSTM," International Journal of Information Technology, vol. 13, no. 5, pp. 2069-2074, 2021.
- [7] 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.
- [8] M. H. Alam, M. M. Rahoman dan M. A. K. Azad, "Sentiment analysis for Bangla sentences using convolutional neural network," dalam 2017 20th International Conference of Computer and Information Technology (ICCIT), 2017.
- [9] K. Becker, V. P. Moreira dan A. G. dos Santos, "Multilingual emotion classification using supervised learning: Comparative experiments," Information Processing & Management, vol. 53, no. 3, pp. 684-704, 2017.
- [10] E. Tighe dan C. Cheng, "Modeling personality traits of filipino twitter users," Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, pp. 112-122, 2018.
- [11] W. Chansanam dan K. Tuamsuk, "Thai Twitter Sentiment Analysis: Performance Monitoring of Politics in Thailand using Text Mining Techniques," International Journal of Innovation, Creativity and Change, vol. 11, no. 2, pp. 436-452.
- [12] F. Djatmiko, R. Ferdiana dan M. Faris, "A Review of Sentiment Analysis for Non-English Language," dalam 2019 International Conference of Artificial Intelligence and Information Technology (ICAIIT), Yogyakarta, Indonesia, 2019.
- [13] K. Sailunaz dan R. Alhajj, "Emotion and sentiment analysis from Twitter text," Journal of Computational Science, vol. 36, 2019.
- [14] M. Rumelli dan D. Akkuş, "Sentiment Analysis in Turkish Text with Machine Learning Algorithms," dalam 2019 Innovations in Intelligent Systems and Applications Conference (ASYU), Izmir, Turkey, 2019.
- [15] A. Chakrabarty dan S. Roy, "A framework for medical text mining using a feature weighted clustering algorithm," dalam 2013 1st International Conference on Emerging Trends and Applications in Computer Science, Shillong, India, 2013.
- [16] M. Muchtar, W. Kembaren dan F. Repelita, "Translation Techniques and Quality in The English Version of Nganting Manuk Text," International Journal on Language, Research and Education Studies, vol. 2, no. 2, 2018.
- [17] G. Woollams, A Grammar of Karo Batak, Sumatra, Australia: Australian National University Canberra, 1996.
- [18] B. Tarigan, R. Sofyan dan R. N. Rosa, "Derivational morphology of Karonese ecolexicon," dalam Seventh International Conference on Languages and Arts (ICLA 2018), 2019.
- [19] S. B. Gurusinga, "Phonological Dialect Differences of Karonese Language in Medan, North Sumatra," Jurnal Culture (Culture, Language, and Literature Review), vol. 7, no. 2, pp. 263-275, 2020.
- [20] V. A. Fitri, R. Andreswari dan 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.
- [21] X. Lin, "Chinese Text Sentiment Analysis Based on Improved Convolutional Neural Networks," dalam 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), Beijing, China, 2019.
- [22] Y. Cahyono dan 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.
- [23] I. M. K. Karo, M. F. M. Fudzee, S. Kasim dan A. A. Ramli, "Sentiment Analysis in Karonese Tweet using Machine Learning," Indonesian Journal of Electrical Engineering and Informatics (IJEEI), vol. 10, no. 1, pp. 219-231, 2022.
- [24] S. S. Kumar, M. A. Kumar dan K. P. Soman, "Sentiment analysis of tweets in malayalam using long short-term memory units and convolutional neural nets," dalam International Conference on Mining Intelligence and Knowledge Exploration, 2017.
- [25] 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.

- [26] W. A. Qader, M. M. Ameen dan B. I. Ahmed, "An overview of bag of words; importance, implementation, applications, and challenges," dalam 2019 International Engineering Conference (IEC), 2019.
- [27] S. Qaiser dan 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.
- [28] R. a. D. H. a. S. Y. Sinnott, "Chapter 15—a case study in big data analytics: exploring twitter sentiment analysis and the weather," Big Data, pp. 357-388, 2016.
- [29] I. M. K. R. R. R. A. W. & A. B. Z. Karo, "A Hybrid Classification Based on Machine Learning Classifiers to Predict Smart Indonesia Program," dalam In 2020 Third International Conference on Vocational Education and Electrical Engineering (ICVEE) IEEE, 2020.
- [30] N. Cristianini dan J. Shawe-Taylor, An introduction to support vector machines and other kernel-based learning methods, Cambridge university press, 2000.

- [31] A. Al-Anazi dan . 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.
- [32] I. M. K. Karo, A. Khosuri dan R. Setiawan, "Effects of Distance Measurement Methods in K-Nearest Neighbor Algorithm to Select Indonesia Smart Card Recipient," dalam 2021 International Conference on Data Science and Its Applications (ICoDSA). IEEE, 2021.
- [33] R. Choudhary dan H. K. Gianey, "Comprehensive review on supervised machine learning algorithms," dalam 2017 International Conference on Machine Learning and Data Science (MLDS). IEEE, 2017.