ICIC Express Letters
Part B: Applications
Volume **15**, Number **1**, January **2024**

ICIC International ©2024 ISSN 2185-2766
pp. **101–108**

EMOTION CLASSIFICATION OF INDONESIAN TWITTER SOCIAL MEDIA TEXT USING SOFT VOTING ENSEMBLE METHOD

Yaquut Al Arsy Ilahi Rifai[1,*] and Derwin Suhartono[2]

[1]Computer Science Department, BINUS Graduate Program – Master of Computer Science
[2]Computer Science Department, School of Computer Science
Bina Nusantara University
Jl. K. H. Syahdan No. 9, Kemanggisan, Palmerah, Jakarta 11480, Indonesia
dsuhartono@binus.edu
*Corresponding author: yaquut.rifai@binus.ac.id

Received March 2023; accepted June 2023

Abstract. *Classifying emotion in text data is a beneficial task in various industries, various models are developed to tackle emotion classification tasks, but they are often in resource-rich languages. Research on Natural Language Processing (NLP) for the Indonesian language is limited. The objective of this study is to address the scarcity of research in the field of Indonesian emotion classification on Twitter data by introducing a soft voting ensemble approach, which integrates a Support Vector Machine (SVM) and three variants of IndoBERT models to leverage the strengths of each individual model and mitigate their weaknesses. As hypothesized, the result showed that the proposed ensemble model performed better compared to the singular model, with an accuracy score of 0.8023, higher than the accuracy of the SVM model, which was 0.6272, or the IndoBERT model, having an accuracy score of 0.7886 in Indonesian classification task.*

Keywords: Indonesian language, Natural language processing, Emotion classification, Ensemble, Soft voting

1. Introduction.

In the field of Natural Language Processing (NLP), understanding human languages through computational methods has been a prominent area of focus [1]. The integration of machine learning in NLP has made document classification, translation, summarization, and information extraction from text widely used in various industries. The detection and classification of emotions in texts play a vital role in analyzing large volumes of textual data more effectively [2]. Emotions, as an ongoing state of mind, are characterized by mental, physical, and behavioral symptoms [3], and have been modeled through various frameworks such as Eckman's Emotion Model, which categorized emotions into six distinct labels: happiness, anger, fear, disgust, sadness, and surprise [4].

The utilization of social media has experienced an unprecedented increase in recent years, providing a platform for individuals to engage in social interactions and disseminate information [5]. Information technology advances have enabled the expression of emotions through text-based posts on social media, particularly Twitter, serving as a rich data source for emotion classification. This has significant implications, allowing governmental entities to monitor public sentiment towards policies or political events. The field of NLP has made significant strides in recent years, producing a variety of models for a range of tasks [6-9]. However, these studies have largely focused on high-resources languages. Multiple languages, such as Indonesian, suffered from a limited amount of research due to their low-resource language being released publicly [10]. The task of understanding emotions in social media texts poses a significant challenge, not only for machines but also for humans, as the expression of emotions in these texts is often vague and indistinct, while the machine requires an accurate benchmark for the modeling of emotions [11].

DOI: 10.24507/icicelb.15.01.101

The contribution of this paper can be summarized as follows. We investigated machine learning models for emotion classification in Indonesian social media texts, specifically Twitter, and proposed an ensemble method to improve performance. The paper is organized as follows. Section 2 reviews the related work. Section 3 presents the dataset and methodology used for building our model. Section 4 explains the experiment and result. Finally, the last section discusses the conclusion.

2. Related Work.

2.1. **Non-Indonesian language models.** The study conducted by [12] employed two approaches for emotion classification: The Rule-Based Approach (RBA) and Machine Learning Approach (MLA). RBA was used for pre-processing and creating a knowledge base for MLA to classify emotions into four categories. MLA outperformed RBA with an accuracy of 88% compared to RBA's slightly over 85%. In [13], a study with two tasks was conducted: an offline training task, and an online classification task. The offline training task used Emotex to develop models for classifying emotion, while the online classification task used EmotexStream, a two-stage framework, to classify live streams of tweets in real time by first distinguishing between tweets with emotions and those without using a binary classifier, followed by a fine-grained emotion classification using Emotex. [14] proposed a hybrid pre-trained method combining BERT and a rule-based approach to improve emotion classification performance. They compared its performance to traditional machine learning models such as Naïve Bayes and LSTM using the Kaggle dataset. The hybrid model outperformed the traditional machine learning models in terms of accuracy; however, this study lacks consideration of how emotions are incorporated in the model. The study by [7] proposed a framework called REM, which used a predefined table to transform emoticons into related emotional words for improving emotion classification. The framework includes four steps of emoticon transformation, integer sequence production, vector padding, and LSTM classification. The study found that the proposed method showed higher proficiency in text containing emoticons while reduced performance in texts without emoticons. In [8], the authors studied the importance of emoticons in emotion classification by using Plutchik's wheel of emotion model. They collected and pre-processed 3.6 million tweets and annotated them using a list of eight emotions. Emojis that were not on the list were substituted with the corresponding words, and multiple emojis with different emotions were ignored. The proposed LSTM network outperformed five other classifiers, achieving an accuracy of 91.9%. The study by [15] proposed an ensemble model combining SVM and BERT for emotion recognition in tweets, which outperformed singular models with an accuracy of 0.91. This suggests that ensemble models have the potential to enhance emotion recognition tasks. Building upon these findings, the current study modifies the technique for the Indonesian language to further explore the potential of ensemble models.

2.2. **Indonesian language models.** Research on the Indonesian language in the field of NLP is scarce compared to other languages such as English, Chinese, or German [10]. Previous studies have focused on comparing rule-based and statistical methods [16]. The experiment showed that the rule-based method has a lower accuracy score of 63.172% compared to the statistical method with an accuracy of 71.740%. To handle an imbalanced dataset, the authors proposed using the Synthetic Minority Oversampling Technique (SMOTE), which resulted in an improved f-measure value of 83.202% for the statistical method. A study on Indonesian fairy tales [17] evaluated multiple machine learning algorithms, including NB, K-Nearest Neighbors (KNN), and SVM-Sequential Minimal Optimization (SVM-SMO) for text emotion detection. The dataset was pre-processed and emotion classification was conducted on four algorithms, with SVM-SMO achieving the best accuracy of 86% using TF-IDF weighting. [18] conducted a study on sentiment classification using LSTM, and benchmarked the result against NB and RB algorithms. The study classified public opinion into four emotional categories: Afraid, Anger, Sad, and Happy. RB was found to produce the best result of 68%, compared to Multinomial Naïve Bayes which has 66%.

3. **Methodology.** In this study, we present a novel methodology for emotion classification in Indonesian text data from Twitter social media, utilizing a soft-voting ensemble of Support Vector Machine (SVM) and three variations of IndoBERT models. Our approach begins by feeding the dataset into multiple pre-processing methods tailored to each machine learning algorithm. Subsequently, emotion classification is performed through the utilization of log probability vectors, which are then combined through vector addition to obtain the final result of the soft-voting ensemble method.

3.1. **Dataset.** The dataset used in this study is the EmoT dataset, which was originally collected and provided by [3]. The dataset contains a total of 4,401 tweets which are distributed among five emotion classes: Anger (1,101 samples), Fear (649 samples), Happy (1,017 samples), Love (637 samples), and Sadness (997 samples). The original usernames contained within the tweets, as represented by the @ symbol, have been substituted with the placeholder term "[USERNAME]". The pre-processing phase of the study involved replacing URLs and hyperlinks with the generic identifier "[URL]", and sensitive numerical information, such as phone numbers and courier tracking numbers, with the placeholder "[SENSITIVE-NO]". The research by [10] had the goal of benchmarking NLP models in various tasks in Indonesian languages, and the dataset has been pre-partitioned into training (80%), validation (10%), and testing (10%) subsets according to their study.

3.2. **Support vector machine.** The pre-processing phase of the SVM model begins with traditional machine learning pre-processing techniques: case folding, stripping extra space, character cleansing, and removing stop words. The removal of stop words has been proven to improve the performance of machine learning algorithms, and NLTK's list of stop words is used for this operation. The integration of *Satstrawi's* Indonesian stemming model will be experimented with as an additional factor. Tokenization and the utilization of the Term-Frequency Times Inverse Document-Frequency (TF-IDF) weighting method were employed to assess the significance of individual tokens in the dataset.

3.3. **IndoBERT.** Four variations of IndoBERT were tested in the training of the dataset: indobert-base-uncased [19], the original IndoBERT; indobertweet-base-uncased [20], a variant of IndoBERT trained on a Twitter dataset; indobert-classification, a fine-tuned indobert model on the indoNLU dataset; and indonesian-roberta-base-emotion-classifier, an IndoBERT version specifically trained for the emotion classification task. The selections of these four models were informed by the unique requirements of the task at hand, which was to classify emotions in tweets written in Indonesian. The original IndoBERT version, indobert-base-uncased, served as a benchmark for the study, providing a baseline level of performance for comparison with the other three models. The second model, indobertweet-base-uncased, was selected due to its specific training on the Twitter dataset, which was deemed essential given the nature of the task at hand, which involved the classification of emotion in tweets. The third model, indobert-classification, was chosen due to its fine-tuned nature on the dataset. Finally, the indonesian-roberta-base-emotion-classifier was chosen for its specific training for the emotion classification task, ensuring that the model was equipped with the necessary expertise to complete the task.

3.4. **Soft-voting ensemble.** The superiority of ensemble methods over singular machine learning models in various tasks has been demonstrated through scientific evidence [21]. The ensemble model utilized in this study was composed of two machine learning models, namely Support Vector Machine (SVM), and four variations of IndoBERT. The probabilities of each emotion class were calculated by aggregating the log probability vectors produced by the two models in the ensemble. It is important to note that, given the log probabilities have a range of values between $(-\infty, 0)$, the combination of the two probability vectors can result in two possible outcomes: while good results are likely to become even better, bad results are prone to deteriorate further.

4. **Experiments and Results.** The purpose of our research was to gain an understanding of the efficacy of the proposed model through a thorough evaluation. To achieve this objective, we first examined the result of the individual machine learning model that displayed the highest level of performance on the selected EmoT dataset for the emotion classification task. Subsequently, we conducted a comparative analysis of the SVM and four selected versions of IndoBERT. However, considering the nature of the proposed model that good results could be further improved, while poor results may worsen, we reduced the number of IndoBERT versions to three before ultimately presenting the results of the ensemble model.

4.1. **SVM model.** The experiment compared the performance of two SVM models, one stemmed and the other unstemmed, with similar pre-processing techniques and TF-IDF feature extraction. The stemmed model showed better performance, with an accuracy of 0.6723 on the test data set compared to the unstemmed model's accuracy of 0.6576. Therefore, the stemmed model was selected for the ensemble model. The result in Table 1 showcased the model has difficulties in detecting fear and sadness while having an easier time classifying love and anger. A confusion matrix was calculated to evaluate the model's ability to detect various emotions and is presented in Figure 1.

Table 1. Precision, recall, F1-score for stemmed SVM model

	Anger	**Fear**	**Happy**	**Love**	**Sadness**
Precision	0.60	0.87	0.61	0.78	0.52
Recall	0.82	0.51	0.56	0.70	0.51
F1-score	0.69	0.64	0.58	0.74	0.51

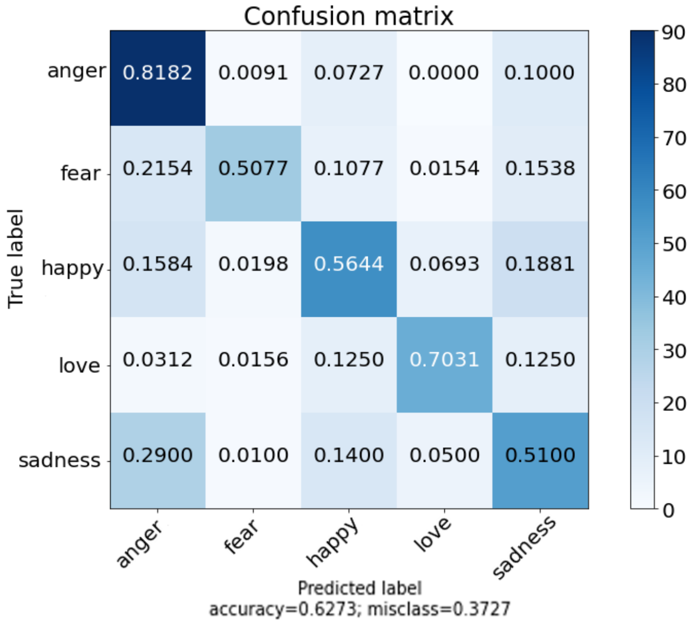

Figure 1. Confusion matrix for stemmed SVM model

4.2. **IndoBERT model.** Four pre-trained variants of IndoBERT were evaluated through a comprehensive analysis before selecting the three most promising models to further advance the emotion classification task. The models were trained for two epochs, 5 and 10, with a batch size of 32. The maximum padding length was set to 185, which was the longest sequence in the dataset, and stemming was used as a variable to assess the model's performance. A consistent pre-processing technique was applied to the model.

The indobert-base-uncased model's unstemmed set with 5-epoch training achieved an average accuracy of 0.6954, while the stemmed set reached 0.7272. Stemming improved the model's accuracy, precision, recall, and F1-score. However, for the 10-epoch training, the unstemmed set scored 0.7159, while the stemmed set had 0.6363. The impact of stemming is dependent on the number of training epochs, with stemming improving the model's performance for 5-epoch training, but not significantly affecting or even decreasing the performance for 10-epoch training. The indobertweet-base-uncased model experiment demonstrates an improved performance over the indobert-base-uncased model with regard to accuracy, precision, recall, and F1-score. The result shows that indobertweet model's accuracy increased with a lower epoch, with an accuracy of 0.7886 for 5 epochs and 0.784 for 10 epochs with the unstemmed set, and 0.7386 and 0.7522 with the stemmed set. The result suggests that the use of stemming had a small negative effect on the accuracy of the indobertweet model, with the unstemmed set performing slightly better on the testing data. In the indobert-classification model experiment, the relationship between stemming and epoch is rather similar to the indobertweet model. Although the model has a testing accuracy of 0.7431 for both 5-epoch and 10-epoch on the unstemmed set, the stemmed set has an accuracy of 0.6863 for 5-epoch and 0.6931 for 10-epoch, so far having lower relative performance compared to the previous two models when the models are stemmed. The last model is the indonesian-roberta-base-emotion-classifier; it was observed that in a lower epoch, its performance varied but relatively inferior to the indobertweet and indobert-base model. As the number of epochs increased, a slight boost in performance was observed in the testing data. However, the effect of stemming was found to have a negative impact on this model's performance, as evident in the decrease in accuracy for both high and low epochs. To better understand the relationship between epochs, stemming, and model performance, a comparison between IndoBERT models can be observed in Table 2.

Table 2. Comparative performance between IndoBERT models

Models	**Epoch**	**Test accuracy**	
		Unstemmed	**Stemmed**
indobert-base-uncased	5	0.6954	0.7272
	10	0.7159	0.6363
indobertweet-base-uncased	5	**0.7886**	**0.7386**
	10	0.784	0.7522
indobert-classification	5	0.7431	0.6863
	10	0.7431	0.6931
indonesian-roberta-base-emotion-classifier	5	0.725	0.6931
	10	0.7409	0.7113

4.3. **Ensemble model.** To evaluate the performance of the ensemble model, the log probability vectors were computed for both the validation and test sets using SVM and IndoBERT models. The result of the ensemble model was then obtained by aggregating the probability vectors from both models. The emotion class predicted for a tweet was determined as the one with the highest probability value in the respective probability vector in accordance with the soft voting ensemble theorem. Despite choosing the best

TABLE 3. Ensemble model performance on test set

	Epoch	Accuracy	Precision	Recall	F1-score
Unstemmed	5	**0.8023**	**0.8082**	**0.8067**	**0.8057**
	10	0.7659	0.7699	0.7728	0.7708
Stemmed	5	0.7704	0.7807	0.7744	0.7750
	10	0.7363	0.7460	0.7421	0.7433

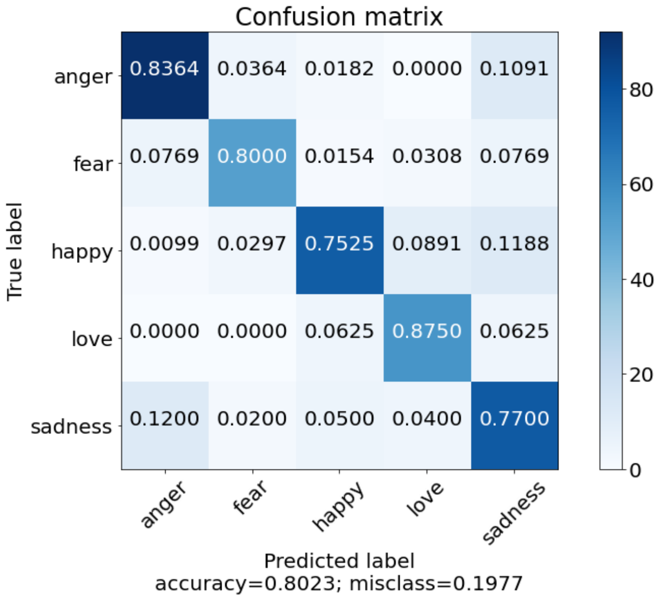

FIGURE 2. Confusion matrix for ensemble model

models to construct the ensemble model, it is still tested on similar conditions to the singular models. The results are shown in Table 3.

The ensemble model demonstrates a high-level performance, with a score of 0.8023 in accuracy, and 0.8082 in precision. An analysis of the confusion matrix, as depicted in Figure 2, reveals that the model's performance is commendable across all emotion classes, with the minimum accuracy score of 0.7525 being observed in the prediction of the happy emotion class.

4.4. **Comparative analysis.** As illustrated in Table 4, the ensemble model yielded the most optimal result, surpassing both the SVM and IndoBERT models with a remarkable accuracy of 0.8023. The result provides support for the hypothesis that utilizing an ensemble approach enhances the performance of language models.

TABLE 4. Comparative performance between ensemble vs singular models

	Accuracy	Precision	Recall	F1-score
SVM	0.6272	0.6723	0.6206	0.633
Indobertweet	0.7886	0.7922	0.7918	0.7918
Proposed ensemble	**0.8023**	**0.8082**	**0.8067**	**0.8057**

5. **Conclusion.** The present study aimed to investigate the performance of machine learning models in the classification of emotions expressed in an Indonesian social media text, specifically Twitter. The study focused on improving the performance of singular machine learning models through a soft voting ensemble method. The result showed that the ensemble model outperformed the singular model, with an accuracy score of 0.8023.

These findings support the earlier conjecture that ensemble models can improve the performance of language models by combining multiple models of different architecture. While our study may not have achieved state-of-the-art results, it has still made a valuable contribution to the field of NLP in Indonesian language. Our findings have shed new light on the challenges and limitations that exist in the current approach and have provided insights into potential directions for future research. For instance, the data used in the study was limited in terms of volume, emotion distribution, and representativeness, which could affect the generalizability of the result. Future work should focus on improving the pre-processing techniques used in the study, such as stemming, and increasing the data quality to further evaluate the performance of the models, or combining models with near-similar performance.

Acknowledgment. This work was supported by Bina Nusantara University. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] D. W. Otter, J. R. Medina and J. K. Kalita, A survey of the usages of deep learning for natural language processing, *IEEE Transactions on Neural Networks and Learning Systems*, vol.32, no.2, pp.604-624, 2021.

[2] T. Daouas and H. Lejmi, Emotions recognition in an intelligent Elearning environment, *Interactive Learning Environments*, vol.26, no.8, pp.991-1009, 2018.

[3] M. S. Saputri, R. Mahendra and M. Adriani, Emotion classification on Indonesian Twitter dataset, *2018 International Conference on Asian Language Processing (IALP)*, Bandung, Indonesia, pp.90-95, 2018.

[4] M. A. Riza and N. Charibaldi, Emotion detection in Twitter social media using long short-term memory (LSTM) and fast text, *International Journal of Artificial Intelligence and Robotics*, vol.3, no.1, pp.15-26, 2021.

[5] P. S. Dandannavar, S. R. Mangalwede and P. M. Kulkarni, Social media text – A source for personality prediction, *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, Belgaum, India, pp.62-65, 2018.

[6] J. Tanwijaya and D. Suhartono, Towards personality identification from social media text status using machine learning and transformer, *ICIC Express Letters, Part B: Applications*, vol.13, no.3, pp.233-240, 2022.

[7] J. Islam, M. A. H. Akhand, Md. A. Habib, M. A. S. Kamal and N. Siddique, Recognition of emotion from emoticon with text in microblog using LSTM, *Advances in Science, Technology and Engineering Systems Journal*, vol.6, no.3, pp.347-354, 2021.

[8] M. Krommyda, A. Rigos, K. Bouklas and A. Amditis, An experimental analysis of data annotation methodologies for emotion detection in short text posted on social media, *Informatics*, vol.8, no.1, 19, 2021.

[9] C. Aswin and W. A. Wella, Bidirectional encoder representations from transformers for cyberbullying text detection in Indonesian social media, *International Journal of Innovative Computing, Information and Control*, vol.17, no.5, pp.1599-1615, 2021.

[10] B. Wilie et al., IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding, *Proc. of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, Suzhou, China, pp.843-857, 2020.

[11] S. A. Deborah, T. T. Mirnalinee and S. M. Rajendram, Emotion analysis on text using multiple kernel Gaussian, *Neural Processing Letters*, vol.53, pp.1187-1203, 2021.

[12] M. Suhasini and S. Badugu, Two step approach for emotion detection on Twitter data, *International Journal of Computer Applications*, vol.179, no.53, pp.12-19, 2018.

[13] M. Hasan, E. Rundensteiner and E. Agu, Automatic emotion detection in text streams by analyzing Twitter data, *International Journal of Data Science and Analytics*, vol.7, no.1, pp.35-51, 2019.

[14] S. Madhuri and S. V. Lakshmi, Detecting emotion from natural language text using hybrid and NLP pre-trained models, *Turkish Journal of Computer and Mathematics Education*, vol.12, no.10, pp.4095-4103, 2021.

[15] I.-A. Albu and S. Spînu, Emotion detection from tweets using a BERT and SVM ensemble model, *arXiv.org*, arXiv: 2208.04547, 2022.

[16] A. R. Atmadja and A. Purwarianti, Comparison on the rule based method and statistical based method on emotion classification for Indonesian Twitter text, *2015 International Conference on Information Technology Systems and Innovation (ICITSI)*, Bandung, Indonesia, pp.1-6, 2015.

[17] Muljono, N. A. S. Winarsih and C. Supriyanto, Evaluation of classification methods for Indonesian text emotion detection, *International Seminar on Application for Technology of Information and Communication*, Semarang, Indonesia, pp.130-133, 2016.

[18] K. Hulliyah, N. S. A. A. Bakar, A. R. Ismail and M. O. Pratama, A benchmark of modeling for sentiment analysis of the Indonesian presidential election in 2019, *2019 7th International Conference on Cyber and IT Service Management (CITSM)*, Jakarta, Indonesia, pp.1-4, 2019.

[19] F. Koto et al., IndoLEM and IndoBERT: A benchmark dataset and pre-trained language model for Indonesian NLP, *arXiv.org*, arXiv: 2011.00677, 2020.

[20] F. Koto, J. H. Lau and T. Baldwin, IndoBERTweet: A pretrained language model for Indonesian Twitter with effective domain-specific vocabulary initialization, *arXiv.org*, arXiv: 2109.04607, 2021.

[21] F. M. Al-Kharboush and M. A. Al-Hagery, Features extraction effect on the accuracy of sentiment classification using ensemble models, *International Journal of Science and Research*, vol.10, no.3, pp.228-231, 2021.