Bulletin of Electrical Engineering and Informatics
Vol. 10, No. 5, October 2021, pp. 2780~2788
ISSN: 2302-9285, DOI: 10.11591/eei.v10i5.3157

Performance comparison of TF-IDF and Word2Vec models for emotion text classification

Denis Eka Cahyani[1], Irene Patasik[2]
[1]Department of Mathematics, Universitas Negeri Malang, Indonesia
[2]Department of Informatics, Universitas Sebelas Maret, Indonesiat

Article Info

Article history:

Received Apr 14, 2021
Revised Jul 27, 2021
Accepted Aug 30, 2021

Keywords:

Emotion
Support vector machine
Text classification
TF-IDF
Word2Vec

ABSTRACT

Emotion is the human feeling when communicating with other humans or reaction to everyday events. Emotion classification is needed to recognize human emotions from text. This study compare the performance of the TF-IDF and Word2Vec models to represent features in the emotional text classification. We use the support vector machine (SVM) and Multinomial Naïve Bayes (MNB) methods for classification of emotional text on commuter line and transjakarta tweet data. The emotion classification in this study has two steps. The first step classifies data that contain emotion or no emotion. The second step classifies data that contain emotions into five types of emotions i.e. happy, angry, sad, scared, and surprised. This study used three scenarios, namely SVM with TF-IDF, SVM with Word2Vec, and MNB with TF-IDF. The SVM with TF-IDF method generate the highest accuracy compared to other methods in the first dan second steps classification, then followed by the MNB with TF-IDF, and the last is SVM with Word2Vec. Then, the evaluation using precision, recall, and F1-measure results that the SVM with TF-IDF provides the best overall method. This study shows TF-IDF modeling has better performance than Word2Vec modeling and this study improves classification performance results compared to previous studies.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Denis Eka Cahyani
Department of Mathematics
Universitas Negeri Malang
Jl. Semarang No.5, Sumbersari, Lowokwaru, Malang, Jawa Timur 65145, Indonesia
Email: denis.eka.cahyani.fmipa@um.ac.id

1. INTRODUCTION

Emotion is the human feeling when communicating with other humans or reaction to everyday events [1]. Human emotions can be expressed in the form of facial expressions, voices, and text [2]. Recently, people used to convey their emotions in the text through social media. Text in social media can be classify as emotion or no emotion. Then the text that contains emotions according to Ekman is divided into six types, namely happy, sad, angry, scared, surprised, and disgusted [3]. The classification of emotion types from texts is of concern in the field of human-computer interaction (HCI), information retrieval (IR), and has been implement in many domains [4], [5]. Emotion classification in the text have received much attention to recognize human emotions in the text [6]-[8].

Text classification has a feature extraction stage to change an unstructured textual format into structured data so that data can be processed with machine learning algorithms for classification [9]. Feature extraction plays an important role in classification because the selection of an effective and appropriate method can affect classification performance [10]. The most widely used feature extraction technique is the

vector space model using the term frequency-inverse document frequency (TF-IDF) model approach [11]. Another feature extraction technique that is widely used is word embedding with the Word2Vec model approach [12]. TF-IDF modeling produces data with high dimensions, while Word2Vec modeling produces low dimensional data [13]. This is related to the computation time of the classification process and classification performance.

Several feature extraction techniques that can be used (TF-IDF and Word2Vec) make researchers often confused about which feature extraction technique is suitable for their research. Improper use of feature extraction techniques will result in longer computation time and not optimal classification performance results. Based on this background, it is important to do research related to comparison of the performance of the TF-IDF and Word2Vec models in classification. The goal of the research to obtain the best feature extraction model used for the text classification process. The best feature extraction is required for faster computation time in the classification process and and improved classification performance results [13].

The main contribution of this paper is present performance comparison of TF-IDF and Word2Vec models for emotion text classification. This paper improves a performance evaluation of research previously. Several previous studies such as the comparison of the emotions of Commuterline and Transjakarta users using the Multinomial Naïve Bayes method and the TF-IDF model were conducted by Cahyani [14]. While the use of Word2vec as feature extraction has not been used in that study [14]. Then the research on tweet emotion detection uses two stages of classification with the support vector machine (SVM) method and the maximum entropy method and the TF-IDF model conducted by [15]. Furthermore, the analysis of sentiment analysis from Twitter Messages using Word2vec by testing four classifiers, namely Gaussian Naive Bayes, Bernoulli Naive Bayes, SVM, Logistic Regression with two different test models, namely Skip Gram and CBOW in the Word2Vec algorithm was carried out by Acosta [16]. Then, the utilization of the Word2Vec model for sentiment analysis of product reviews with the SVM method was carried out by Fauzi [17]. This paper was different from previous studies because previous studies only used one of the TF-IDF or Word2Vec techniques to perform the classification process. Previous studies have not compared the performance of using TF-IDF and Word2vec models for classification processes in the same time with same data for detecting text emotions. Performance comparisons are needed so that we can find the best model that can be used to produce optimal text classification performance.

This paper discusses the performance comparison of the TF-IDF and the Word2vec model for classification of emotions in text. The classification algorithm used in this paper is the SVM and compared with the MNB (Multinomial Naïve Bayes) algorithm which was carried out in previous studies [14]. This research emotion classification is applied to Commuterline and Transjakarta tweet data which use Indonesian language.

2. RESEARCH METHOD

2.1. Preprocessing

The research method of the emotion text classification in this study is shown in Figure 1. The preprocessing stage is the process of preparing text data before it is processed in the system. Preprocessing is used in this study to select data so that the processed data becomes more structured. The preprocessing has four-step i.e. case folding, filtering, normalization, stop words removal, and stemming. Case folding is a task of converting text become lowercase. Filtering is a task of filtering the attributes of tweets i.e. links, mentions, URL, Normalization is a task of changing non-standard words into standard words. Stop words removal is a task of eliminating common word that have no meaning. Stemming is a task of removing the affixes in word [18], [19].

2.2. TF-IDF and Word2Vec model

In this stage, we perform modeling of TF-IDF and Word2Vec. TF-IDF is a method of weighting a word/term which gives a different weight to each term in a document based on the frequency of terms per document and the frequency of terms in all documents [20]. TF-IDF is used in this study because it provides better performance, especially in improving recall and precision values [21]. There is four-step in the TF-IDF model. The first step is the calculation of the frequency of occurrence of each word in each document (TF). It is shown in (1).

$$tf_t = 1 + \log(tf_t) \quad (1)$$

where; tf_t: number of occurrences of term t

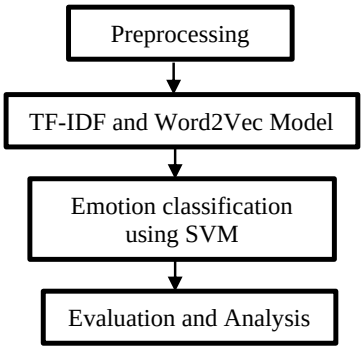


Figure 1. Research method

The second step is the calculation of the number of documents containing a specific word (DF). Then, the third step is the calculation of inverse DF (IDF). The calculation is shown in (2).

$$idf_t = log(\frac{D}{df_t}) \quad (2)$$

where: idf_t: inverse document frequency
D: number of document
df_t: the number of document that contains term t

The last step is the calculation of TF-IDF. TF-IDF is the multiplication of the TF results with the IDF calculation results for each word. The calculation is shown in (3).

$$W_{t,d} = tf_t \times idf_t \quad (3)$$

where: W: weigth of *term* (t) in document (d)
tf_t: number of occurrences of term t
idf_t: inverse document frequency that contains term t

The TF-IDF model will compare with the Word2Vec model. Word2Vec is the neural network that represents words in vector form [22]. Word2Vec is used in this study because it provides better performance for the semantic task in determining the association of a word with other similar words. For example, man is associated with boy or woman is associated with girl [23]. Word2Vec has two models i.e. continuous bag-of-words (CBOW) model and the continuous Skip-gram model. This study using Skip-gram because can better represent sparse words in data than the CBOW model [24]. The architecture of Skip-gram model is shown in Figure 2.

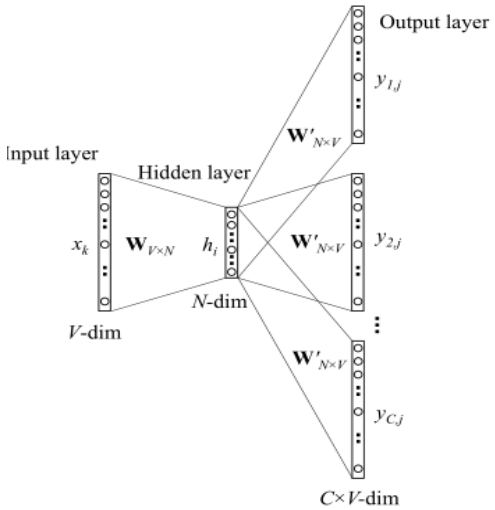


Figure 2. The architecture of Skip-gram model [25]

In the Skip-gram architecture, the model uses the current word as input to predict the surrounding context, where the Skip-gram will study the probability distribution of words in the context with a predefined window. The Skip-gram model has input layer, hidden layer, and output layer [25]. The input layer on Word2Vec is a one-hot vector, where one input word from the given vocabulary will be 1 and the other word will be 0. Each neuron in the input layer represents one word in the vocabulary. In the hidden layer, the number of neurons represents the number of dimensions of the word vector. The activation function in the hidden layer is linear, so the hidden layer neuron value is the input value multiplied by the weight value. The activation function in the hidden layer is shown in (4). Then, the value of the hidden layer is multiplied by a different weight value in the output layer that the function is shown in (5).

$$h = \mathrm{W}^T x \tag{4}$$

where:
h: hidden layer
W^{T}: transpose of weigth
x: input vector

$$u_j = \mathrm{W}'^T h \tag{5}$$

where: u_j: output line j to the hidden layer
W'^{T}: transpose of the weight from the hidden layer to the output layer

The number of neurons used in the output layer is the same as the number of neurons in the input layer that represents the target word. The output layer uses the Softmax activation function, where the Softmax activation function is shown in (6).

$$y_j = \frac{\exp(u_j)}{\sum_{j\prime=1}^{V} \exp(u_{j\prime})} \tag{6}$$

where: y_j: softmax output line j
u_j': output of all lines
V: number of vocabulary

2.3. Emotion classification using SVM

The resulting weigh of TF-IDF and word vector in the previous stage was utilized as the classification features. This study uses the SVM classification method. SVM is a classification method that is widely used in the field of text classification because of the superiority of its performance [26], [27]. SVM classification creates an ideal dividing line or hyperplane in a higher dimensional component space to map information with minimal risk [28]. If the existing data cannot be separated linearly (non-linearly), SVM is modified using the Kernel function, where the $\vec{x}$ data is mapped by the function $\Phi(\vec{x})$ to a vector space with a higher dimension. Furthermore, the learning process on SVM in finding support vector points relies on the multiplication of the dot product from the transformed data. Since the transformation is not easy to understand, the dot product calculation can be replaced with a kernel function. The kernel function is shown in (7).

$$K\left(\overrightarrow{x_i}, \overrightarrow{x_j}\right) = \Phi(\overrightarrow{x_i}).\Phi\left(\overrightarrow{x_j}\right) \tag{7}$$

$$f\left(\Phi(\vec{x})\right) = \sum_{i=1,\overrightarrow{x_i} \epsilon SV}^{n} a_i y_i K(\vec{x}, \overrightarrow{x_i}) + b \tag{8}$$

The classification results of the $\vec{x}$ data is shown from (8), where α_i is Lagrange multipliers, which is zero or positive ($\alpha_i \geq 0$), y_i is the class of test data x_i, b is bias, n is the number of samples in the training set, and SV is a support vector, which is a subset of the training set that has a Lagrange multipliers value greater than 0 ($\alpha_i > 0$). The kernel functions that can be used in SVM are Linear, Polynomial, Sigmoid, and Radial Basis Function (RBF) [29]. This study uses a linear kernel function because have good performance, fast, and only require few parameter compared with other kernels [30].

In this study, the SVM classification applies the 10-fold cross-validation technique. The 10-fold cross-validation is a technique that uses the entire dataset as training data and testing data where the classification process is carried out 10 times with various forms of training and testing data [31].

2.3. Evaluation and analysis

At the evaluation stage, the calculation of accuracy precision, recall, and f1-measure are performed as shown in (9)-(12) [32].

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (9)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

$$F1 - Measure = 2 * \frac{(Precision*Recall)}{(Precision+Recall)} \quad (12)$$

In the analysis stage, we compare the results of the SVM with TF-IDF and SVM with Word2Vec classification. The results were also compared with the methods used in previous studies.

3. RESULTS AND DISCUSSION

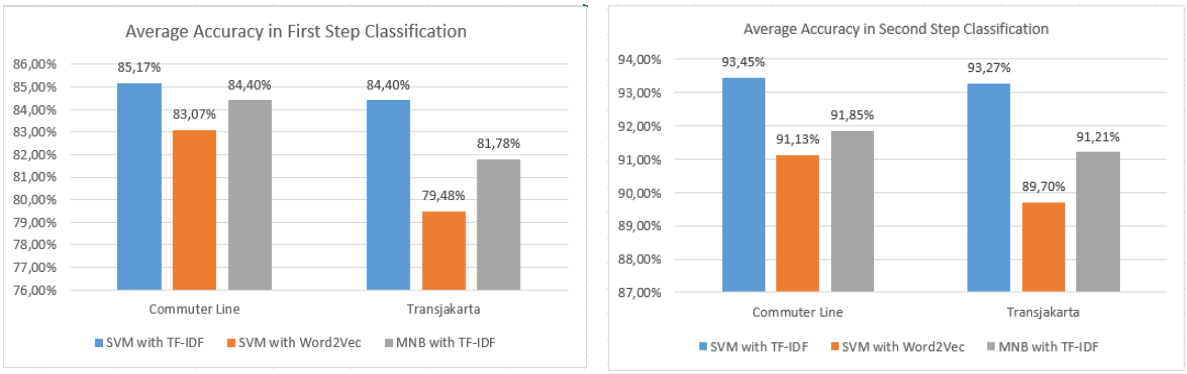
This study uses data crawling from tweets of Transjakarta and Commuterline users. Query search in data collection uses the official Transjakarta (@PT_Transjakarta) and Commuterline (@CommuterLine) accounts. Tweet data was obtained from January 1, 2017 to September 30, 2017. All of the dataset is in Indonesian language. This experiment used python programming language with the some library i.e. scikit, numpy, pandas and gensim. The experiment in this study is a continuation of previous research [14], so this research experiment uses the same data. Table 1 shows the experimental dataset in this study.

Table 1. Experimental dataset

Class	Category of Emotions	Commuter line	Transjakarta
No Emotion	-	57,134	27,144
Emotion		20,395	10,649
	Happy	4,289	2,633
	Angry	15,365	7,619
	Sad	507	265
	Fear	190	116
	Surprised	44	16

The classification is divided into two steps. The first step classifies the tweet data into emotion and no emotion. The classification results in this study also are compared with the results of previous studies [14]. The result data in the first step classification that contains emotion tweets then processed in the second step classification. The second step classifies tweets that contain emotions into five types of emotions i.e. happy, angry, sad, scared, and surprised. Figure 3 shows a comparison of average accuracy in the first step and second step classification between SVM with TF-IDF, SVM with Word2Vec and MNB with TF-IDF that conducted in previous studies [14]. We not combine MNB with Word2Vec because Word2Vec vectors sometimes contain negative values, MNB classifier does not allow for negative values in the document vectors. It should be possible to scale all vectors uniformly to avoid negative values but this result in poor performance [33].

Figure 3 shows that, the SVM with TF-IDF method generates the highest accuracy compared to other methods for Commuterline and Transjakarta data, both in the first step and second step classification. Then followed by the MNB method with TF-IDF, and the last is SVM method with Word2Vec. This shows that TF-IDF modeling has better performance than Word2vec modeling. Also in general, the accuracy generated by the commuter line data is better than the Transjakarta data for each method. This is because the number of Commuterline data is bigger than Transjakarta data so that the features of Commuterline for the classification process are more diverse. With the many various features in the commuter line data, the resulting accuracy value on the commuter line data is higher than the transjakarta data.



(a) (b)

Figure 3. Comparison of average accuracy in; (a) first step and (b) second step classification

The classification also measures precision, recall, and F1-measure on Commuter line and Transjakarta data. The precision, recall and F1-measure values affect how well the system performs in recognizing an emotion. Figure 4 shows the results of the comparison precision, recall and F1-measure in first step classification. Figure 4 shows the SVM with TF-IDF method provides the best overall precision, recall and F1-measure. This shows that classification using the SVM method with TF-IDF succeed generates the system work properly to recognize emotion and no-emotion data. Furthermore, the second order resulted in the MNB method with TF-IDF although the results with the first order were not much different. Meanwhile, the classification of SVM with Word2vec in third place has a significant difference when compared to the classification of SVM with TF-IDF. This also proves that TF-IDF modeling has better performance than Word2Vec modeling. In general, the precision, recall, and F1-measure generated by the commuter line data is better than the Transjakarta data for each method.

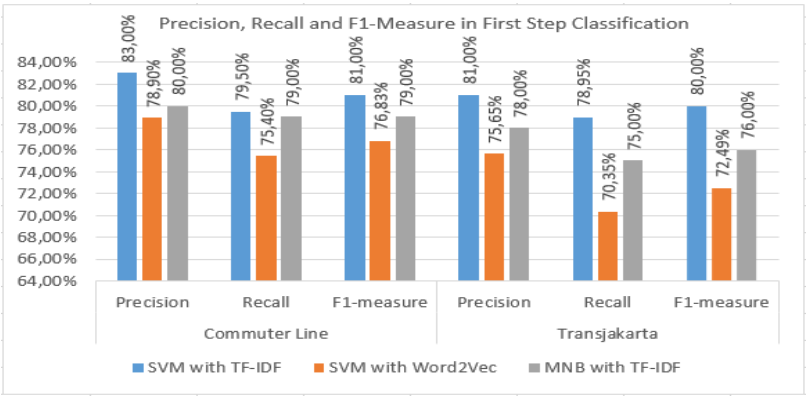
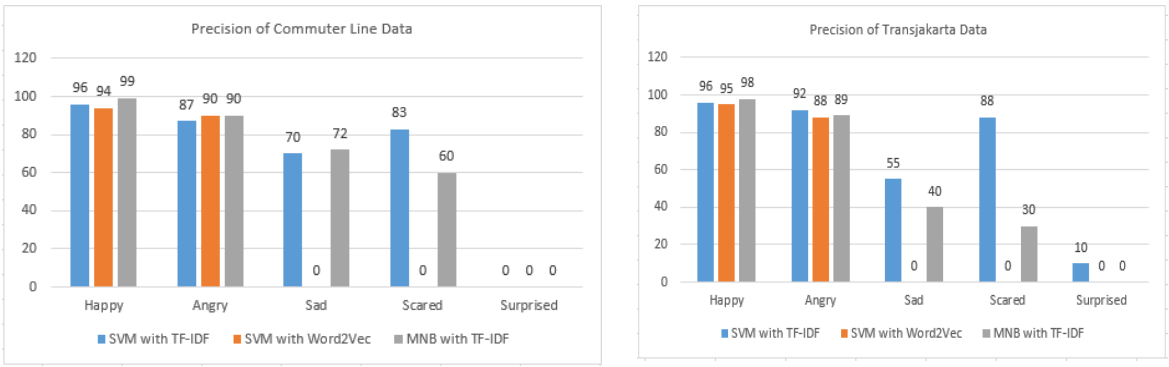


Figure 4. Comparison of precision, recall and F1-measure in first step classification

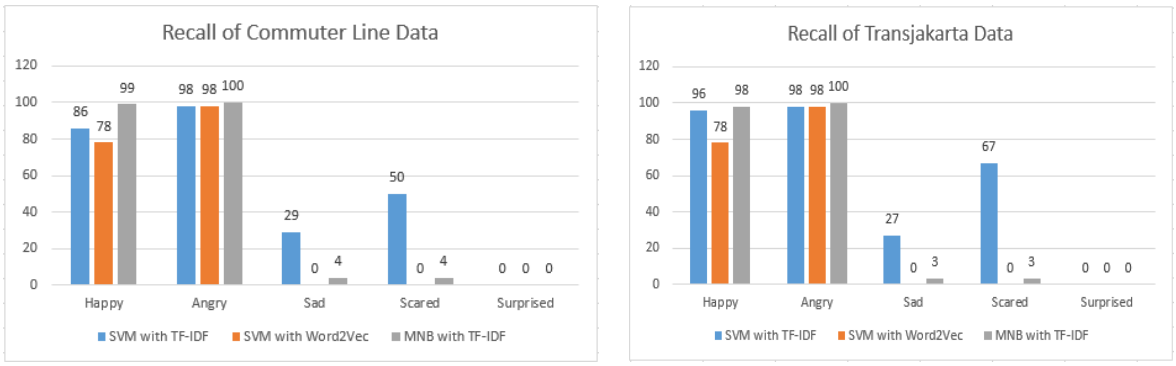
The precision, recall and F1-measure were also generated in the second step classification. The average precision, recall, and F1-measure values on the Commuterline and Transjakarta data are presented in Figure 5-7. In Figure 5, the precision value shows the system performance in the three methods is good enough to recognize happy and angry emotions. However, the SVM with Word2Vec method does not succeed in recognizing the emotions of sad, scared and surprised. This is because the data for the emotional class for sad, scared and surprised have a small number. Word2Vec requires a large number of data to learn word representations and to place words that are similar to a closer position so that Word2vec cannot recognize emotions with small data. For the surprised emotions, the three methods fail to recognize emotions correctly on the commuter line data, while for the Transjakarta data the SVM with TF-IDF method has a low precision value. The precision of surprised emotions is low for all metode because the number of surprised emotions is a minority of data which has a large difference in the number of other emotions so that when the surprise emotion is classified, it is classified into other emotions.



(a) (b)

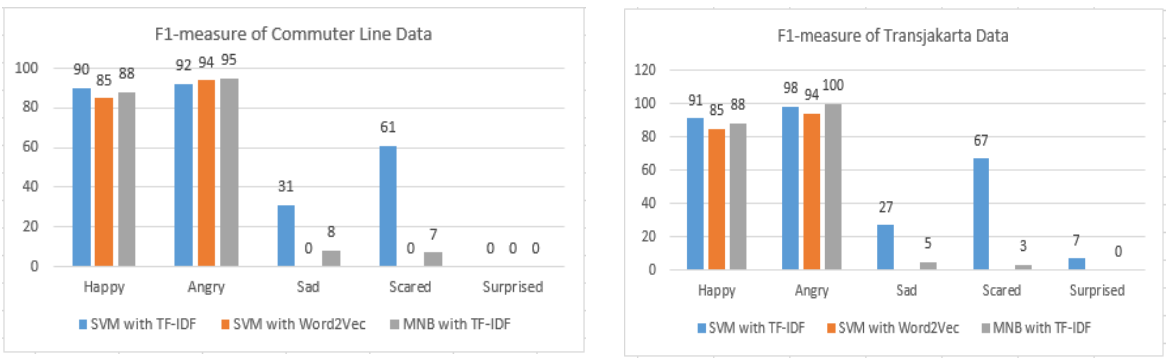
Figure 5. Comparison of precision in second step classification: (a) Commuter line (b) Transjakarta

In Figure 6 and Figure 7, the recall and F1-Mmasure values show that the performance of the three methods is good enough to recognize happy and angry emotions. However, for sad and fearful emotions, there are significant differences where the SVM with TF-IDF method is better at recognizing sad and fearful emotions compared to MNB with TF-IDF. The precision, recall, and F1-measure values in the MNB with TF-IDF method are low, which means that the MNB with TF-IDF method is less able to recognize sad and fearful emotions. Meanwhile, in the SVM with Word2Vec method, the recall and f1-measure values are zero, which means that this method fails to recognize sad, scared, and surprised emotions. This is because the data for the emotional class for sad, scared and surprised have a small number. Then based on the recall value on both the data and the F1-measure value on the Commuter line data, the three methods cannot recognize surprised emotions because the value obtained is zero. Meanwhile, the F1-measure value of the SVM with TF-IDF method on the Transjakarta data has an F1-measure value, although it is low. This means that the SVM with TF-IDF method can identify surprised emotions even though not optimal.



(a) (b)

Figure 6. Comparison of recall in second step classification: (a) Commuter line (b) Transjakarta



(a) (b)

Figure 7. Comparison of F1-measure in second step classification: (a) Commuterline (b) Transjakarta

The best method for measuring precision, recall and F1-measure is SVM with TF-IDF. This is because the SVM with TF-IDF method can recognize the five types of emotions, including recognizing surprised emotions where other methods fail to recognize surprised emotions. On the other hand, the SVM with Word2Vec method can only recognize happy and angry emotions and cannot recognize other emotions. So this shows the TF-IDF model's performance is better than the Word2Vec model for recognizing every type of emotion (happy, angry, sad, scared, surprised) based on precision, recall and F1-measure values.

TF-IDF model's performance is better than the Word2vec model because the number of data in each emotion class is not balanced and there are several classes that have a small number of data. The number of surprised emotions is a minority of data which has a large difference in the number of other emotions. In the small data, Word2Vec can not collect the semantic and syntactic information of words properly. Word2Vec need large training data to learn the word representation. Meanwhile, TF-IDF modeling can generate good accuracy even with a small number of data.

Evaluation of classification performance in this study improves classification performance results compared to previous studies [14]. The SVM with TF-IDF method used in this study gave better results than the MNB with TF-IDF method in previous studies. In the first and second steps accuracy evaluation, the accuracy value of the SVM with TF-IDF method is better than the MNB and TF-IDF methods. Likewise in the evaluation of precision, recall and F1-measure, the SVM with TF-IDF method is superior to the MNB method with TF-IDF. So that in this study, we have the advantage of improving the results of evaluation of accuracy, precision, recall and F1-measure for emotion text classification.

4. CONCLUSION

In this study we compared the performance of the TF-IDF and Word2Vec models to represent features in the emotional text classification. We use the SVM and MNB methods for classification of emotional text on Commuterline and Transjakarta tweet data. The classification is divided into two steps, namely the first step to determine whether a tweet contains emotions or does not contain emotion, and the second step is to determine a tweet that contains emotions into five types of emotions (happy, angry, sad, scared and surprised). In this study we used three scenarios of classification methods, namely SVM with TF-IDF, SVM with Word2Vec and MNB with TF-IDF. The SVM with TF-IDF method generate the highest accuracy compared to other methods in the first dan second steps classification, then followed by the MNB with TF-IDF, and the last is SVM with Word2Vec. Then, the evaluation using Precision, Recall and F1-Measure results that The SVM with TF-IDF provides the best overall method in the first and second steps classification. The SVM with TF-IDF method succeed to recognize emotion and no-emotion data in first step classification and succeed recognize the five types of emotions in second step. This shows that TF-IDF modeling has better performance than Word2vec modeling in classification emotion text. Evaluation of classification performance in this study using SVM with TF-IDF improves classification performance results compared to previous studies that using MNB with TF-IDF.

In the future work, the researchers are expected to use balanced data on each emotion class and large amounts of data. A large and balanced amount of data in each class is needed to improve the performance of the feature extraction technique so that it affects classification performance. Futhermore, the researchers can also combine TF-IDF and Word2Vec as feature extraction for text classification.

REFERENCES

[1] E. F. Pace-Schott *et al.*, "Physiological feelings," *Neuroscience and Biobehavioral Reviews*. 2019, doi: 10.1016/j.neubiorev.2019.05.002.

[2] M. M. Saad, N. Jamil, and R. Hamzah, "Evaluation of support vector machine and decision tree for emotion recognition of malay folklores," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 7, no. 3, pp. 479-486, 2018, doi: 10.11591/eei.v7i3.1279.

[3] L. Jenkins, "Does Personality effect Facial Emotion Recognition? A Comparison between the older Ekman Emotion Hexagon Test and a newly created Measure," *Madridge J. Neurosci.*, 2017, doi: 10.18689/mjns-1000107.

[4] M. Jeon, "Chapter 1-Emotions and Affect in Human Factors and Human–Computer Interaction: Taxonomy, Theories, Approaches, and Methods," *Emotions and Affect in Human Factors and Human-Computer Interaction*, pp. 3-26, 2017, doi: 10.1016/B978-0-12-801851-4.00001-X.

[5] L.-A.-M. Bostan and R. Klinger, "An Analysis of Annotated Corpora for Emotion Classification in Text," *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 2104-2119.

[6] J. Hofmann, E. Troiano, K. Sassenberg, and R. Klinger, "Appraisal Theories for Emotion Classification in Text," *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 125-138, 2021, doi: 10.18653/v1/2020.coling-main.11.

[7] M. Kumari and P. Chaudhary, "Deep Learning Technique Based Text Emotion Classification System Using Genetic Algorithm Technique," *Int. J. Eng. Sci. Res. Technol.*, vol. 10, no. 1, pp. 28-40, 2017, doi:

10.29121/ijesrt.v10.i1.2021.2.
[8] X. Shi, X. Kang, P. Liao, and F. Ren, "Building Label-Balanced Emotion Corpus Based on Active Learning for Text Emotion Classification," in *Studies in Computational Intelligence*, Aug. 2021, vol. 917, pp. 13–24, doi: 10.1007/978-3-030-56178-9_2.
[9] R. Feldman and J. Sanger, "The text mining handbook: advanced approaches in analyzing unstructured data," *Choice Rev. Online*, 2007, doi: 10.5860/choice.44-5684.
[10] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341-348, 2019, doi: 10.1016/j.procs.2019.05.008.
[11] Y. Zhang, Y. Zhou, and J. T. Yao, "Feature Extraction with TF-IDF and Game-Theoretic Shadowed Sets," In: Lesot MJ. *et al.* (eds) *Information Processing and Management of Uncertainty in Knowledge-Based Systems. IPMU 2020. Communications in Computer and Information Science*, vol 1237, 2020, doi: 10.1007/978-3-030-50146-4_53.
[12] E. M. Alshari, A. Azman, S. Doraisamy, N. Mustapha, and M. Alkeshr, "Improvement of sentiment analysis based on clustering of Word2Vec features," *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, 2017, pp. 123-126, doi: 10.1109/DEXA.2017.41.
[13] W. Zhu, W. Zhang, G. Z. Li, C. He, and L. Zhang, "A study of damp-heat syndrome classification using Word2vec and TF-IDF," *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2016, pp. 1415-1420, doi: 10.1109/BIBM.2016.7822730.
[14] D. E. Cahyani, F. Meilani, and S. W. Sihwi, "Emotions comparison of commuter line and transjakarta users based on twitter using multinomial naïve bayes classifier," *2019 5th International Conference on Science and Technology (ICST)*, 2019, pp. 1-6, doi: 10.1109/ICST47872.2019.9166171.
[15] J. E. The, A. F. Wicaksono, and M. Adriani, "A two-stage emotion detection on Indonesian tweets," *2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2015, pp. 143-146, doi: 10.1109/ICACSIS.2015.7415174.
[16] J. Acosta, N. Lamaute, M. Luo, E. Finkelstein, and A. Cotoranu, "Sentiment Analysis of Twitter Messages Using Word2Vec," *Proc. Student-Faculty Res. Day, CSIS, Pace Univ.*, 2017.
[17] M. A. Fauzi, "Word2Vec model for sentiment analysis of product reviews in Indonesian language," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 9, no. 1, pp. 525-530, 2019, doi: 10.11591/ijece.v9i1.pp525-530.
[18] N. Indurkhya and F. J. Damerau: "Text Preprocessing David D. Palmer," in *Handbook of Natural Language Processing*, 2020.
[19] L. Hickman, S. Thapa, L. Tay, M. Cao, and P. Srinivasan, "Text Preprocessing for Text Mining in Organizational Research: Review and Recommendations," *Organ. Res. Methods*, Nov. 2020, doi: 10.1177/1094428120971683.
[20] J. Leskovec, A. Rajaraman and J. D. Ullman, "Mining of massive datasets," Cambridge University Press, 2011.
[21] T. Tokunaga and M. Iwayama, "Text categorization based on weighted inverse document frequency," *Tech. Rep. Tokyo Inst. Technol.*, pp. 5-31, 1994.
[22] Liu Wensen, Cao Zewen, Wang Jun and Wang Xiaoyi, "Short text classification based on Wikipedia and Word2vec," *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*, 2016, pp. 1195-1200, doi: 10.1109/CompComm.2016.7924894.
[23] C. Nicholson, "A Beginner's Guide to Word2Vec and Neural Word Embeddings | Pathmind," *https://wiki.pathmind.com/word2vec*. https://wiki.pathmind.com/word2vec (accessed Jul. 20, 2021).
[24] A. J. Landgraf and J. Bellay, "Word2vec skip-gram with negative sampling is a weighted logistic PCA," *arXiv*. 2017.
[25] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," *Proceedings of the 26th International Conference on Neural Information Processing Systems*, vol. 2, pp. 3111–3119, 2013.
[26] N. S. A. Yasmin, N. A. Wahab, A. N. Anuar, and M. Bob, "Performance comparison of SVM and ANN for aerobic granular sludge," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 8, no. 4, pp. 1392-1401, 2019, doi: 10.11591/eei.v8i4.1605.
[27] L. K. Ramasamy, S. Kadry, and S. Lim, "Selection of optimal hyper-parameter values of support vector machine for sentiment analysis tasks using nature-inspired optimization methods," *Bulletin of Electrical Engineering and Informatics (BEEI)*, vol. 10, no. 1, pp. 290-298, 2021, doi: 10.11591/eei.v10i1.2098.
[28] P. K. Sethy, S. K. Behera, P. K. Ratha, and P. Biswas, "Detection of coronavirus disease (COVID-19) based on deep features and support vector machine," *Int. J. Math. Eng. Manag. Sci.*, 2020, doi: 10.33889/IJMEMS.2020.5.4.052.
[29] R. Gholami and N. Fakhari, "Support Vector Machine: Principles, Parameters, and Applications," in *Handbook of Neural Computation*, pp. 515-535, 2017, doi: 10.1016/B978-0-12-811318-9.00027-2.
[30] A. McCallum and K. Nigam, "A Comparison of Event Models for Naive Bayes Text Classification," *AAAI/ICML-98 Work. Learn. Text Categ.*, 1998, doi: 10.1.1.46.1529.
[31] T. Fushiki, "Estimation of prediction error by using K-fold cross-validation," *Stat. Comput.*, vol. 21, pp. 137-146 2011, doi: 10.1007/s11222-009-9153-8.
[32] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining (New International Editon)," Pearson Education India, 2014.
[33] Ashok Shilakapati, "Word Embeddings and Document Vectors-When in Doubt, Simplify," Accessed on: Jul. 21, 2021 [Online], Available: https://towardsdatascience.com/word-embeddings-and-document-vectors-when-in-doubt-simplify-8c9aaeec244e.