Accredited Ranking SINTA 2
Decree of the Director General of Higher Education, Research, and Technology, No. 158/E/KPT/2021
Validity period from Volume 5 Number 2 of 2021 to Volume 10 Number 1 of 2026

Published online on: **http://jurnal.iaii.or.id**

JURNAL RESTI
(Rekayasa Sistem dan Teknologi Informasi)
Vol. 6 No. 1 (2022) 78 - 84 ISSN Media Electronic: 2580-0760

Feature Expansion Word2Vec for Sentiment Analysis of Public Policy in Twitter

Alvi Rahmy Royyan[1], Erwin Budi Setiawan[2]
[1,2]Informatics, School of Computing, Telkom University
[1]alvirahmyr@student.telkomuniversity.ac.id, [2]erwinbudisetiawan@telkomuniversity.ac.id*

Abstract

Social media users, especially on Twitter, can freely express opinions or other information in the form of tweets about anything, including responding to a public policy. In a written tweet, there is a limit of 280 characters per tweet and this allows for problems such as vocabulary mismatches. Therefore, in this study, the feature expansion Word2vec method was applied to overcome when the vocabulary mismatches occur. This study implements and compares the Twitter sentiment analysis using the feature expansion Word2vec method and the baseline model. To perform classification on this sentiment data, two different machine learning algorithms including Support Vector Machine (SVM) and Logistic Regression (LR) are used to compare the model. The result is feature expansion Word2Vec with SVM classifier has a slightly better performance which succeeded in increasing the system accuracy up to 0,99% with 78,99% accuracy score, rather than LR classifier which achieved 78,31% accuracy score.

Keywords: sentiment analysis, feature expansion, word2vec, public policy

1. Introduction

Public policy is a regulation or set of rules that are legally allocated by the government that affects everyone in a country. The government, which has the authority to make public policy, is expected to be able to make a policy that solves problems and brings goodness to various parties, especially the community. However, as a matter of fact, some of the public policies made not only have positive effects, but also negative effects. As a result of this, various pros and cons responses were expressed by users of social media networks especially in one of the well-known microblogging sites, namely Twitter. A microblog is a web that allows users to exchange content such as short sentences, images, or video links [1]. In uploading a tweet on Twitter, users are only given a sentence up to 280 characters. This can increase the use of word variations and emoticons so that vocabulary mismatches are possible.

Nowadays people use Twitter as an alternative platform to respond and express their opinion on a problem that occurs in their surrounding environment. The increasing number of Twitter users and more than 500 million tweets per day sent [2], has attracted the attention of research academics widely and has produced many diverse studies to obtain important information about an event. One of them is a Twitter sentiment analysis which is currently a popular topic for research [3]. Text sentiment analysis is an automated process to determine whether a text segment contains an objective or opinion content, and then it can determine the polarity of the sentiment [4]. Sentiment analysis is useful, as it collects and classifies public opinion by analyzing large social data. Twitter sentiment classification aims to classify these sentiments whether the tweet polarity is positive, negative, or neutral [5].

Fauzi *et al.* [6], presented an experiment to analyze the sentiment of 772 reviews of products from the FemaleDaily website in Bahasa Indonesia using the SVM classification with the Word2Vec as a word embedding method. This method is compared with other methods such as Bag-of-Words (BOW) with TF-IDF, Raw TF, and Binary TF. As a result, sentiment analysis using SVM, in general, has a good performance with an average accuracy value is 81% and the highest accuracy is 85% when using the BOW with

Accepted: 26-09-2021 | Received in revised: 18-02-2022 | Published: 27-02-2022

TF-IDF. However, the accuracy value when using Word2Vec with SVM has the lowest score which only reaches 70%. They evaluated the low accuracy caused by the small amount of dataset to train the Word2Vec model. A small data train makes it difficult for Word2Vec to capture the semantic and syntactic information very well and the words cannot be moved closer into the position of words that similar to them because it does not have so many varied examples to moved them all into better positions. Research by Kurniawan *et al.* [7], explored an experiment of Twitter sentiment analysis related to natural disasters in Indonesia using the Continuous Bag-of-Words (CBOW) and Skip-Gram models as word embedding techniques from Word2Vec and using SVM classification. They found that the differences between CBOW and Skip-Gram architectural model and the number of dimensions also affect the classification results. This study compares the two Word2Vec model techniques with various dimensions which produce the highest precision value of 0,644, recall 0,58, and F1-Score value of 0,611 in the Skip-Gram model with dimension size of 100. They suggest using a greater amount of data to increase data variation and apply preprocessing steps that are more suitable with the dataset. Erwin *et al.* [8] presented an experiment of Twitter sentiment analysis using the Logistic Regression classification reached 98,81% for a hybrid method that combined basic features with feature expansion based on tweet-based features. They found that the proposed method was able to increase the accuracy system than the baseline. In research by S. P. Sheela [9] conducted an experiment of sentiment analysis and prediction for online review without ratings like from Amazon's customer review. This paper used various classifiers model to group the reviews into two categories as reviews with ratings or blank ratings, and then make a prediction of sentiments for the reviews without ratings. This paper observed that the Logistic Regression and Naïve Bayes predictions of opinions are much similar than the Multinomial Naïve Bayes and Bernoulli classifiers. Sheela stated that Logistic Regression performance is better than the Naïve Bayes, it achieved higher precision value of 0,92, recall 0,93, and F1 Score value of 0,93, and accuracy score of 93%. Then, Erwin *et al.* [10] introduced a method called feature expansion using word embedding Word2Vec for tweet topic classification. Writing a limited sentence length such as a tweet can reduce the meaning of the original sentence, so this method is proposed to reduce vocabulary mismatches. This study shows that the proposed method can affect and produce an accuracy value of 58,86% and increase the accuracy up to 0,38% when using the Logistic Regression.

Based on various researches that have been described and to the best of the author's knowledge, there has been no research on Twitter sentiment analysis using tweets in Bahasa Indonesia about public policies that implement the feature expansion using Word2Vec. To perform classification on this data set, two different machine learning algorithms including Support Vector Machine (SVM) and Logistic Regression (LR) are used to compare the model. Thereafter, to improve the system's accuracy, the feature expansion method implemented to overcome the problem of limited character length in a tweet that can increase the use of word variations and allow for vocabulary mismatches.

The rest of this paper is organized as follows. Section 2 describes the feature expansion Word2Vec method for sentiment analysis with two different classifier using SVM and LR to compare the model. Section 3 provides the results and discussion of experiments. Section 4 concludes the results.

2. Research Methods

Figure 1 shows the system plan of the sentiment analysis using feature expansion Word2Vec, which consists of data crawling, pre-processing, implementing the feature extraction or the base method, then the feature expansion using Word2Vec, classification process performs by using SVM and LR, and finally evaluating the performance.

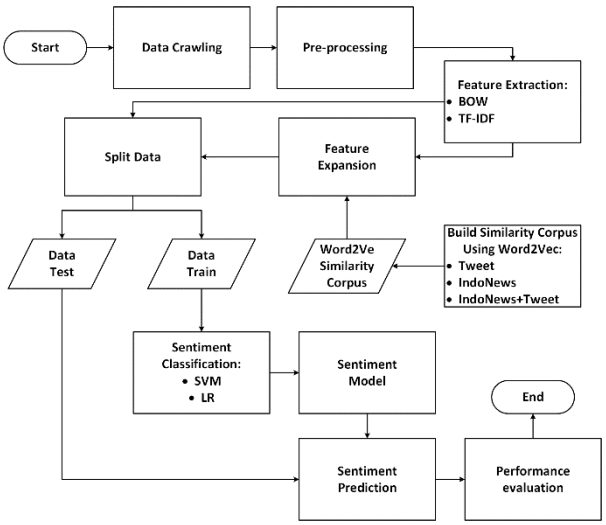

Figure 1. Sentiment Analysis System Using Feature Expansion Word2Vec

2.1 Data Crawling and Labeling

The data from Twitter is gathered by crawling. Crawling is the process of retrieving small and large data on the web that comes from a number of searched keywords and it can be stored [11]. This crawling technique is carried out using the help of the Twitter Application Programming Interface (Twitter API) which has been provided by Twitter's developer. Twitter API can be accessed through authentication

DOI: https://doi.org/10.29207/resti.v6i1.3525

requests which each request must be made by authorized Twitter users.

Classification is a process of building a model to describe or classify data where the data already has its own label or target class. In this study, the collected tweet data does not yet have a label or target class, therefore a manual labelling process is needed. We classify sentiment into two types of target class, as positive and negative. This labelling process for each class involves 3 to 4 other people helped with the definition of each label as follows, the negative label means that the tweet contains more harsh words with the intention of insulting, bringing down, and tending to reject the government, this sentence usually has a tone of anger, disappointment, and sadness, the positive label means that the tweet contains more appreciation, support, and seeing the benefits and positive sides of a public policy, this sentence usually has a tone of gratitude, pride, and happiness.

2.2 Data

We use 16597 tweets in Bahasa Indonesia that have been collected from Twitter. This data consists of several different keywords related to public policies in Indonesia such as, #uuciptakerja, #mositidakpercaya, #omnibuslaw, #dewanperampokrakyat, #uuite, #ppkm, #ruukpk, and #reformasidikorupsi. Table 1 provides samples of tweets.

Table 1. Sample of Tweets

Tweet	Sentiment
*Ekonomi **hancur**, pandemi makin **ganas**, hukum **hancur**, pemerintahan **amburadul**.. #PresidenTerburukDalamSejarah*	Negative
*Bersama UU Cipta Kerja menjadikan Indonesia menjadi **lebih baik** dan **mensejahterakan** para pekerja juga buruh https://t.co/qCJSODU2k4*	Positive

The amount of data distribution that has been labeled with sentiment can be seen in Figure 2.

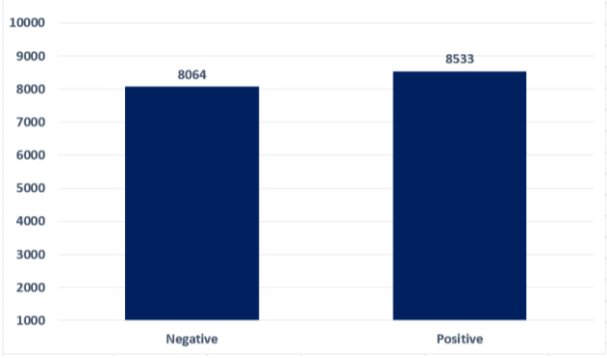

Figure 2. Data Distribution of Tweets

In the feature expansion process, this research used three corpuses. The first corpus (Tweet Corpus), contains around 10765 words taken from tweets in Bahasa Indonesia that were previously mentioned. The second corpus, contains around 178568 words taken from several mainstream media (IndoNews) such as Republika, CNN Indonesia, SindoNews, Kompas, Tempo, Detik.com, and Liputan6. The third corpus (IndoNews+Tweet Corpus), contains around 179636 words taken from the Tweets data combined with IndoNews data. The composition of the IndoNews data which is used to make a similarity corpus with the word embedding Word2Vec, can be seen in Table 2.

Table 2. IndoNews Data

Media	Quantity
Republika	53812
CNN Indonesia	29349
SindoNews	22401
Kompas	15055
Tempo	13702
Detik.com	7974
Liputan6	251
Total	142555

2.3 Data Preprocessing

To clear tweets from noise, misspelling, a series of preprocessing steps are required. The purpose of doing this process is to clean the data and improve the quality of the data when it is used to build a sentiment analysis model [12]. This process is assisted by Python libraries named String, Natural Language Toolkit (NLTK), and Sastrawi library.

There are six steps in this text preprocessing. **Case Folding** is a technique of converting sentences to lowercase. **Data Cleaning** is a process of removing all numbers, punctuation, URLs (http://, www…com), hashtags (#), and username tags (@username). **Word Normalization** is a technique of changing abbreviated words, slang words, typos, and informal words (*alay*) into formal words helped with a manually created word dictionary. **Stopwords Removal** is a technique of removing commonly used words and words that do not have special meanings such as pronouns, prepositions, and conjunctions [3]. **Stemming** is a technique of returning all words to their basic form by removing the prefix, infix, and suffix. For the stemming process, a special library for processing Indonesian text is used, namely Sastrawi. **Tokenization** is a technique of separating sentences into word for word called a token.

2.4 Bag-of-Words (BOW)

The Bag-of-Words (BOW) as a feature extraction technique is very general, simple, and flexible for sentences and documents. BOW is a text representation that describes the occurrence of words in a document. The reason for its name is because the model does not pay attention to word order or structure information in the document, but only to words that are known to appear in the document [13].

DOI: https://doi.org/10.29207/resti.v6i1.3525

2.5 Term Frequency – Inverse Document Frequency (TF-IDF)

The TF-IDF consists of TF which states the frequency of each word that appears in the document and the IDF value which results in a higher weight for words that rarely appear and a lower weight for words that occur frequently [14]. The frequency of words in a document indicates how important the word in a given document is.

2.6 Word2Vec

Word embedding is a distributed representation consisting of word properties in real number vectors that take syntactic features and semantic word relationships [15]. Word embedding is an attribute or feature learning technique in Natural Language Processing (NLP) where words that have the same, similar, or relative context, mathematically will be grouped into vector spaces [16]. One of model that can be used to produce word embedding with length *N* dimensions is Word2Vec [17]. Word2Vec which was developed by Mikolov *et al.* (2013) is one of the continuous learning tools to produce word embedding [18]. Depending on the number of predicted words, there are two types of Word2Vec models, the Continuous Bag-of-Words (CBOW) model and the Skip-Gram model [19]. Mikolov *et al.* in their research [20], has explained how word embedding Word2Vec works. Word2Vec model can be seen in Figure 3.

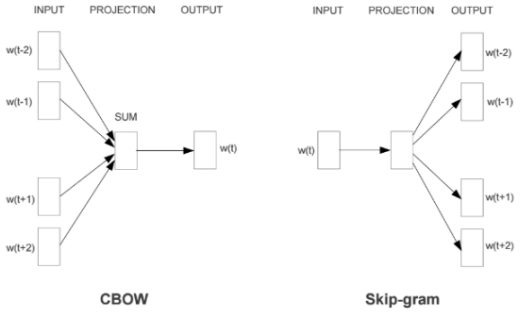

Figure 3. Architecture Model of Word2Vec

2.7 Feature Expansion

The basic idea of feature expansion is to reformulate a word by adding new words that have been previously stored using certain techniques [21]. As explained in the previous section, we use word embedding Word2Vec to solve the problem of vocabulary mismatch. In this study, the idea is to substitute the zero weight value in vector representation with a semantically similar word. Word2Vec is used to group words that are semantically similar into a single vector space, then generate a corpus contains a collection of similar words. One example of the results of word similarity can be seen in Table 3.

Table 3. Similarity Corpus Word of Twitter Top 10

Words	Twitter
Rank 1	akun
Rank 2	mikroblog
Rank 3	kicau
Rank 4	instagram
Rank 5	microblogging
Rank 6	linimasa
Rank 7	hastag
Rank 8	jejaring
Rank 9	cuit
Rank 10	fanpage

In the feature expansion process, this study uses a method similar to that done by Erwin *et al.* [10]. For example, the procedure for implementing feature expansion using similarity corpus IndoNews that was previously made using Word2Vec is illustrated on the feature vector representation the word "twitter" having zero value, but in one of the tweet document contains the sentence "… akun aktivis ancam bebas ekspresi". The word represented with a zero value then get checked on similarity corpus IndoNews, resulted in word "twitter" has lists of similar word as in Table 3. Therefore, the value of "twitter" which was originally zero is replaced with 1, as illustrated in Table 4. It applies to the BOW method, when using TF-IDF we replace it with the weight value. This feature expansion process is carried out on all words according to each corpus in the experiment section.

Table 4. Vector Representation on Document 1 Before and After Feature Expansion

Before	akun	aktivis	ancam	bebas	ekspresi	twitter
	1	1	1	1	1	0
After	akun	aktivis	ancam	bebas	ekspresi	twitter
	1	1	1	1	1	**1**

2.8 Logistic Regression (LR)

Logistic Regression is a well-known machine learning algorithm after Linear Regression. The working of LR is to describe and estimate the relationship between the dependent variable (usually known as the variable "Y") and one of the independent variables (variable "X") or a series of independent variables [22].

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad (1)$$

In equation (1), the left-hand side is called the log probability function (logit or log-odds), $1 - p$ is called odds which basically describes ratio of the probability of success to the probability of failure. LR algorithm requires the output to be a class and it is limited from 0 to 1, for example 0 (no) and 1 (yes).

2.9 Support Vector Machine (SVM)

Support Vector Machine is a classification method in supervised learning related to prediction algorithms, both in analyzing classification and regression cases. The standard SVM model was proposed in 1993 based on the theory developed by Vapnik and Chervonenkis

DOI: https://doi.org/10.29207/resti.v6i1.3525

[19]. SVM is one of the most popular classification models, widely applied for text processing [19]. SVM is considered to have the ability to compute large data sets, classifying non-linear data, *etc*. [23].

SVM algorithm tries to find the line that separates and categorizes between classification groups with the widest margin called a hyperplane [24]. The concept of SVM classification is to find the best hyperplane as the separator [25]. The best hyperplane is the line that has the largest distance or the largest margin between the hyperplane and the data point that is closest to the hyperplane line called support vector, which can be seen in Figure 4.

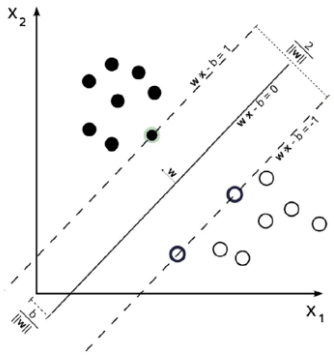

Figure 4. SVM Optimal Margin between Two Classes [26]

3. Results and Discussions

There were three scenarios using Logistic Regression and Support Vector Machine to obtain optimal performance values. The first scenario was to search for a model that will be used as a baseline using Bag-of-Words (BOW). The second scenario was an experiment to measure the effect of feature extraction using TF-IDF on the baseline model. Then, the third scenario was to measure the effect of feature expansion using Word2Vec with three different corpuses that have been created previously. In scenario 3, each corpus is carried out testing for Top 1, Top 5, Top 10, Top 15, and Top 20 features. Top 1 feature means that it will take a feature or word size of one word from each corpus that has the highest similarity value with a word in the corpus. Likewise, the same is done for the Top 5, Top 10, Top 15, and Top 20 features.

To get the optimal accuracy value in each scenario, searching for optimal parameters and the random ratio 70:30, 80:20, 90:10 on the training data and test data is carried out for each classification algorithm. In each classification system, the program execution is repeated 5 times and the average accuracy value is taken. The result of the first scenario is shown in Table 5.

Based on Table 5, the ratio of 90:10 on training data and test data gets the best accuracy for each two classification algorithms compared to other ratios, then for the next test scenario, the ratio of 90:10 on training data and test data will be used.

Table 5. Baseline for First Scenario

Classifier	Ratio	Accuracy (%)
Baseline (Logistic Regression)	70:30	76,76
	80:20	77,05
	90:10	**77,69**
Baseline (Support Vector Machine)	70:30	77,54
	80:20	77,89
	90:10	**78,21**

As stated before, the second scenario was an experiment to measure the effect of feature extraction using TF-IDF compared to the baseline model, this result is shown in Table 6.

Table 6. Second Scenario Using Baseline and TF-IDF

Model	Ratio	Accuracy (%)	
		Logistic Regression	Support Vector Machine
Baseline	90:10	77,69	78,21
Baseline + TF-IDF		77,98 (+0,38)	**78,80 (+0,75)**

The result of the second scenario has shown that the implementation of TF-IDF for feature extraction has a better performance with 78,80% accuracy score which increase up to 0,75% on SVM classifier, rather than using the baseline model.

Thereafter, the implementation of feature expansion Word2Vec with three different corpuses that have been created on the third scenario can be shown in Table 7-10. The Tweet corpus, IndoNews corpus, and IndoNews+Tweet corpus columns, respectively describe the accuracy from feature expansion experiments using Tweet similarity corpus, IndoNews similarity corpus, and combination of IndoNews and Tweet similarity corpus.

Table 7. Third Scenario Using Feature Expansion with Baseline on LR

Feature	Accuracy (%)		
	Tweet Corpus	IndoNews Corpus	IndoNews + Tweet Corpus
Top 1	76,53 (-1,49)	77,67 (-0,02)	77,57 (-0,16)
Top 5	75,28 (-3,10)	77,39 (-0,39)	76,96 (-0,93)
Top 10	74,77 (-3,75)	76,83 (-1,10)	77,25 (-0,56)
Top 15	74,66 (-3,89)	76,77 (-1,18)	77,01 (-0,87)
Top 20	74,43 (-4,19)	76,60 (-1,40)	76,94 (-0,96)

Table 7 shows the performance of feature expansion with Baseline on LR classifier. All accuracy decreases when using the entire corpus and the lowest accuracy was achieved by Top 20 with Tweet corpus.

Table 8 shows the performance of feature expansion with Baseline on SVM classifier. The highest increase of 0,07% was only achieved by Top 1 using IndoNews+Tweet corpus.

DOI: https://doi.org/10.29207/resti.v6i1.3525

Table 8. Third Scenario Using Feature Expansion with Baseline on SVM

Feature	Accuracy (%)		
	Tweet Corpus	IndoNews Corpus	IndoNews + Tweet Corpus
Top 1	77,42 (-1,01)	78,20 (-0,01)	**78,27 (+0,07)**
Top 5	76,89 (-1,69)	78,11 (-0,13)	77,93 (-0,36)
Top 10	76,11 (-2,69)	77,55 (-0,84)	77,42 (-1,01)
Top 15	75,65 (-3,27)	77,16 (-1,35)	77,14 (-1,36)
Top 20	75,83 (-3,04)	77,47 (-0,95)	77,20 (-1,29)

Table 9. Feature Expansion with TF-IDF on LR

Feature	Accuracy (%)		
	Tweet Corpus	IndoNews Corpus	IndoNews + Tweet Corpus
Top 1	78,31 **(+0,81)**	78,23 **(+0,70)**	78,25 (+0,73)
Top 5	78,18 (+0,64)	78,22 (+0,68)	78,20 (+0,67)
Top 10	78,11 (+0,54)	78,16 (+0,60)	78,28 **(+0,76)**
Top 15	78,07 (+0,50)	78,19 (+0,65)	78,14 (+0,59)
Top 20	78,11 (+0,54)	78,17 (+0,62)	78,14 (+0,59)

Unlike two previous table, based on Table 9, feature expansion with TF-IDF on LR classifier, all the accuracy score has increased. The highest accuracy achieved by Top 1 feature using Tweet corpus with 78,31% accuracy score, also this value increases up to 0,81% compared to the baseline.

Table 10. Feature Expansion with TF-IDF on SVM

Feature	Accuracy (%)		
	Tweet Corpus	IndoNews Corpus	IndoNews + Tweet Corpus
Top 1	78,86 **(+0,82)**	78,87 (+0,84)	78,82 (+0,78)
Top 5	78,73 (+0,67)	78,82 (+0,78)	78,88 (+0,85)
Top 10	78,73 (+0,67)	78,81 (+0,76)	78,89 (+0,87)
Top 15	78,77 (+0,72)	78,90 **(+0,89)**	78,99 **(+0,99)**
Top 20	78,75 (+0,69)	78,88 (+0,85)	78,94 (+0,93)

In Table 10 shows the accuracy score of feature expansion with TF-IDF on SVM classifier. All the accuracy score has also increased and the highest increase of 0,99% was achieved by Top 15 feature using combination of IndoNews and Tweet corpus.

In short of based on the three scenarios, a model using the Support Vector Machine (SVM) classifier shows a more stable and produces highest accuracy score, as can be seen in Figure 5.

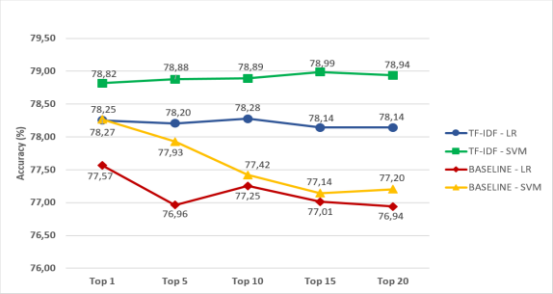

Figure 5. Accuracy Comparison Using Feature Expansion with IndoNews+Tweet Corpus on LR and SVM

The highest accuracy was achieved by Top 15 feature using IndoNews+Tweet corpus on SVM classifier with 78,99% accuracy score and this value increases up to 0,99% compared to the baseline model.

4. Conclusion

This research has been studied for Twitter sentiment analysis using feature expansion Word2Vec. To perform classification on the data set, two different machine learning algorithms such as Support Vector Machine (SVM) and Logistic Regression (LR) are used to compare the model. Based on the three scenarios that have been carried out, it can be concluded that feature expansion using Word2Vec method is proven to increase the accuracy values in this sentiment analysis system, specifically to complete missing vocabulary or reduce vocabulary mismatches in limited sentences or texts such as tweets on Twitter. The use of three different corpuses as a collection of word variation also showed various in increasing or decreasing in accuracy, the combination of IndoNews and Tweet (IndoNews+Tweet) is the similarity corpus with the highest increase in accuracy compared to the baseline. For further research, collecting more data set with another topic and various feature expansion using other word embedding techniques performed by different classification algorithms can be implemented.

Reference

[1] A. M. Kaplan and M. Haenlein, "The early bird catches the news: Nine things you should know about micro-blogging," *Bus. Horiz.*, vol. 54, no. 2, pp. 105–113, 2011, doi: 10.1016/j.bushor.2010.09.004.

[2] Ying Lin, "10 Twitter Statistics Every Marketer Should Know in 2021 [Infographic]," Jan. 25, 2021. https://id.oberlo.com/blog/twitter-statistics (accessed Mar. 01, 2021).

[3] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," *IEEE Access*, vol. 7, pp. 163677–163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[4] Z. Jianqiang, G. Xiaolin, and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," *IEEE Access*, vol. 6, pp. 23253–23260, 2018, doi: 10.1109/ACCESS.2017.2776930.

[5] Z. Jianqiang and G. Xiaolin, "Comparison research on text pre-processing methods on twitter sentiment analysis," *IEEE Access*, vol. 5, pp. 2870–2879, 2017, doi: 10.1109/ACCESS.2017.2672677.

[6] M. A. Fauzi, "Word2Vec model for sentiment analysis of product reviews in Indonesian language," *Int. J. Electr. Comput. Eng.*, vol. 9, no. 1, p. 525, 2019, doi: 10.11591/ijece.v9i1.pp525-530.

[7] F. W. Kurniawan and W. Maharani, "Indonesian Twitter Sentiment Analysis Using Word2Vec," *2020 Int. Conf. Data Sci. Its Appl. ICoDSA 2020*, pp. 31–36, 2020, doi: 10.1109/ICoDSA50139.2020.9212906.

[8] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion for sentiment analysis in twitter," 2018, doi: 10.1109/EECSI.2018.8752851.

[9] S. P. Sheela, "Sentiment Analysis and Prediction of Online Reviews with Empty Ratings," *Int. J. Appl. Eng. Res.*, vol. 13, no. 14, 2018.

[10] E. B. Setiawan, D. H. Widyantoro, and K. Surendro, "Feature expansion using word embedding for tweet topic

DOI: https://doi.org/10.29207/resti.v6i1.3525

classification," 2017, doi: 10.1109/TSSA.2016.7871085.
[11] J. Eka Sembodo, E. Budi Setiawan, and Z. Abdurahman Baizal, "Data Crawling Otomatis pada Twitter," 2016, doi: 10.21108/indosc.2016.111.
[12] R. Ferdiana, F. Jatmiko, D. D. Purwanti, A. S. T. Ayu, and W. F. Dicka, "Dataset Indonesia untuk Analisis Sentimen," *J. Nas. Tek. Elektro dan Teknol. Inf.*, vol. 8, no. 4, p. 334, 2019, doi: 10.22146/jnteti.v8i4.533.
[13] Y. Goldberg, *Neural network methods for natural language processing (Synthesis Lectures on Human Language Technologies)*, vol. 10, no. April. 2017.
[14] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int. J. Comput. Appl.*, vol. 181, no. 1, 2018, doi: 10.5120/ijca2018917395.
[15] D. Dessì, R. Helaoui, V. Kumar, D. R. Recupero, and D. Riboni, "TF-IDF vs word embeddings for morbidity identification in clinical notes: An initial study," *CEUR Workshop Proc.*, vol. 2596, pp. 1–12, 2020.
[16] M. S. R. Hitesh, V. Vaibhav, Y. J. A. Kalki, S. H. Kamtam, and S. Kumari, "Real-time sentiment analysis of 2019 election tweets using word2vec and random forest model," *2019 2nd Int. Conf. Intell. Commun. Comput. Tech. ICCT 2019*, pp. 146–151, 2019, doi: 10.1109/ICCT46177.2019.8969049.
[17] H. Imaduddin, Widyawan, and S. Fauziati, "Word embedding comparison for Indonesian language sentiment analysis," *Proceeding - 2019 Int. Conf. Artif. Intell. Inf. Technol. ICAIIT 2019*, pp. 426–430, 2019, doi: 10.1109/ICAIIT.2019.8834536.
[18] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," 2013.
[19] M. M. Truşcă, "Efficiency of SVM classifier with Word2Vec and Doc2Vec models," *Proc. Int. Conf. Appl. Stat.*, vol. 1, no. 1, pp. 496–503, 2020, doi: 10.2478/icas-2019-0043.
[20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, "Distributed representations ofwords and phrases and their compositionality," 2013.
[21] J. Tang, Y. Wang, K. Zheng, and Q. Mei, "End-to-end learning for short text expansion," 2017, doi: 10.1145/3097983.3098166.
[22] C. Y. J. Peng, K. L. Lee, and G. M. Ingersoll, "An introduction to logistic regression analysis and reporting," *J. Educ. Res.*, vol. 96, no. 1, 2002, doi: 10.1080/00220670209598786.
[23] L. Sravani, A. S. Reddy, and S. Thara, "A Comparison Study of Word Embedding for Detecting Named Entities of Code-Mixed Data in Indian Language," *2018 Int. Conf. Adv. Comput. Commun. Informatics, ICACCI 2018*, pp. 2375–2381, 2018, doi: 10.1109/ICACCI.2018.8554918.
[24] S. Chandra Satapathy and A. Joshi, *Smart Innovation, Systems and Technologies 107 Information and Communication Technology for Intelligent Systems*, vol. 2. 2018.
[25] U. Rofiqoh, R. S. Perdana, and M. A. Fauzi, "Analisis Sentimen Tingkat Kepuasan Pengguna Penyedia Layanan Telekomunikasi Seluler Indonesia Pada Twitter Dengan Metode Support Vector Machine dan Lexion Based Feature," *J. Pengemb. Teknol. Inf. dan Ilmu Komput. Univ. Brawijaya*, vol. 1, no. 12, pp. 1725–1732, 2017, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/628.
[26] W. Z. Lu and D. Wang, "Learning machines: Rationale and application in ground-level ozone prediction," *Appl. Soft Comput. J.*, vol. 24, 2014, doi: 10.1016/j.asoc.2014.07.008.

DOI: https://doi.org/10.29207/resti.v6i1.3525