Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

Optimal Number Data Trains in Hoax News Detection of Indonesian using SVM and Word2Vec

Muhammad Sulthon Asramanggala, Sri Suryani Prasetyowati, Yuliant Sibaroni*

Informatics, Faculty of Informatics, Telkom University, Bandung, Indonesia
Email: [1]sulthonasr@student.telkomuniversity.ac.id, [2]srisuryani@telkomuniversity.ac.id, [3]yuliant@telkomuniversity.ac.id
Correspondence Author Email: yuliant@telkomuniversity.ac.id
Submitted: **26/05/2023**; Accepted: **27/06/2023**; Published**: 29/06/2023**

Abstract−Along with the development of the era of technological development also has an increase. Information dissemination occurs very quickly on social media, especially Twitter. On Twitter, only some news circulating is necessarily accurate information. Lots of information that is spread is hoax news that irresponsible individuals apply. In this research, the author will build a system to determine the optimal amount of data trained in the hoax news classification process. In this study, the authors will use the support vector machine and word2vec algorithms to classify hoax and non-hoax news on the system to be created. In this study, five experiments were carried out with the number of train data used as many as 5000, 10000, 15000, 20000, and 25000. 5000 data train results in an accuracy of 77.28%, 10000 data train produce an accuracy of 79.68%, data 15,000 trains produce an accuracy of 79.892%, 20,000 data trains produce an accuracy of 80,416%, and 25,000 data trains produce an accuracy of 81,184%, by using a combination of unigram with token full token selection. This research aims to build a hoax detection system that can determine the optimal amount of data training to use. Also, this research is used to see the performance of the Support Vector Machine algorithm with Word2Vec in detecting hoax news.

Keywords: Hoax, Classification, Support Vector Machine, Word2Vec, Twitter

1. INTRODUCTION

The rapid development of technology lately has made it easier for people to get information quickly. But not infrequently, the information received is hoax news. Hoax is a term for news or information in the form of fake news or slander. Hoaxes are made to lead readers' opinions to match the information being reported [1]. The most important media outlets for spreading fake news are Radio 1.20%, print media 5 %, and television 8.70%. The most used channel to spread fraud is social media (Facebook, Twitter, Instagram, and Path), the most used, 92.40%, especially Twitter[2]. Nowadays, Twitter is widely used by Indonesians because it is easy to find information and easy-access. The Ministry of Communication and Information of the Republic of Indonesia stated that in the last three years, there have been 9,546 cases of hoaxes spreading in Indonesia. With so many instances of spreading hoaxes, especially those on social media Twitter, research will be needed to detect hoaxes.

Research on the detection of hoaxes on social media has been carried out by several researchers, namely Munirul et al. in 2020 [3], I. Kencana Wintang et al in 2020 [4], Ismayanti et al in 2021 [5], B.P Nayoga et al in 2021 [6], and P. N Anggreyani et al in 2021 [7]. SVM and TF-IDF classifiers are used in research [3], feed-forward and back-propagation classifiers with TF-IDF and Word2Vec vectorization are used in research [4], Logistic Regression classification classifier, Support Vector Machine (SVM), Random Forest with the Word2Vec expansion feature is used in research [5], the support Vector Machine (SVM) and Naive Bayes classifier is used in research [6], and the LSTM and CNN classifiers are used in research [7].

In research [3], 120 data were used, 30 test data were used, and the results were obtained using the Support Vector Machine method with TF-IDF weighting to get an accuracy of 60%. In research [4], the results of feed-forward and back-propagation using TF-IDF vectorization increased the highest performance compared to Word2Vec with an accuracy of 78.76%. TF-IDF worked longer than Word2Vec, but the performance results showed that TF-IDF provided the highest accuracy. In research [5], 21,588 data trains were used, and 5396 test data obtained using the Support Vector Machine method with Word2Vec yielded the highest accuracy value of 87.34%. In research [6], 1000 datasets were used, the SVM model gave values for accuracy of 92.6%, precision of 92.48%, recall of 92.59%, and f1 macro of 92.45%, and after using the dropout, the SVM model experienced a slight increase in the value of accuracy, precision, recall, and f1 macro. Finally, research [7] showed that LSTM-CNN could achieve 79.71% accuracy by using 16 units in a layer by combining dropouts and regularizers. Using LSTM-CNN with Word2Vec can process large amounts of data to detect hoaxes.Support Vector Machine (SVM) is a method in machine learning that works based on the principle of structural risk Minimization (SRM), which aims to find the best hyperplane that separates the two classes in the input [8]. The SVM method can classify problems linearly, but currently, SVM has developed to solve issues non-linearly by finding the optimal hyperplane [9]. In other words, the SVM method is very suitable for detecting hoax news, separating hoax and non-hoax news.

In some research conducted to detect hoax news, one method plays an important role, namely word2vec. Word2vec is a word embedding algorithm that maps every word in text into vectors. The word2vec algorithm was created by Mikolov et al. in 2013 [10]. The working principle of word2vec is to predict the meaning of words according to the chances of their appearance and to be able to do associations to determine the relationship between words and other words [11]. The advantage of Word2Vec is that it can represent the contextual similarity of two words in the resulting vector. Word2vec is a feedforward neural network model consisting of a hidden and fully connected

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

layer [10]. Word2Vec has three parameters that influence the learning model process: architecture, evaluation method, and dimensions. Each type of the three parameters that Word2Vec influences the performance of deep learning accuracy [12].

Based on the research above, it can be concluded that there is no optimal data train measure for hoax classification. Therefore, in this research, research will be carried out to determine the optimal amount of data trained in hoax classification.

2. RESEARCH METHODOLOGY

2.1 Research Stages

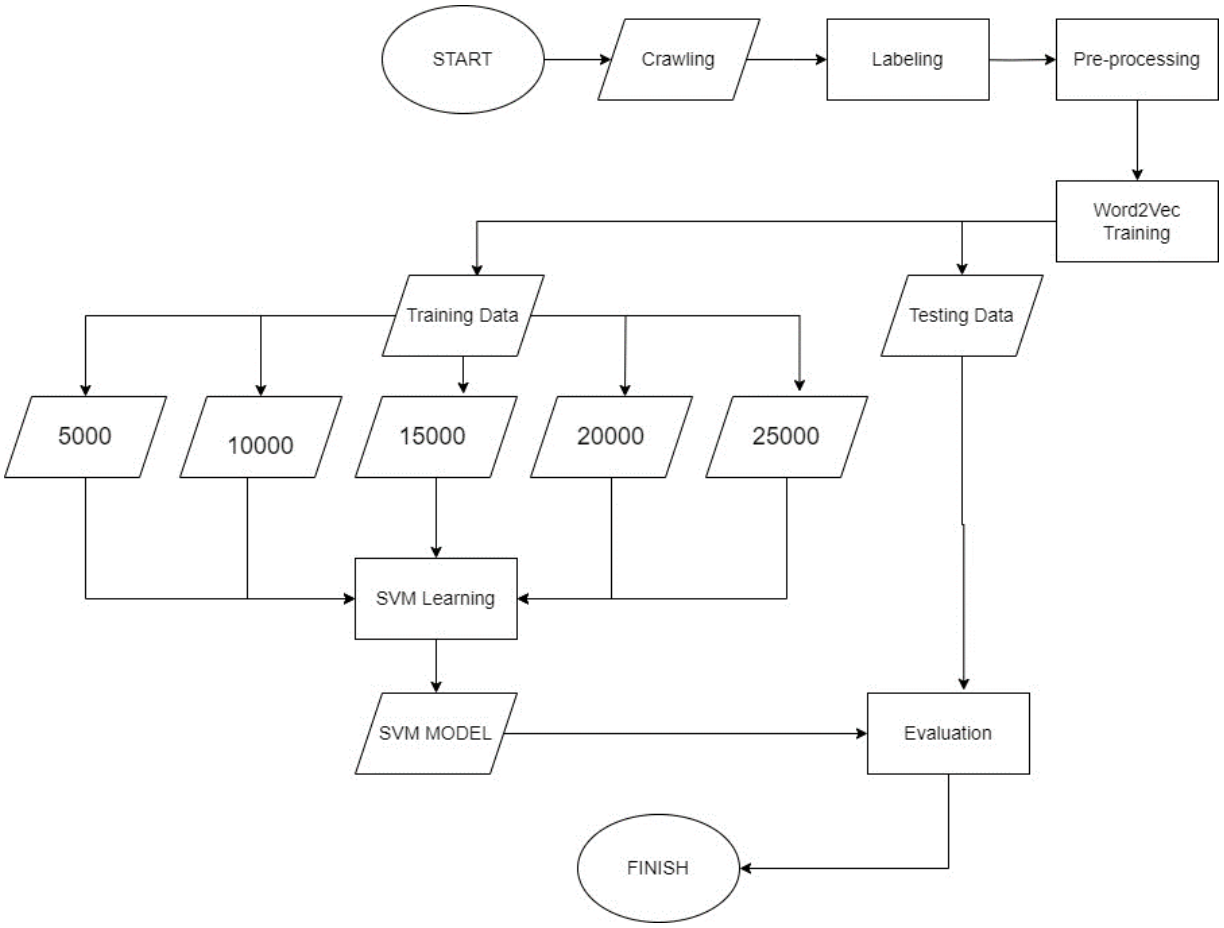

Figure 1. Research Stages

The work process in this research begins with retrieving the Indonesian news media dataset from Twitter users, then labels the data to separate the data into 2 classes: hoaxes and non-hoaxes. Furthermore, preprocessing will be carried out because the collected datasets are still raw data. After the data is clean, the training process will be carried out using Word2Vec before separating the training data and testing data. In the data training process, 5 schemes will be used with different amounts of training data, that is: 5000,10000,15000,20000 and 25000. Then the SVM learning process will be carried out to train the most optimal SVM model. After the modeling, an evaluation and testing process will be conducted to obtain a good and optimal amount of training data.

2.2 Dataset

The dataset used in this study is taken from social media Twitter. The dataset will be used in the form of Twitter user tweets about Indonesian press. Each dataset usually consists of several variables: user id, username, time, tweets from the user, and the number of followers. The amount of data obtained during crawling is 25,000 tweet data taken in the range of January 2022 - February 2023.

2.3 Labeling

Labeling is the process of labeling the data to be studied by the system. The data will be stored in a CSV file format with the label "valid" for fake news and "hoax" for fake news. The following is an example of the data labeling process. The following is an example of the labeling process:

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

Table 1. Labeling

Label	Text	Description
Valid	Satuan Brimob Polda Metro Jaya bersama Relawan HIMA IPB Bogor ikut serta dalam rangka Penyaluran Bantuan Kemanusiaan bagi Korban Bencana Alam Gempa Bumi di wilayah Cianjur, Jawa Barat.	Verified news is true
Valid	samator group turut dukung program vaksinasi covid nasional	Verified news is true
Hoax	Telah terjadi gempa bumi dengan titik episetrum 11 km dari pantai pangandaran dengan kekuatan 5 SR, diklaim lebih parah dari gempa Cianjur dan akan berpotensi menimbulkan tsunami di sekitar pantai selatan jawa.	Verified news is Fake
hoax	Telah terjadi gempa bumi dengan titik episetrum 11 km dari pantai pangandaran dengan kekuatan 5 SR, diklaim lebih parah dari gempa Cianjur dan akan berpotensi menimbulkan tsunami di sekitar pantai selatan jawa.	Verified news is fake

2.4 Preprocessing

Data preprocessing ensures the data is ready for use by machine learning algorithms or other data analysis, resulting in more accurate models and more meaningful results. There are several stages carried out in the preprocessing process:
a. *Remove emoji* is done to remove emoji characters contained in tweets.
b. *Case Folding* is done to change all uppercase letters to lowercase.
c. *Normalization* is done to change the abbreviation words according to the normalization word dictionary.
d. *Clean tweet* done to remove punctuation, hashtags, and links that do not affect research results.
e. *Stemming* is done to change the word into its base word, used sastrawi library in Python for this process.
f. *Tokenization* is done to change the text into smaller token tokens.

2.5 N-Gram

The language used daily is not a collection of individual words but a collection of sequences of individual words that have meaning [14]. N-Gram is a combination of unigram and bigram, so the features obtained from both will be used as one [15]. For example, the sentence "Gempa Cianjur adalah bencana nasional" can be broken down into:
a. Unigram : gempa, cianjur, adalah, bencana, nasional
b. Bigram : gempa cianjur, cianjur adalah, adalah bencana, bencana nasional

2.6 Word2Vec Train

After pre-processing the dataset, the training will be carried out using Word2Vec. Word2Vec aims to find hidden interactions in a word and then visualize the word in a vector form. Wod2vec relies on locale information from the language. The surrounding words influence the learned semantics of a particular word. This model can study linguistic patterns as linear relationships between word vectors [10].

Word2Vec has three parameters that influence the learning model process: architecture, evaluation method, and dimensions. Each type of the three parameters that Word2Vec influences the performance of deep learning accuracy[12]. The Word2Vec model will go through three stages: vocabulary builder, context builder, and neural network (skip-gram architecture) [16].

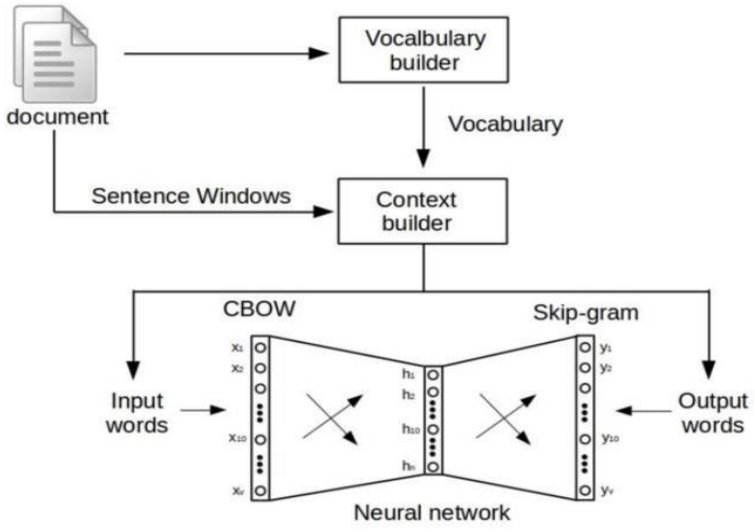

Figure 2. Arshitecture Word2Vec

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

In this research, the writer will use one of the architectures in Word2Vec, the skip-gram, which is used to predict the context (output) around the current word (input), bounded by the window. In one sentence the skip-gram model can predict the words that are around the word [17]. Here is an illustration :

Table 2. Word2Vec Train

Text	Word Pair	Caption
gempa cianjur adalah bencana nasional	(gempa, cianjur), (gempa, adalah)	Input : gempa Target : cianjur, adalah
gempa cianjur adalah bencana nasional	(cianjur, gempa), (cianjur, adalah), (cianjur, bencana)	Input : cianjur Target : gempa, adalah, bencana
gempa cianjur adalah bencana nasional	(adalah , gempa), (adalah , cianjur), (adalah , bencana), (adalah , nasional)	Input : adalah Target : gempa, cianjur, bencana, nasional
gempa cianjur adalah bencana nasional	(bencana , cianjur), (bencana , adalah), (bencana , nasional)	Input : bencana Target : cianjur, adalah, nasional
gempa cianjur adalah bencana nasional	(nasional, bencana), (nasional, adalah)	Input : nasional Target : bencana, adalah

2.7 SVM Model

Support Vector Machine is a machine learning method that works according to the Structural Risk Minimization (SRM) principle, which aims to find the best hyperplane that separates two classes in the input space [18]. The hyperplane is made in such a way as to separate the nearest data point (support vector) from each class [19]. The main advantage of this approach is that it can work in very high-dimensional areas.

In simple terms, SVM is trained to create a model that will be used for classification and prediction [20]. The algorithm goes through a training process by creating a hyperplane line by taking the vector closest to the margins of the two labels. This support vector is also used to represent data sets for predictions. The following is an illustration of the Support Vector Machine algorithm architecture.:

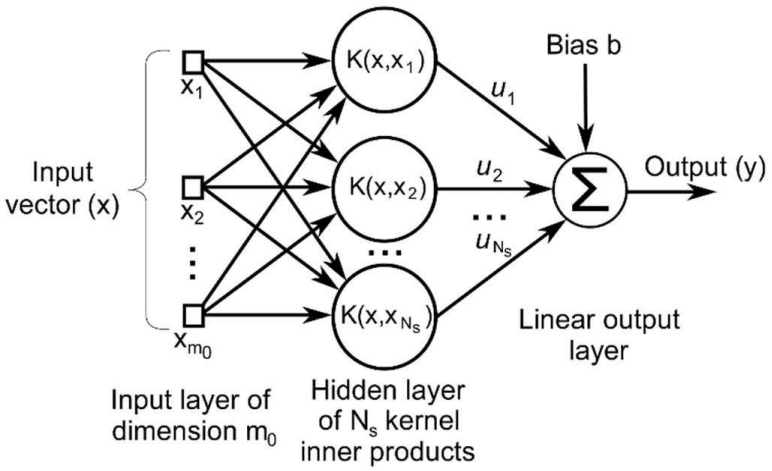

Figure 3. Achitecture SVM

Support Vector Machine (SVM) is a non-parametric method usually used in data classification and image processing. The level of accuracy in this method is taken from the parameters and the kernel; the user can determine the parameters, and each parameter will have a different impact on the kernel. The Support Vector Machine (SVM) method has two ways to solve it: linear and non-linear [20]. The level of accuracy in this method is taken from the parameters and the kernel, the following is the kernel that is commonly used in Support Vector Machine modeling:

a. Linear Kernel which functions for linear data classification, can be calculated with the following equation:

$$f(x) = w * x + b \quad (1)$$

b. The Kernel Polynomial which is used when the data is not linearly separated, can be calculated with the following equation:

$$f(x) = (gamma * (x * x') + coef0)^d \quad (2)$$

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

c. Kernel RBF (Radial Basic Function) is used to solve data classification problems that cannot be separated linearly, which can be calculated by the following equation::

$$f(x) = exp(-gamma* || x - x' ||^2) \quad (3)$$

2.8 Evaluation

After the system has been run, at this stage, an evaluation will be carried out to get the most optimal amount of train data. In this final project, the writer will use the k-fold cross-validation method. K-fold Cross Validation is a validation technique to determine the best model [21]. The goal of Kfold Cross Validation is to assess how statistical analysis results will generalize to independent data sets. This technique is mainly used to make model predictions and ensure how accurate a predictive model is when it is run in practice because the data will be divided into train data and test data [22].

Table 3. 5 Fold Cross Validation

Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Table 4 shows how the k-fold cross-validation works. In the first iteration, make the first partition the testing data and the other partition the training data. The second iteration makes the second partition testing data and the other partitions' training data. It will continue the iteration process until all data has been divided. These experiments' results will record the model performance evaluation value using the confusion matrix.

3. RESULTS AND DISCUSSION

3.1 Extraction Features of Word2Vec

In this study, the authors used Word2Vec as an extraction feature. Parameters used in Word2Vec vector size = 100, window = 8, min count = 10, sg = 1. Examples of extraction results can be seen in Table 5 below.

Table 4. Word2Vec Result

Text	Word2Vec Vector
bekas	0.9417843818664551
kereta	0.9226388931274414

In the table above, it can be seen that the vector weight for the word used is 0.94178, and the vector weight for the word train is 0.92226.

3.2 Testing

Test results in this study were carried out using a dataset of 27,000 taken from social media Twitter using the Support Vector Machine method with Word2Vec weighting. Six cases of testing were carried out in this study. Each test in the study was carried out by cross-validation with a fold value of 5 and will display the accuracy value for each fold.

3.2.1 Testing with 5000 train data

In the first test using 5000 data train aims to see the results of the highest accuracy. In this test, the N-Gram method is added to predict the word that will appear, then a token selection is added to see changes in each case. In Table 6 it can be seen that the highest average accuracy is obtained by unigram with the full token model producing an average of 77.28%

Table 5. Accuracy Data Train 5000

Fold	Unigram				Bigram			
	3 word	6 word	10 word	full token	3 word	6 word	10 word	full token
1	52,11%	52,78%	50,13%	75,4%	52,33%	50,46%	50,4%	73,9%
2	51,5%	51,54%	48,63%	76,5%	50,48%	59,22%	54,09%	74,9%
3	49,49%	49,53%	48,08%	79,2%	48,27%	47,67%	51,91%	76,6%
4	47,16%	47,52%	50,54%	77,6%	49,83%	47,52%	53%	76,3%
5	50,83%	50,92%	49,45%	77,7%	51,5%	51,85%	52,73%	75,7%

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

Average	50,218%	50,458%	49,366%	77,28%	50,482%	51,344%	52,426%	75,48%

3.2.2 Testing with 10000 train data

The first test using 10,000 data trains aims to see the results of the highest accuracy. In this test, the N-Gram method is added to predict the word that will appear, then a token selection is added to see changes in each case. In Table 7 it can be seen that the highest average accuracy is obtained by unigram with the full token model producing an average of 79.68%

Table 6. Accuracy Data Train 10000

Fold	Unigram				Bigram			
	3 word	6 word	10 word	full	3 word	6 word	10 word	full
1	74,93%	74,62%	79,26%	78,9%	73,17%	74,27%	78,11%	78,75%
2	73,7%	75,29%	78,57%	79,4%	74,24%	74,81%	79,14%	78,9%
3	75,25%	76,32%	79,49%	80,1%	74,29%	74,81%	78,11%	79,25%
4	75,73%	77,43%	81,68%	81,6%	74,88%	76,05%	79,72%	80,1%
5	73,69%	75,15%	78,45%	78,4%	72,19%	73,22%	76,49%	77,4%
Average	74,66%	75,762%	79,49%	79,68%	73,754%	74,632%	78,314%	78,88%

3.2.3 Testing with 15,000 train data

The first test using a data train of 15,000 aims to see the results of the highest accuracy. In this test, the N-Gram method is added to predict the word that will appear, then a token selection is added to see changes in each case. In Table 8 it can be seen that the highest average accuracy is obtained by unigram with the full token model producing an average of 79.892%

Table 7. Accuracy Data Train 15,000

Fold	Unigram				Bigram			
	3 word	6 word	10 word	full	3 word	6 word	10 word	full
1	49,4%	50,28%	48,06%	80,9%	50,01%	48,85%	49,75%	80,03%
2	49,91%	49,38%	47,01%	79,7%	49,98%	50,61%	51,61%	79,53%
3	49,24%	49,38%	51,61%	78,5%	48,74%	49,14%	52,41%	78,3%
4	48,81%	50,85%	49,75%	79,63%	50,61%	50,95%	51,85%	80,03%
5	49,06%	50,59%	50,2%	80,73%	48,81%	50,68%	50,36%	79,76%
Average	49,284%	50,096%	49,326%	79,892%	49,63%	50,046%	51,196%	79,53%

3.2.4 Testing with 20,000 train data

In the first test using 20,000 data train aims to see the results of the highest accuracy. In this test, the N-Gram method is added to predict the word that will appear, then a token selection is added to see changes in each case. In Table 9 it can be seen that the highest average accuracy is obtained by unigram with the full token model producing an average of 80.416%

Table 8. Accuracy Data Train 20,000

Fold	Unigram				Bigram			
	3 word	6 word	10 word	full	3 word	6 word	10 word	full
1	50,42%	49,58%	50,43%	80,12%	49,92%	49,09%	50,1%	79,37%
2	48,65%	49,01%	49,46%	79,67%	49,31%	49,17%	49,62%	79,35%
3	50,01%	49,47%	48,73%	80,25%	51,49%	51,25%	49,48%	80,27%
4	51,57%	51,05%	51,5%	81,12%	51,65%	51,71%	51,15%	80,52%
5	51,44%	51,25%	51,47%	80,92%	51,49%	51,61%	52,33%	80,92%
Average	50,418%	50,072%	50,318%	80,416%	50,772%	50,566%	50,536%	80,086%

3.2.5 Testing with 25,000 train data

The first test using 25,000 data trains aims to see the results of the highest accuracy. In this test, the N-Gram method is added to predict the word that will appear, then a token selection is added to see changes in each case. In Table 10 it can be seen that the highest average accuracy is obtained by unigram with the full token model producing an average of 81.184%.

Table 9. Accuracy Data Train 25,000

Fold	Unigram				Bigram			
	3 word	6 word	10 word	full	3 word	6 word	10 word	full
1	49,84%	49,82%	49,63%	80,36%	50,22%	50,06%	50,34%	80%

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

2	49,67%	48,92%	49,78%	81,86%	51,08%	50,54%	50%	81,08%
3	49,92%	50,25%	50,08%	80,82%	49,79%	50,54%	51,57%	79,94%
4	49,52%	50,46%	51,74%	80,86%	50,6%	51,49%	50,83%	80,6%
5	49,06%	49,35%	49,8%	82,02%	49,03%	49,56%	51,07%	81,6%
Average	49,602%	49,76%	50,206%	81,184%	50,144%	50,438%	50,762%	80,644%

3.3 Analysis of Test Results

Based on the tests carried out with the same model but with different amounts of data trained, it produces a fairly good accuracy for each amount of data trained when using the full token model unigram and token. When the test is carried out with a data train of 10,000 accuracies in each case, there is an accuracy value that is relatively not much different. When the data train is used, 5000, 15,000, 20,000, and 2,500 accuracies are produced when using a 3-word, 6-word token model. and 10 words both in unigram and bigram, the accuracy can be in the range of 49% - 50%. The following is a table and graph of the average accuracy comparison:

Table 10. Comparasion Accuracy

Data train	Unigram				Bigram			
	3 word	6 word	10 word	full	3 word	6 word	10 word	full token
5000	50,21%	50,45%	49,36%	77,28%	50,48%	51,34%	52,42%	75,48%
10000	74,66%	75,76%	79,49%	79,68%	73,75%	74,63%	78,31%	78,88%
15000	49,28%	50,09%	49,32%	79,89%	49,63%	50,04%	51,19%	79,53%
20000	49,60%	50,07%	50,31%	80,41%	50,77%	50,56%	50,53%	80,08%
25000	49,60%	49,76%	50,20%	81,18%	50,14%	50,43%	50,76%	80,64%
Max	81,18 %							

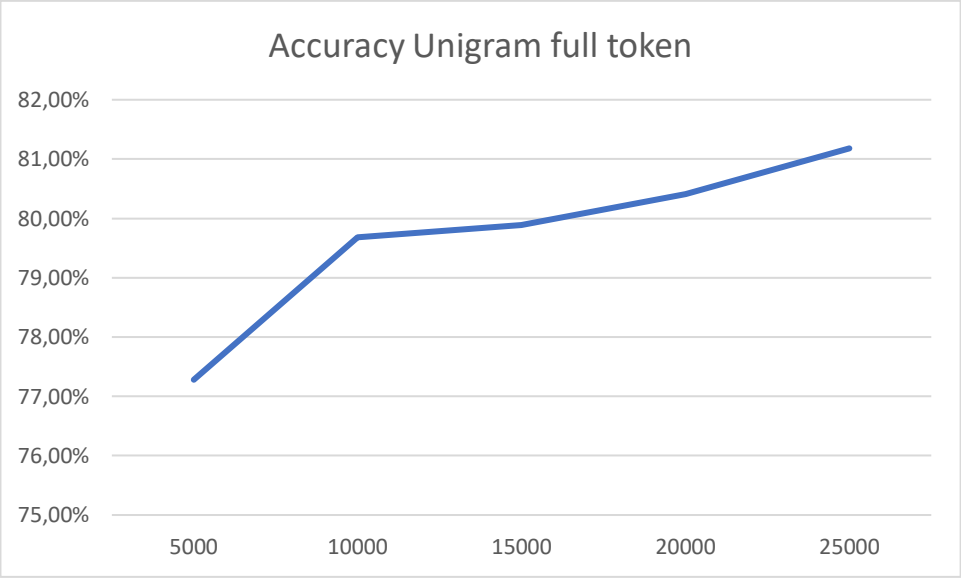

Figure 4. Graph of SVM accuracy for various data train sizes using the full token unigram

It can be seen from table 11 and Figure 4 that there is a phenomenon that when the amount of train data is large, the trend accuracy is higher. From Figure 4, it can also be concluded that the optimal data train used in this study is 25000 unigrams with full token.

3. CONCLUSION

This study aims to determine the optimal number of data trains in hoax classification using the SVM and Word2Vec classifiers using datasets taken from social media Twitter. This research also determines the optimal number of token selections. Token selection is a tokenization method where the model will use all the tokens in the dataset with the minimum number of tokens according to those set in the model. As with the 3-word token selection example, the model will use all datasets with at least 3 tokens in each data and the full token; the model will use all existing data without a minimum token. In this study, the most optimal token selection is the full token which produces an accuracy of 81.18% for unigram and 80.64% for bigram. From this, it can be seen that using unigram is better than bigram. To achieve optimal SVM classifier results with accuracy above 80%, at least more than 20,000 training data is required. A sufficient amount of training data allows the model to learn more complex patterns and produce more accurate decisions. In this case, the more training data used, the better the performance of the SVM model. Suggestions for

Building of Informatics, Technology and Science (BITS)
Volume 5, No 1, Juni 2023 Page: 21−28
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v5i1.3516

further research are to use extraction features and other classifiers to improve accuracy when using 3-word, 6-word, and 10-word selection tokens.

REFERENCES

[1] A. Afriza and J. Adisantoso, "Metode Klasifikasi Rocchio untuk Analisis Hoax Rocchio Classification Method for Hoax Analysis," *J. Ilmu Komput. Agri-Informatika*, vol. 5, no. 1, pp. 1–10, 2018, [Online]. Available: http://journal.ipb.ac.id/index.php/jika

[2] R. N. Rahayu and Sensusiyati, "Analisis Berita Hoax Covid - 19 Di Media Sosial Di Indonesia," *J. Ekon. Sos. Hum.*, vol. 1, no. 9, p. 63, 2020.

[3] Munirul, Ula, M. M. Alvanof, and R. Triandi, "Analisa Dan Deteksi Konten Hoax Pada Media Berita," *J. Teknol. Terap. Sains 4.0 Univ. Malikussaleh*, vol. 1, p. 2, 2020.

[4] I. Kencana Wintang, Crisanadenta; Setiawan Budi, Erwin; Kurniawan, "JURNAL RESTI Hoax Detection on Twitter using Feed-forward and Back-propagation," *RESTI J. (System Eng. Inf. Technol.*, vol. 4, no. 10, pp. 655–663, 2020.

[5] F. Ismayanti and E. B. Setiawan, "Deteksi Konten Hoax Berbahasa Indonesia Di Twitter Menggunakan Fitur Ekspansi Dengan Word2vec," *eProceedings …*, vol. 8, no. 5, pp. 10288–10300, 2021, [Online]. Available: https://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15697%0Ahttps://openlibrarypublications.telkomuniversity.ac.id/index.php/engineering/article/view/15697/15410

[6] B. P. Nayoga, R. Adipradana, R. Suryadi, and D. Suhartono, "Hoax Analyzer for Indonesian News Using Deep Learning Models," *Procedia Comput. Sci.*, vol. 179, no. 2020, pp. 704–712, 2021, doi: 10.1016/j.procs.2021.01.059.

[7] P. N. Anggreyani and W. Maharani, "Hoax Detection Tweets of the COVID-19 on Twitter Using LSTM- CNN with Word2Vec," vol. 6, pp. 2432–2437, 2022, doi: 10.30865/mib.v6i4.4564.

[8] D. A. Pisner and D. M. Schnyer, "Support vector machine," *Mach. Learn. Methods Appl. to Brain Disord.*, pp. 101–121, 2019, doi: 10.1016/B978-0-12-815739-8.00006-7.

[9] I. M. Mubaroq and E. B. Setiawan, "The Effect of Information Gain Feature Selection for Hoax Identification in Twitter Using Classification Method Support Vector Machine," *Indones. J. …*, vol. 5, no. September, pp. 107–118, 2020, doi: 10.21108/indojc.2020.5.2.499.

[10] A. Nurdin, B. Anggo Seno Aji, A. Bustamin, and Z. Abidin, "Perbandingan Kinerja Word Embedding Word2Vec, Glove, Dan Fasttext Pada Klasifikasi Teks," *J. Tekno Kompak*, vol. 14, no. 2, p. 74, 2020, doi: 10.33365/jtk.v14i2.732.

[11] D. E. Latumaerissa, "Studi Ekstraksi Fitur Data Teks Rencana Pelaksanaan Pembelajaran Memanfaatkan Model Word2Vec," *J. Linguist. Komputasional*, vol. 4, no. 2, p. 34, 2021, doi: 10.26418/jlk.v4i2.54.

[12] D. I. Af'idah, Dairoh, S. F. Handayani, and R. W. Pratiwi, "Pengaruh Parameter Word2Vec terhadap Performa Deep Learning pada Klasifikasi Sentimen," *J. Inform. Jurunal Pengemb. IT*, vol. 6, no. 3, pp. 156–161, 2021.

[13] S. Khomsah and Agus Sasmito Aribowo, "Model Text-Preprocessing Komentar Youtube Dalam Bahasa Indonesia," *Rekayasa Sist. dan Teknol. Informasi, RESTI*, vol. 4, no. 10, pp. 648–654, 2020.

[14] I. Fahrur Rozi, A. Taufika Firdausi, and K. Islamiyah, "Analisis Sentimen Pada Twitter Mengenai Pasca Bencana Menggunakan Metode Naïve Bayes Dengan Fitur N-Gram," *J. Inform. Polinema*, vol. 6, no. 2, pp. 33–39, 2020, doi: 10.33795/jip.v6i2.316.

[15] M. Hakiem, M. A. Fauzi, and Indriati, "Klasifikasi Ujaran Kebencian pada Twitter Menggunakan Metode Naïve Bayes Berbasis N-Gram Dengan Seleksi Fitur Information Gain," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2443–2451, 2019, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/4682

[16] J. Patihullah and E. Winarko, "Hate Speech Detection for Indonesia Tweets Using Word Embedding And Gated Recurrent Unit," *IJCCS (Indonesian J. Comput. Cybern. Syst.*, vol. 13, no. 1, p. 43, 2019, doi: 10.22146/ijccs.40125.

[17] E. Suryati, A. Ari Aldino, N. Penulis Korespondensi, and E. Suryati Submitted, "Analisis Sentimen Transportasi Online Menggunakan Ekstraksi Fitur Model Word2vec Text Embedding Dan Algoritma Support Vector Machine (SVM)," vol. 4, no. 1, pp. 96–106, 2023, [Online]. Available: https://doi.org/10.33365/jtsi.v4i1.2445

[18] I. M. Parapat and M. T. Furqon, "Penerapan Metode Support Vector Machine (SVM) Pada Klasifikasi Penyimpangan Tumbuh Kembang Anak," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.* , vol. 2, no. 10, pp. 3163–3169, 2018, [Online]. Available: http://j-ptiik.ub.ac.id

[19] D. F. N. Anisa, I. Mukhlash, and M. Iqbal, "Deteksi Berita Online Hoax Covid-19 Di Indonesia Menggunakan Metode Hybrid Long Short Term Memory dan Support Vector Machine," *J. Sains dan Seni ITS*, vol. 11, no. 3, 2023, doi: 10.12962/j23373520.v11i3.83227.

[20] D. Maulina and R. Sagara, "Klasifikasi Artikel Hoax Menggunakan Support Vector Machine Linear Dengan Pembobotan Term Frequency-Inverse Document Frequency," *J. Mantik Penusa*, vol. 2, no. 1, pp. 35–40, 2018.

[21] Y. Jung, "Multiple predicting K-fold cross-validation for model selection," *J. Nonparametr. Stat.*, vol. 30, no. 1, pp. 197–215, 2018, doi: 10.1080/10485252.2017.1404598.

[22] F. Tempola, M. Muhammad, and A. Khairan, "Perbandingan Klasifikasi Antara KNN dan Naive Bayes pada Penentuan Status Gunung Berapi dengan K-Fold Cross Validation," *J. Teknol. Inf. dan Ilmu Komput.*, vol. 5, no. 5, p. 577, 2018, doi: 10.25126/jtiik.201855983.