Erie Kresna Andana et al., International Journal of Advanced Trends in Computer Science and Engineering, 8(1.3), 2019, 356 -

by Erie Kresna Andana Dosen Ft

ISSN 2278-3091

Volume 8, No.1.3, 2019

International Journal of Advanced Trends in Computer Science and Engineering

Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse6481.32019.pdf

https://doi.org/10.30534/ijatcse/2019/6481.32019

Comparative Analysis of Text Classification Using Naive Bayes and Support Vector Machine in Detecting Negative Content in Indonesian Twitter

Erie Kresna Andana[1,2], Muhaini Othman[1], Rosziati Ibrahim[1]

[1]Universiti Tun Hussein Onn Malaysia, Malaysia, gi180017@siswa.uthm.edu.my

[1]Universiti Tun Hussein Onn Malaysia, Malaysia, muhaini@uthm.edu.my

[1]Universiti Tun Hussein Onn Malaysia, Malaysia, rosziati@uthm.edu.my

[2]Muhammadiyah University of Surabaya, Indonesia

ABSTRACT

Research on the detection of social media content, especially Twitter, has been done. Twitter content detection is based on classifying content or words made by users (tweets) into two groups, namely positive and negative. Research to detect negative content or harsh words in Indonesian tweets is still rare. There are several studies that have been conducted, the detection of negative words is only limited to certain categories, such as pornography, hate speech, and others, so that if the negative word only includes one category, then if there are other negative words that do not belong to that category, this word will not be detected. This is a challenge for researchers to classify texts in Indonesian. In some research, to be able to detect and separate negative words and positive words in Indonesian, the Naive Bayes (NB) and Support Vector Machine (SVM) proved to produce better performance among other algorithms. Therefore, this paper aims to analyze the comparison of the results that have been achieved about detecting negative content on Indonesian twitter using NB and SVM. First, this paper will briefly explain the NB and SVM, then proceed with an explanation of the general research framework that has been carried out. In the results and discussion section, a comparison of the results achieved by existing researchers is explained. And based on these results, another approach will be proposed to detect negative content on Indonesian twitter.

Key words : Indonesian twitter, text classification, Naive Bayes, Support Vector Machine

1. INTRODUCTION

Nowadays social media sites are the most popular destination for internet surfers. Some are even very interesting and often visited by a wide audience [1]. Based on a survey of social media in 2018 [2], ranking first in social media applications is Facebook, which has 2.07 billion active users, and then followed by Instagram (800 million), Twitter (330 million), LinkedIn (500 million), Pinterest (200 million), and also Snap-Chat (178 million).

Social media was originally intended to facilitate users to exchange ideas, opinions, information, and respond to problems that occur. With the growing popularity of social media, new digital platforms continue to be developed and used where individuals interact and share information in the form of news and opinions [3]. People can actively express their opinions about products on social media, blogs, and comments on the websites visited [4, 5]. Social media also provides various services for the users [6]. But this also becomes a tool to disseminate other negative and destructive things through various content on social media.

Indonesia is one of the countries with the largest number of social media users. According to the Ministry of Communication and Information written in Indonesian web news, in 2013, the number of Indonesian Twitter users was 19.5 million [7]. Twitter is one of the fastest-growing social media in Indonesia. Twitter is familiar to school-age teenagers, even elementary schools. This is a social problem in Indonesia [8], that children also have gadgets and they have social media accounts, most of which have twitter, to figures in the class of government officials until the president, and political figures also use twitter. With the many Twitter users in Indonesia, it has the effect of the possibility of uncontrolled communication in the use of words or the number of netizens who communicate in abusive languages. Therefore, a tool is needed that can detect and state that the words written in their tweets are included in positive or negative content. This is to protect fellow Twitter users themselves and so that people can be wiser in sharing news through Twitter.

Research on the detection of abusive or negative language content on social media has carried out in recent years. By observing the written text, we can determine the direction of sentiment that is addressed in that opinion, whether it involves emotions and leads to negative sentiments or towards positive [9]. Turaob *et al.* [10] researched on the detection of abusive language in Thai. This research uses a dataset of posts and comments on Facebook. And as a text classification algorithm, using Multinomial Naive Bayes (MNB). Chen *et al.* [11], in their research, emphasized the importance of classifying coarse language on social media so that what is conveyed through social media becomes better and more friendly for children and adolescents. They conducted research using datasets taken from English Youtube comments and used Naive Bayes (NB) and Support Vector Machine (SVM) as classifiers.

Meanwhile, research on the detection of Indonesian negative content or abusive language on social media is still rare. However, there are some researchers have classified texts on Indonesian social media. Wongso *et al*. [12] conducted researched on text classification on Indonesian language news. As a dataset was news text taken from Indonesian language news web pages and using TF-IDF and Singular Value Decomposition (SVD) for feature extraction. Alfina *et al*. [13] researched the detection of hate speech in Indonesian. This research used datasets taken from Indonesian twitter and uses Naive Bayes (NB), Support Vector Machine (SVM), Bayesian Logistic Regression, and Random Forest Decision Tree (RFDT) as its text classification algorithm. And other researchers who examined negative content in Indonesian obtained the NB and SVM can produce a good level of accuracy, as will be discussed in this paper.

This paper presents a Comparative Analysis of Text Classification Using Naive Bayes and Support Vector Machine in Detecting Negative Content in Indonesian Twitter to explain the comparison of several studies that have been done before, about the detection of negative words in Indonesian using the two algorithms above. Evaluation of the algorithm in text classification, used various parameters, namely accuracy, precision, recall, and F1-score. This paper is organized follows: Section 2 explains the text classification algorithm that is widely used by negative content researchers in Indonesian, namely Naive Bayes and Support Vector Machine. Section 3 describes the research framework that was carried out in the text classification on Indonesian twitter. Section 4 discusses the comparison of the results that have been obtained by the researchers. And the conclusions are presented in Section 5.

2. TEXT CLASSIFICATION

Text classification is the process of recognizing text documents into two or more categories [14]. The most common form is binary classification, or classifying text in documents into only two categories. Text classification is often the first step in processing documents and then the document is further processed. However, text classification can also be the only process that is the purpose of research in text documents. An example that is often encountered is spam filtering. The purpose of text classification is not only to extract information from text. The basic approach of text classification is to get a feature set from extracting features that it can obtain a general description of the text in a document, and then apply a certain algorithm to group the text into appropriate categories according to the research objectives [9].

2.1. Naive Bayes

Text classification using the Naive Bayes algorithm is a probabilistic classification based on the Bayes Theorem assuming that no words are related to each other (each word is independent) [12]-[15]. First step is to calculate the probability of each class:

$$P(c) = \frac{f_c}{f_d} \tag{1}$$

Where f_c is the number of training document(s) labeled with c class and f_d is the number of all training documents. Then, the probability of the document for each class is calculated by this formula:

$$P(c|d) = P(c)\prod_{i=1}^{n} P(x_i \in d|c) \tag{2}$$

From the formula above, the class with the highest probability results will be assigned as class d for the document.

2.2. Support Vector Machine (SVM)

The concept of SVM can be explained simply as an effort to find the best hyperplane that functions as a separator of two classes in input space [16]-[17]. Pattern that are members of two classes: +1 and -1 and share discrimination boundaries. Margin is the distance between the hyperplane and the closest pattern of each class. This closest pattern is called support vector. The effort to find the location of the hyperplane is the core of the learning process in SVM. The available data is denoted as $D = \{(x_i, y_i), (x_i, y_i)\}, x_i \in R^n$, while the labels are denoted $y_i \in \{-1, +1\}$ for i=1,2, ..., l, which l is the amount of data. It is assumed that the two classes -1 and +1 can be completely separated by a dimension d hyperplane, which is defined in the following formula :

$$\omega^T.X + b = 0 \tag{3}$$

Patterns that belong to class -1 (negative samples) can be formulated as a pattern that satisfies the following inequalities:

$$\omega^T.X + b \leq -1 \tag{4}$$

While the pattern that includes class +1 (positive sample) satisfies the following inequalities:

$$\omega^T.X + b \geq +1 \tag{5}$$

3. FRAMEWORK OF TEXT CLASSIFICATION

In each text classification research, each document is always converted into the smallest unit of words or terms before being processed with a text classification algorithm. The process of downloading document data from the beginning to processing by the text classification algorithm, generally through a sequence written as a research framework as follows:

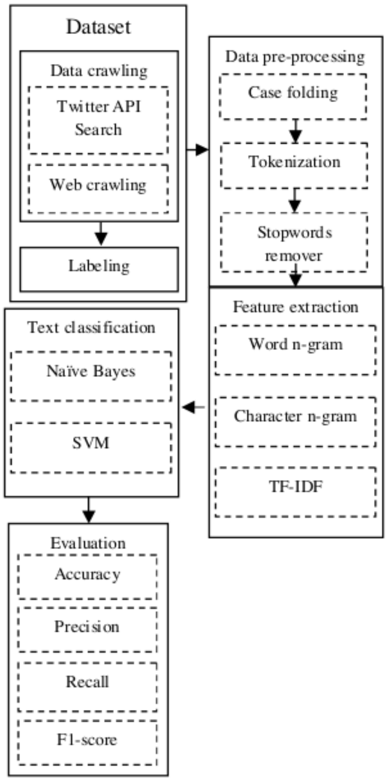

Figure 1: Research framework

As shown in Figure 1 above, research begins by downloading data in a tweet or news from Twitter using a tool provided by Twitter, namely: Application Programming Interface (API) [18]-[19]-[20]-[21]. Twitter was chosen as the source of data in this study because of reasons such as tweets generally public and short (140-280 characters). By utilizing the API, the application is made to retrieve tweet data from Twitter, based on keywords in the form of negative words in Indonesian, then save them into the database. In other research, data was also taken from sources other than Twitter, for example by crawling Indonesian websites containing negative words [20]-[22]. This test is done to assess the performance of text classification algorithms that are being researched can apply in general.

After collecting datasets, data is filtered by deleting the same or duplicated tweets and simultaneously labeling the data into 2 labels, namely: negative and positive content. And after the labeling process, the next step is data pre-processing and normalization of text, which is to ensure that text in the form of tokens that will be processed in the text classification algorithm will be verified as words in Indonesian. The pre-processing and normalization steps are responsible for cleaning inputs and converting data into forms that can be used by the text classification algorithm [23]. The pre-processing consists of case-folding, tokenizing, and stopword remover.

Case-folding is the process to convert all documents into lowercase. And tokenizing is the process of deciphering the original description in the form of sentences into words by removing punctuation characters between two sentences, such as dot (.), comma (,), and space. The tokenizing process also deletes several unnecessary attributes in tweetes such as username, re-tweet (RT), uniform resource locator (URL) address, hashtag (#), and emoticons from the dataset. While stopword remover is removing vocabulary that is not a feature (unique word) of a document. For example "di", "oleh", "pada", "sebuah", "karena" and so on. Before the stopword removal process is done, a stopword list must be created. If included in the stoplist, the words will be removed from the description so that the words left in the description are considered words that characterize the contents of a document or keywords. Stopword remover aims to make dataset/tokens processed by the text classification algorithm smaller in size [12]-[13]-[19]-[20]-[21].

The next step is feature extraction which aims to get words that will be processed in text classification, in the form of original vocabulary from an Indonesian dictionary. The challenge in Indonesian text classification is that Twitter users write their tweets in various forms. There are formal forms of tweets, but many are informal (slang), deleted vowels, excessive vocal repetitions, consonant replacements with other consonants that have sound similarities, letters replaced with numbers, or vocabulary usage from foreign languages. Repeated writing of vocals or letters usually shows or expresses the feelings/emotions or mood of the author. This feeling is caused by the events or events he wrote or because of the closeness of the relationship between the existing users, both individuals, and groups [25].

To make improvements to informal word forms or slang, a slang dictionary is arranged. Likewise, for the use of foreign language vocabulary, to make improvements to words before being processed in an algorithm, it is necessary to build a dictionary of foreign words that are often used by Indonesian users. The meaning of the actual word can be found in this dictionary. And then to detect the form of words from slang words and words in foreign languages, word n-grams are used, namely, word unigram and word bigrams [13]-[20]-[21]. Whereas to be able to detect deleted vowel writing, excessive vocal repetition, consonant replacements with other consonants that have sound similarity, and substitution of letters with numbers, is to use n-gram char, namely char trigrams and char quadgrams, so that letters are written incorrectly can be found and replaced with the actual word [21].

Then, the next step is weighting each word using the TF-IDF method [10]-[12]-[19]-[20]-[22]. TF-IDF is a statistical measure used to evaluate how important a word is in a document or in a group of words. The frequency of occurrence of words in a given document shows how important the word is in the document. Word weight is greater if it often appears in documents and gets smaller if it appears in many documents. Wongso *et al*. [12] also use Singular Value Decomposition (SVD) to complete the TF-IDF performance. The purpose of using this SVD is used to reduce data dimensions that are too large. And Izzah *et al*. [20] added using the Most Common Porn words (MCP) and the Most Common Non-Porn words (MCNP) as feature extraction.

After that, the process continued with the classification of text using the Naive Bayes or SVM algorithm. Naive Bayes can calculate probability values and can be used to determine a word including negative or positive. While SVM can separate 2 classes by drawing a hyperplane with a maximum margin value so that it can be used to determine the negative or positive class. The performance of the two algorithms is measured by calculating the values of accuracy, precision, recall, and F1-score.

4. RESULT AND DISCUSSION

The following are the results of several methods that have been used to detect negative content on Indonesian twitter. To observe the performance of the algorithms used, we also include research on text classifications on Indonesian language news websites.

Table 1: The methods for Indonesian text classification

Problem	Method	Feature extraction	Objectives
Text classification in Indonesian news articles [12]	- Multinomial Naive Bayes (MNB) - Bernoulli Multivariate Naive Bayes (BNB) - Gaussian Naive Bayes (GNB) - Support Vector Machine (SVM)	TF-IDF, SVD, and combination of them	Classify texts in Indonesian news into five labels, namely economy, health, technology, sports, and politics
Identify pornographic text on twitter that is written multilingual [19]	- Decision Tree (DT) - Naive Bayes (NB) - Support Vector Machine (SVM)	TF-IDF	Classify texts written in multilingual so that using datasets in Indonesian, English, and a mixture of both
Classification pornography content on twitter [20]	- Naive Bayes (NB) - Support Vector Machine (SVM)	TF-IDF, word unigram, bigrams, MCP, MCNP, and combination of them	Classify words into two labels, porn and non-porn
Detection of abusive language in Indonesian social media [21]	- Naive Bayes (NB) - Support Vector Machine (SVM) - Random Forest Decision Tree (RFDT)	Word unigram, bigrams, trigrams, char trigrams, quadgrams, and combination of them	Classify words into three labels: not abusive, abusive but not offensive and offensive. And two classification labels are not abusive and abusive

As seen in Table 1, some research has used Naive Bayes and SVM algorithms for the classification of Indonesian twitter text and gives good results. Likewise the use of these algorithms to classify news on Indonesian language news websites. And the results of the experiments can be seen in Table 2 below:

Table 2: Experimental results

Problem	Method	Evaluation parameters	Results
Text classification in Indonesian news articles [12]	- Multinomial Naive Bayes (MNB) - Bernoulli Multivariate Naive Bayes (BNB) - Gaussian Naive Bayes (GNB) - Support Vector Machine (SVM)	Precision and recall	- Precision=98,4%, recall=98,4% using combination of MNB with TF-IDF - Precision=98,2%, recall=98,2% using combination of BNB with TF-IDF - Precision=97,94%, recall=97,90% using combination of SVM with TF-IDF
Identify pornographic text on twitter that is written multilingual [19]	- Decision Tree (DT) - Naive Bayes (NB) - Support Vector Machine (SVM)	Accuracy and k-fold cross validation	Indonesian datasets: - The highest accuracy=94,01% using NB at k-9 - The average accuracy=92,78% - The SVM highest accuracy=88,99% at k-3 - The average accuracy=87,60% English datasets: - The highest accuracy=85% using SVM at k-4 - The average accuracy=83,33% - The NB highest accuracy=84,50% at k-4 - The average accuracy=82,22%
Classification pornography content on twitter [20]	- Naive Bayes (NB) - Support Vector Machine (SVM)	F1-score Precision, recall, F1-score for the best combination	- F1-score=90,83% at k-10 using combination of NB with word unigram+word bigram+TF-IDF - Precision=96,01%, and recall=86,98%, and F1-score=91,14% at k-10 using combination of SVM with word unigram+word bigram+TF-IDF+MCP+MNCP)
Detection of abusive language in Indonesian social media [21]	- Naive Bayes (NB) - Support Vector Machine (SVM) - Random Forest Decision Tree (RFDT)	F1-score	- Three classes: F1-score=70,06% using combination of NB with word unigram+word bigrams F1-score=67,48% using combination of SVM with word unigram - Two classes: F1-score=86,43% using combination of NB with word unigram F1-score=83,94% using combination of SVM with word unigram

In research on the classification of Indonesian language news texts, news groupings into 5 categories, namely economic, health, technology, sports, and politics, the dataset used is news from Indonesian websites. In this research, as can be seen in Table 1, Wongso *et al*. [12] used TF-IDF and Singular Value Decomposition (SVD) for the features extraction. SVD is used to reduce data dimensions that are too large. The text classification algorithm used is Naive Bayes, which consists of Multinomial Naive Bayes (MNB), Multivariate Bernoulli Naive Bayes (BNB), and Gaussian Naive Bayes (GNB), and also uses Support Vector Machine. This research resulted in the combination of TF-IDF and MNB achieving the highest precision and recall values, which was around 98.4% followed by TF-IDF and BNB, which was around 98.2%. This result is supported by other studies that MNB tends to be better at the level of precision and recall than BNB if the dataset contains a large amount of vocabulary.

Whereas the other combinations, namely TF-IDF and GNB and the combination of TF-IDF and SVD and MNB, give the wrong results. The process of using a combination of TF-IDF and GNB results in an error. While the results given by a combination of TF-IDF and SVD and MNB cannot meet the requirements for the running of the MNB algorithm. MNB cannot process vectors that have negative values. While the combination of TF-IDF and SVD and GNB gave the worst results because the data in the experiment was not continuous. GNB relies on the assumption that

continuous values associated with each class are distributed according to the Gaussian Distribution.
The use of SVD in this research does not aim to increase accuracy but to compensate for large processing times while maintaining good algorithm performance. In this research, SVD is not able to influence the performance of algorithms to be better like other existing research. The feature reduced by SVD causes a loss of its unique value which is a representation of data in a document. And consequently further in reducing the accuracy of the algorithm used. SVD is supported to make extraction of information more efficient but does not always perform well. In addition to measuring algorithm performance from precision and recall, it is also measured from processing time. The combination of TF-IDF and MNB produces the fastest processing time.

Another case with research using multilingual, in this case the dataset used is twitter in Indonesian, English, and a mixture of Indonesian and English as well as often written in tweets of users from Indonesia [19]. This research is done by dividing the dataset with different parameters on the composition of training and testing data. The method used is K-Folds Cross-Validation with k = 2 to k = 10. The aim is to determine the optimal fold value for the classification of pornographic tweets. The results show that the composition of training and testing data affects the accuracy.

The experimental results are shown in Table 2, for the dataset in Indonesian, the classification method that has the best accuracy is Decision Tree of 94.49%, using a comparison of training and testing data at k-7. In the Naive Bayes method, the highest accuracy value is in k-9 of 94.01%. And SVM produces the best accuracy in k-3 of 88.99%. and this results in the most optimal average accuracy value of 92.78%.

In the dataset in English, the SVM has the highest average accuracy value of 83.33% compared to the other two classification methods. In the Naive Bayes, the highest accuracy is in multiples of 2, the Decision Tree at k-3, while the SVM at k-4. In the Indonesian English Language dataset, the highest average accuracy value produced by the SVM is 72.72%. The highest accuracy value was generated by the Naive Bayes at k-7, the Decision Tree at k-10, while SVM at k-8.

The level of accuracy of text classifiers is influenced by differences in the composition of training and testing data. The greater the number of multiples that indicate the more amount of training data, causing the value of accuracy tends to decrease.

For Indonesian datasets, SVM produces the lowest accuracy value of the other two methods. This is because grammar in Indonesian is more complex than English grammar which has special rules for writing certain sentence patterns. And this is also in line with previous research, that classification tests involving Indonesian and English datasets, using SVM also produce lower accuracy for Indonesian datasets. Besides being caused by more complicated Indonesian grammar, datasets obtained from Indonesian twitter users also affect the accuracy of the results. Indonesian twitter users tend to use non-standard and informal words, so that in the pre-processing process after word cleaning and standardization, the tokens/words produced do not match the dictionary.

While Izzah *et al*. [20] detected negative content, in case, pornographic words, which are on twitter. The dataset used in addition to twitter is also an Indonesian website. Some combinations of features extraction used are unigram, unigram + bigrams, TF-IDF, the most common pornographical words (MCP), and the most common non-pornographical words (MCNP). And the algorithms used are NB and SVM.

As shown in Tabel 2, the highest F1-score produced by the SVM algorithm was 91,14%. F1-score is obtained if SVM is combined with Unigram + Bigrams + MCP + MCNP + TF-IDF. Precision values are 96.01% and recall is 86.98%. While the NB algorithm only produces an F1-score of 89.04%. NB produced the highest F1-score of 90.83% with feature extraction using a combination of Unigram + Bigrams + TF-IDF. This model was also tested using various K-fold Validations and got the highest score on k-10. And they use balanced data in the second experiment. Based on the experimental results on balanced data, the F1-score from the model increased by 0.84% to 91.98%. This happens because the features of the two classes are balanced, so the classification algorithm can classify into the appropriate class easily.

Then, to observe whether the algorithm used applies generally, an algorithm is tested using a dataset from the website. Experiments on the dataset from this website show that the SVM algorithm can achieve a high precision value of 98.33%. This is because the words that appear most often in the previous experimental dataset also exist in the current dataset. The reset should be done on the dataset. In this experiment, the recall value was 81% and F1-score was 84.73%.

And the other research shows that to detect abusive words that include all abusive words, dirty words, hate speech, and various other abusive words in Indonesian, a new modification is needed in the dataset. Ibrohim *et al*. [21] classifies abusive words by introducing new datasets. They divided abusive words into several types so that they included most of the negative words that often appeared on Indonesian twitter. And in this research, they conducted two experimental scenarios. In the first scenario, dividing tweets becomes three classes: not abusive, abusive not offensive and offensive. This is done following the explanation from previous Indonesian language research, that abusive words is not always an offensive. For the second scenario, only classifies the dataset into two classes that are not abusive and abusive to find out whether the classification results would be better if only classifying two labels. In this scenario, tweets labeled abusive but not offensive and offensive language will be labeled as abusive language. For both scenarios, they use the word n-gram and char n-gram features with the NB, SVM and RFDT algorithms as classifiers. The word n-gram used is the word unigram, word bigrams, word trigrams and the combination of all. While the char n-gram used is char trigrams, char quadgrams, and also all combinations.

As the results can be seen in Table 2, for the first scenario, NB with the word unigram and bigrams features give the highest F1-score with a value of 70.06%, followed by NB with the features of word unigram and word bigrams and word trigrams (69.61%) and NB with char quadgrams (69.55%). And for the second scenario, it can be seen that NB with the word unigram feature provides the highest F1-score results with a value of 86.43%, followed by NB with char trigrams and char quadgrams (86.17%) and then followed by NB with the word unigram and word bigrams features (86.12%). From both scenarios, the NB performs better than SVM and RFDT to classify tweets. It also shows that using the word unigram and the word n-gram combination features giving better results.

5. CONCLUSION AND FUTURE WORK

Text classification can be used to detect negative content in sentences written on twitter. The detection of negative content is an interesting object of research because various languages in this world have their uniqueness. For Indonesian sentences, research that has been done is still rare. Besides being rare, writing Indonesian Twitter content, often in the form of mixed words between formal words and slang words. In addition to using mixed languages, another challenge in processing text data on Indonesian twitter is: repeating many vowels, removing vowels, replacing consonants that have similar sounds, replacing letters with numbers that have similar shapes.

For the classification of the Indonesian news websites, NB is better than SVM. On the classification of negative and positive words, in this case, limited to the pornographical words found on Indonesian twitter and web, SVM produces better results than NB with TF-IDF feature extraction with a combination of unigram, bigrams, MCP and MCNP and with vary the composition of training and testing data. But conversely if only using TF-IDF and unigram feature extraction and changing the composition of training and testing data, NB gives better results.

And to recognize informal words or written in slang or other informal word forms, it is necessary to match the slang dictionary which must be included in the next research. The new dataset is also needed to recognize expanded negative words not limited to pornographic words only. And the feature extraction here also needs to use the word n-gram or char n-gram. With the research that has been achieved, by improving the dataset and using word n-grams and char n-grams, it shows that the Naive Bayes algorithm is better than SVM and followed by the RFDT. Henceforth this research still needs to be refined regarding the completeness of the slang dictionary, the use of foreign languages, and the replacement of consonants that have similar sounds that are often written by Indonesian twitter users, so that the performance of the algorithm can be improved.

In addition to the classification of text in two languages compared, or on the classification of multilingual data, for example, Indonesia and English, the same algorithm can produce different accuracy. This is influenced by the grammar factors of the two languages that are very unique and different.

REFERENCES

1. M. Sami, S. Memon, J. Baloch, and S. Bhatti. **An automated framework for finding fake accounts on facebook**, *International Journal of Advanced Studies In Computer Science And Engineering (IJASCSE)* Volume 7 Issue 2, 2018, pp. 8-16
2. A. Lua. **Social media trends report: 10 key insights into the present and future of social media**. *Buffer Soc 2018*, 2018:1, https://blog.bufferapp.com/social-media-trends-2018 (accessed March 27, 2018)
3. W. Housley, H. Webb, M. Williams, R. Procter, A. Edwards, M. Jirotka, P. Burnap, B.C. Stahl, O. Rana, and M. Williams. **Interaction and transformation on social media: the case of twitter campaigns**, *Social Media+ Society* 4.1 2018, pp. 1-12 https://doi.org/10.1177/2056305117750721
4. O. Oueslati, A.I.S. Khalil, and H. Ounelli. **Sentiment analysis for helpful reviews prediction**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, Volume 7 No. 3, May - June 2018, pp. 34-40 https://doi.org/10.30534/ijatcse/2018/02732018
5. H. Xu, Y. Zhang, and R. DeGroof. **A feature-based sentence model for evaluation of similar online products**, *Journal of Electronic Commerce Research*, Volume 19 No. 4, 2018, pp. 320-335
6. P. Deshwal and R. Ranjan. **Facebook blog innovation experience, course and hosteller**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, Volume 8 No. 1.1, 2019, pp. 145-150
7. Kominfo: **Pengguna Internet di Indonesia 63 Juta Orang**., Web Kominfo 2013, https://kominfo.go.id/index.php/content/detail/3415/Kominfo+%3A+Pengguna+Internet+di+Indonesia+63+Juta+Orang/0/berita_satker (accessed May 23, 2018)
8. S.G. Handono, K. Laeheem, and R. Sittichai. **Factors related with cyberbullying among the youth of Jakarta, Indonesia**, *Children and Youth Services Review 99 (2019)*, pp. 235-239 https://doi.org/10.1016/j.childyouth.2019.02.012
9. P. Thakur and R. Shrivastava. **A review on text based emotion recognition system**, *International Journal of Advanced Trends in Computer Science*

and Engineering (IJATCSE), Volume 7 No. 5, September - October 2018, pp. 67-71 https://doi.org/10.30534/ijatcse/2018/01752018

10. S. Turaob and J. Mitrpanont. **Automatic discovery of abusive Thai language**, *International Conference on Asia-Pacific Digital Libraries 2017*, Bangkok, Thailand, 2017, pp. 267-278 https://doi.org/10.1007/978-3-319-70232-2_23
11. X. Chen, M. Vorvoreano, and K. Madhavan. **Mining social media data for understanding students' learning experiences**, *IEEE Trans Learn Technology 2014*, pp. 246-259 https://doi.org/10.1109/TLT.2013.2296520
12. R. Wongso, F.A. Luwinda, B.C. Trisnajaya, and O.R. Rudy. **News article text classification in indonesian language**, *2nd International Conference on Computer Science and Computational Intelligence 2017 (ICCSCI 2017)*, Bali, Indonesia, 2017, pp. 137-143 https://doi.org/10.1016/j.procs.2017.10.039
13. I. Alfina, R. Mulia, M.I. Fanany, and Y. Ekanata. **Hate speech detection in the indonesian language: a dataset and preliminary study**, *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, Bali, Indonesia, 2017, pp. 233-237 https://doi.org/10.1109/ICACSIS.2017.8355039
14. S.A. Salloum, M. Al-Emran, A.A. Monem, and A. Shaalan. **A survey of text mining in social media: facebook and twitter perspectives**, *Advances in Science, Technology and Engineering Systems Journal* Vol. 2 No. 1, 2017, pp. 127-133 https://doi.org/10.25046/aj020115
15. D. Berrar. **Bayes' theorem and Naive Bayes classifier**, *Encyclopedia of Bioinformatics and Computational Biology*, Tokyo Institute of Technology, Tokyo, Japan, 2018, pp. 1-10 https://doi.org/10.1016/B978-0-12-809633-8.20473-1
16. A. Rizaldy and H.A. Santoso. **Performance improvement of Suppport Vector Machine (SVM) with information gain on categorization of Indonesian news documents**, *2017 International Seminar on Application for Technology of Information and Communication (iSemantic)*, Semarang, 2017, pp. 227-232 https://doi.org/10.1109/ISEMANTIC.2017.8251874
17. G. Chen and Z. Ge. **SVM-tree and SVM-forest algorithms for imbalanced fault classification in industrial processes**, *IFAC Journal of Systems and Control*, 2019 https://doi.org/10.1016/j.ifacsc.2019.100052
18. S. Kunal, A. Saha, A. Varma, and V. Tiwari. **Textual dissection of live twitter reviews using Naive Bayes**, *International Conference on Computational Intelligence and Data Science (ICCIDS 2018)*, Gurugram, India, 2018, pp. 307-313 https://doi.org/10.1016/j.procs.2018.05.182
19. E. Barfian, B.H. Iswanto, and S.M. Isa. **Twitter pornography multilingual content identification based on machine learning**, *2nd International Conference on Computer Science and Computational Intelligence 2017 (ICCSCI 2017)*, Bali, Indonesia, 2017, pp. 129-136 https://doi.org/10.1016/j.procs.2017.10.024
20. N. Izzah, I. Budi, and S. Louvan. **Classification of pornographic content on twitter using Support Vector Machine and Naive Bayes**, *2018 4th International Conference on Computer and Technology Applications (ICCTA 2018)*, Istanbul, Turkey, 2018, pp. 156-160 https://doi.org/10.1109/CATA.2018.8398674
21. M.O. Ibrohim and I. Budi. **A dataset and preliminaries study for abusive language detection in Indonesian social media**, *3rd International Conference on Computer Science and Computational Intelligence 2018 (ICCSCI 2018)*, Jakarta, Indonesia, 2018, pp. 222-229 https://doi.org/10.1016/j.procs.2018.08.169
22. T.B. Adji, Z. Abidin, and H.A. Nugroho. **System of negative indonesian website detection using TF-IDF and vector space model**, *2014 IEEE International Conference on Electrical Engineering and Computer Science*, 24-25 November 2014, Bali, Indonesia, 2014, pp. 174-178 https://doi.org/10.1109/ICEECS.2014.7045240
23. I.S. Makki and F. Alqurashi. **An adaptive model for knowledge mining in databases "emo_mine" for tweets emotions classification**, *International Journal of Advanced Trends in Computer Science and Engineering (IJATCSE)*, Volume 7 No. 3, May - June 2018, pp. 52-60 https://doi.org/10.30534/ijatcse/2018/04732018
24. M.A. Jiwanggi and M. Adriani. **Topic summarization of microblog document in Bahasa Indonesia using the phrase reinforcement algorithm**, *5th Workshop on Spoken Language Technology for Under-resourced Languages, SLTU 2016, 9-12 May 2016*, Yogyakarta, Indonesia, 2016, pp. 229-236 https://doi.org/10.1016/j.procs.2016.04.054
25. R. Kumari and M. Sasane. **Emotion analysis using text mining on social networks**, *International Journal of Innovative Research In Technology (IJIRT)*, Volume 2 Issue 1, June 2015, pp. 20-27

Erie Kresna Andana et al., International Journal of Advanced Trends in Computer Science and Engineering, 8(1.3), 2019, 356 -

ORIGINALITY REPORT

21%	%	21%	7%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Muhammad Okky Ibrohim, Indra Budi. "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media", Procedia Computer Science, 2018 Publication	5%
2	Edo Barfian, Bambang Heru Iswanto, Sani Muhamad Isa. "Twitter Pornography Multilingual Content Identification Based on Machine Learning", Procedia Computer Science, 2017 Publication	4%
3	Rini Wongso, Ferdinand Ariandy Luwinda, Brandon Christian Trisnajaya, Olivia Rusli, Rudy. "News Article Text Classification in Indonesian Language", Procedia Computer Science, 2017 Publication	2%
4	Tora Fahrudin, Joko Lianto Buliali, Chastine Fatichah. "Ina-BWR: Indonesian bigram word	1%

rule for multi-label student complaints", Egyptian Informatics Journal, 2019
Publication

5	Submitted to University of Southampton Student Paper	1 %
6	Nur Rosyid Mubatada'i, Ali Ridho Barakbah, Afrida Helen. "Semantic Information Retrival for Scientific Experimental Papers with Knowlege based Feature Extraction", INOVTEK Polbeng - Seri Informatika, 2019 Publication	1 %
7	"Text Classification and Categorization", Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications, 2012. Publication	1 %
8	Submitted to Asia Pacific University College of Technology and Innovation (UCTI) Student Paper	1 %
9	Submitted to Universitas Siswa Bangsa Internasional Student Paper	1 %
10	Arjon Turnip, M. Ilham Rizqywan, Dwi E. Kusumandari, Mardi Turnip, Poltak Sihombing. "Classification of ECG signal with Support Vector Machine Method for	1 %

Arrhythmia Detection", Journal of Physics: Conference Series, 2018
Publication

11	Umroh Makhmudah, Saiful Bukhori, Januar Adi Putra, Bratasena Anggabayu Bhirawa Yudha. "Sentiment Analysis Of Indonesian Homosexual Tweets Using Support Vector Machine Method", 2019 International Conference on Computer Science, Information Technology, and Electrical Engineering (ICOMITEE), 2019 Publication	1 %
12	Submitted to University of Melbourne Student Paper	1 %
13	Rahmat Hendrawan, Adiwijaya, Said Al Faraby. "Multilabel Classification of Hate Speech and Abusive Words on Indonesian Twitter Social Media", 2020 International Conference on Data Science and Its Applications (ICoDSA), 2020 Publication	1 %
14	Aditya Akundi, Bill Tseng, Jiamin Wu, Eric Smith, M Subbalakshmi, Francisco Aguirre. "Text Mining to Understand the Influence of Social Media Applications on Smartphone Supply Chain", Procedia Computer Science, 2018 Publication	<1 %

15	Muhammad Okky Ibrohim, Erryan Sazany, Indra Budi. "Identify Abusive and Offensive Language in Indonesian Twitter using Deep Learning Approach", Journal of Physics: Conference Series, 2019 Publication	<1 %
16	Imam Riadi, Sunardi Sunardi, Panggah Widiandana. "Mobile Forensics for Cyberbullying Detection using Term Frequency - Inverse Document Frequency (TF-IDF)", Jurnal Ilmiah Teknik Elektro Komputer dan Informatika, 2020 Publication	<1 %
17	Yetisgen-Yildiz, M.. "A new evaluation methodology for literature-based discovery systems", Journal of Biomedical Informatics, 200908 Publication	<1 %

Exclude quotes	On	Exclude matches	< 20 words
Exclude bibliography	On		