Detection of Hate Speech in Indonesian Language on Twitter Using Machine Learning Algorithm

Febby Apri Wenando
Informatics Engineering
Universitas Muhammadiyah Riau
Pekanbaru, Indonesia
febbyapri@umri.ac.id

Evans Fuad
Informatics Engineering
Universitas Muhammadiyah Riau
Pekanbaru, Indonesia
evansfuad@umri.ac.id

***Abstract*—Hate speech is an act of communication carried out by an individual or group in the form of provocation or insults to other individuals or groups. Hate speech is prohibited because it can trigger acts of violence and prejudice either from the perpetrators of the statement or victims of the act. This study aims to find the best algorithm for detecting Hate Speech by comparing the Decision Tree, Naive Bayes, Support Vector Machine and Random Forest algorithms using N-Gram-based Word Scoring (TF-IDF) methods, including the Union Gram, Bigram and Trigram using the programming language Python. The results reveal that the Naive Bayes algorithm shows the best results using the Trigram feature with an Accuracy of 88.57%, Precision of 96.75% and Recall of 99.34%.**

Keywords—Machine Learning, text classification, Python, Decision Tree, Naive Bayes, SVM, Random Forest

I. Introduction

Social media is now a communication media that is quite popular among internet users. One of the most popular social media is Twitter. Twitter can be an effective and efficient place of promotion or campaign. The success team of a regional head candidate or president candidate can now use social media in their candidate's campaign. Black campaign is an act of insulting, slandering, pitting, inciting, or spreading hoaxes committed by a candidate, group of people, political parties or supporters of a candidate against their opponents. This is different from expressing criticism of a particular candidate's vision and mission or program, which is not classified as the Black Campaign. But for now after the campaign period has passed, there is a new term called Hate Speech.

Hate Speech is an act of communication carried out by an individual or group in the form of provocation or insults to other individuals or groups in various aspects such as race, color, ethnicity, gender, disability, sexual orientation, citizenship, religion, etc. [1] - [7].

Research on hate speech is mostly done in English texts [3] - [7], and there are still few studies conducted in Indonesian Language texts. Pratiwi, et al. [1] tried to conduct research to detect hate speech in Indonesian texts, but this study only focused on hate speech on religion. The results of this study are creating a new dataset for the detection of hate speech in Indonesian. The dataset used is divided into 2 classes, namely Hate Speech toward religion and non-religion. The quality of the dataset used is inadequate, because it only discusses tweets about religion, and with an unequal number of tweets between the two classes. Therefore, the classification results can wrongly detect tweets related to religion as words that contain Hate Speech.

Alfina, et al. [2] also conducted research on the detection of hate speech using Indonesian language text by creating a new dataset. The dataset was taken from Twitter using a hashtag on the DKI PILGUB by using the hashtags #DebatPilkadaDKI and #SidangAhok. The dataset collected was 710 and divided into 2 classes, namely 520 HS tweets and 260 NONHS tweets. Because the dataset was not balanced between the two classes, the researchers reduced the majority dataset with the Undersampling Method to be balanced, and the result became 260 HS tweets and 260 NONHS tweets. It can be seen that the dataset in this study was reduced by 450, which should be maximized dataset to detect Hate Speech.

Limited use of data sets results in a decrease in the value of accuracy in text classification. The method proposed in this study addresses the problem of limited data sets and compares the performance of the Tf-Idf Scoring method with several machine learning algorithms.

II. Research Method

A. Data Set

The initial stage of this research is data collection. Previously there had been hate speech dateset totaling 713 in Text (.txt) format. Then the format is changed to CSV format.

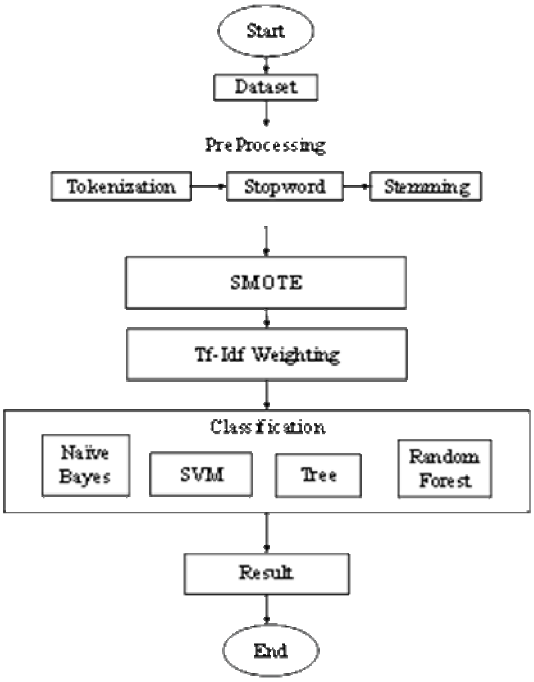

Figure 1. Research Flowchart

B. *Pre-process*

We adopt the preprocessing method used by [8] with a few modifications. Pre-processing is the initial stage in the classification of texts to prepare texts. In a text document, the feature in question is the text content contained in a text document. At this stage there are several stages carried out in this study:

- Case Folding

The Case Folding process is the stage for changing all letters in each document into lowercase letters.

- Tokenizing

Tokenizing is the process for separating words from paragraphs in text content into single words or multiple words.

- Stopword Removal

It is a process to eliminate common or less important words (stop list) that have little effect on the text. At this pre-process stage, the output is a structured text document and stored in a "Bag-of-Word" representation model. This representation is a model used in text classification because of the ease of use in classification purposes [8].

- Stemming

It is a process for finding the root of words. The root search for a word or commonly called basic word can reduce the index results without having to eliminate the meaning. In this study, the Stemmer used was the Tala Stemmer [9]. Tala Stemmer is the adoption of the Porter Stemmer algorithm in English.

- SMOTE

Synthetic Minority Oversampling Technique is a technique for handling imbalanced classes [10]. This study conducted the SMOTE technique aimed at balancing class data.

TABLE 1. CONFUSION MATRIX

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

The work assessment of the model is determined based on accuracy and precision. Accuracy and precision values can be calculated based on the values of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) formulated in the equation.

TABLE 2. Comparison Using SMOTE

Algorithm	**%Accuracy Original Data**	**%Accuracy SMOTE Data**
Naïve Bayes	47.5 %	76.4 %
SVM	50.0 %	76.4 %
Tree	50.0 %	58.8 %
Random Forest	42.5 %	64.7 %

Table 2 shows the results of the accuracy values comparison with the original dataset and using SMOTE. The results show the best accuracy results using the SMOTE technique found in the Weka application [10].

C. *Classification*

- Word Weighting Method

After the pre-process, the token or word produced is given a score that can represent how much influence the score has on a document [11]. The word scoring method used in this study is Term Frequency - Inverse Document Frequency (TF-IDF). The workings of the word scoring in the TF-IDF method are by using 2 scoring parameters, namely the local scoring $tf_{i,j}$ which is the score obtained from the frequency of occurrence of the word i in the document j and global scoring using idf_i which is the score obtained by considering the number of occurrences of the word i (DF_i) for all N documents. Then, the score of the local scoring is multiplied by the score of the global scoring [12]. How to calculate the score is using equation (1):

$$(1) \qquad w_{i,j} = tf_{i,j} \, x \left(\log\left(\frac{N}{DF_i}\right)\right)$$

This study uses a machine learning algorithm to classify text in detecting hate speech using the Weka Data Mining Tool application [12].

III. RESULTS AND DISCUSSION

Evaluation on this research was carried out using the K-fold Cross Validation Method. The first stage is to divide the data into k subset pieces that have the same size. The second stage is to use each subset to become test data and the rest is used for training data to form a model and bring up the evaluation results.

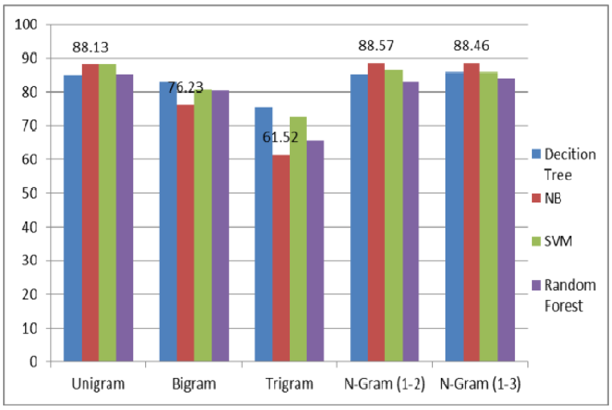

Figure 2. The Results of Accuracy Value Comparison

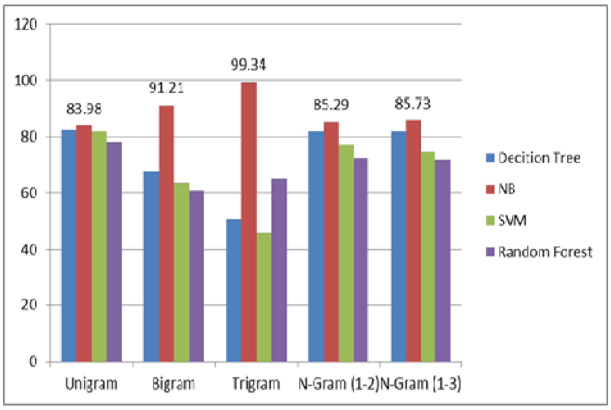

Figure 3. Kappa value comparison results

Figure 2 and Figure 3 show the results of the comparison of the accuracy and kappa values of the hate speech text classification. The results show that the Naive Bayes algorithm has the highest level of accuracy and kappa value, with an accuracy value of 88.57% and a Recall value of 99.34% and using the N-gram scoring method (1-2) and Trigram.

From the results of this study, it can be concluded that the Tf-Idf scoring method based on N-Gram (1-2) and Trigram with the Naïve Bayes algorithm shows the best results, in terms of accuracy and kappa values.

REFERENCES

[1] S. H. Pratiwi, “Detection of Hate Speech against Religion on Tweet in the Indonesian Language Using Naïve Bayes Algorithm and Support Vector Machine,” B.Sc. Tesis, Universitas Indonesia, Indonesia, (2016).

[2] Alfina, Ika & Mulia, Rio & Fanany, Mohamad Ivan & Ekanata, Yudo. Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study. 10.1109/ICACSIS.2017.8355039(2017).

[3] Waseem, Z., & Hovy, D. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In Proceedings of the NAACL student research workshop (pp. 88-93) (2016).

[4] Schmidt, A., & Wiegand, M. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media (pp. 1-10) (2017).

[5] Malmasi, S., & Zampieri, M. Detecting Hate Speech in Social Media. arXiv preprint arXiv:1712.06427. (2017).

[6] Warner, William & Hirschberg, Julia. Detecting hate speech on the world wide web. 19-26. (2012).

[7] Burnap, P., & Williams, M. L. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. Policy & Internet, 7(2), 223-242. (2015).

[8] Wenando, F. A., Adji, T. B., & Ardiyanto, I. Text Classification to Detect Student Level of Understanding in Prior Knowledge Activation Process. Advanced Science Letters, 23(3), 2285-2287. (2017).

[9] F. Z. Tala, “A study of stemming effects on information retrieval in Bahasa Indonesia,” *Inst. Log. Lang. Comput. Univ. Van Amst. Neth.*, (2003).

[10] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, (2002).

[11] I. H. Witten and E. Frank, *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition*. Morgan Kaufmann, (2005).

[12] Witten, Ian H., et al. "Weka: Practical machine learning tools and techniques with Java implementations." (1999).