IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

A comparison of classification algorithms for hate speech detection

To cite this article: T T A Putri *et al* 2020 *IOP Conf. Ser.: Mater. Sci. Eng.* **830** 032006

The Electrochemical Society
Advancing solid state & electrochemical science & technology
242nd ECS Meeting
Oct 9 – 13, 2022 • Atlanta, GA, US
Submit your abstract
Extended abstract submission deadline: April 22, 2022
Connect. Engage. Champion. Empower. Accelerate.
MOVE SCIENCE FORWARD

A comparison of classification algorithms for hate speech detection

T T A Putri[*]**, S Sriadhi, R D Sari, R Rahmadani, and H D Hutahaean**

PTIK-FT, Universitas Negeri Medan, Indonesia

*tansatrisna@unimed.ac.id

Abstract. Freedom of opinion through social media is frequently affect a negative impact that spreads hatred. This study aims to automatically detect Indonesian tweets that contain hate speech on Twitter social media. The data used amounted to 4,002 tweets related to politics, religion, ethnicity and race in Indonesia. The application model uses classification methods with machine learning algorithms such as Naïve Bayes, Multi Level Perceptron, AdaBoost Classifier, Decision Tree and Support Vector Machine. The study also compared the performance of the model using SMOTE to overcome imbalanced data. The results show that the Multinomial Naive Bayes algorithm produces the best model with the highest recall value of 93.2% which has an accuracy value of 71.2% for the classification of hate speech. Therefore, the Multinomial Naïve Bayes algorithm without SMOTE is recommended as the model to detect hate speech on social media.

1. Introduction

Social media development nowadays contributes the freedom of speech effect for people. Freedom of speech to express a feeling and thinking of something through social media as if being a trend that should be done by social media users. Freedom of speech gives impact to the individual to share opinion and belief about anything [1]. However, there are some individuals who abuse this freedom of expression to make offensive comment or promote their beliefs that could give negative impacts to the people [2].

One of the negative impacts from freedom of speech is the number of hate speech that were shared by irresponsible people. Hate speech commonly defined as any communication that underestimates someone or a group with specific characteristics such as race, skin color, ethnicity, gender, sexual orientation, nationality, religion and another characteristics [3]. Hate speech can also be defined as a certain offensive form of language that utilizes point of view about a social group to express hate ideology [4]. Based on the definitions from both of experts, we could conclude that hate speech is any kind of communication which is offensive, underestimating and humiliating an individual or group of people.

Indonesia as the third country with the most Twitter users also faces the problem of hate speech [5]. Hate speech is considered as the highest case happened on social media that was complained to Minstry of Communication and Information of Indonesia[1]. That matter is possible happened in Indonesia since Indonesia consists of race, ethnicity, and religion diversity.

[1] http://m.viva.co.id/digital/digilife/923759-ujaran-kebencian-konten-negatif-terbanyak-masuk-ke-kominfo

Study related to hate speech that happened through social media already done before and become interesting to be discussed. Related studies had done the classification to detect hate speech on Twitter using English text data. This study aimed to obstruct hate speech spread on Twitter [6]. Another study was done that used classification method on Twitter using English text data. This study aimed to detect hate speech that was addressed to black people [7].

As well study related to hate speech that used Indonesian language is still a few. Studyes related to hate speech in Indonesian detected a hate speech related to politic [8]. In collecting the data, used keywords that was related to Jakarta Governor Election 2017. This study stated that Random Forest Decision Tree was the best model with the highest F-measure compared to other models that used Naïve Bayes, SVM and Logistic Regression algorithm. While another study in Indonesian language done in order to detect a hate speech related to religion in Indonesia [9].

Previous study done to detect hate speech that was shared through Twitter social media [6]. Hate speech regarding to this study is a hate speech in any form of speech which is racist and sexist. This study used the known list of criteria related to critical racist speech and used it to classify corpus of collected data. Data that was used in this study is English text data from Twitter. The number of data of this study was more than 16.000 tweets. The goal of this study was defining the impact of various language features and the correlation to n-gram character in order to detect a hate speech. Result showed that n-gram feature with n=4 and accompanied by knowledge of user's gender, has the highest accuracy.

Previous study used Indonesian language worked on hate speech detection related to politic [8]. Features used in this study were word unigram, word bigram, character trigram, character quadragram and negative sentiment. Algorithms that were compared in this study were Naïve Bayes, SVM, Bayesian Logistic Regression and Random Forest Decision Three. Result showed that Random Forest was the highest F-measure model (93.5%). But this research did not show the recall performance of every model that was built. For our research, recall score is the most important performance evaluation since recall is the measure of the ability of a model to define the true positive hate speech.

2. Methods

2.1. Data collection

Data collection in this study done by crawling to Twitter data. Crawling done by using Twitter API (Application Program Interface) owned by Twitter. Usage of Twitter API done on coded program by us using Python programming language. Input for this phase is keywords. And keywords used in this research were "jokowi diktator", "antek komunis", "pki", "anti islam", "rezim jokowi", "islam teroris" and "budha teroris". Based on those keywords, this study is not only about politic but also religion. Data collection done from the beginning of August 2017 to September 2017. We collected data as many as 23.061 tweets and data that could be used in this research was 4.002 tweets.

In order to obtain the best model of hate speech detection, we also compared the model we've resulted with the model of previous study. The previous study provided a data consists of 260 hate labelled and 430 non hate labelled [8]. At last, we will show the result of our comparison.

2.2. Labelling

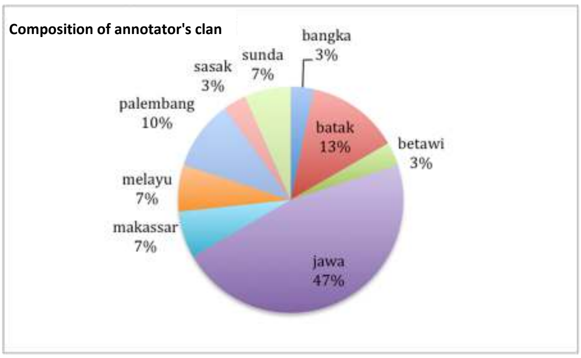

Figure 1. Composition of annotator's clan.

In this study, labelling data done manually and helped by annotators. Annotators composed by 30 volunteers who had a different background even the clan, religion and education. Each of the volunteers labelled 400 tweets and for every tweet, labelled by 3 different annotators.

Data would be labelled by annotators, and would be sent by email in *.xlsx file format. If the annotators had done labelling the data, so the data would be sent back to us by email.

In labelling the data, tweets that were hate speech would be labelled as "hate" and "noHate" for the not hate speech data. And in this phase, we conducted the characteristics of a hate speech according to Surat Edaran Kapolri No: SE/6/X/2015 which is contained by one of these factors such as provocation, incitement, insult and defamation. Thus, the number of hate speech from the data is 2776 and 1226 as the not hate.

Based on Figure I, we could see the diversity of clan of the annotators. We would like to show that the bias possibility of the annotators was a little. The annotators were not all a javanese, so when there was a speech related to javanese or bataknese, not all of them felt offended and labelled it as a hate.

2.3. Preprocessing

The first phase of preprocessing is normalization and case folding. This phase is a process to transform all words to be lower case and standard form. Transforming letter to be lower case aims to synchronize all of the words that will be processed to feature extraction phase. If the words are not synchronized then words such as "Makan" and "makan" will be categorized as a different word though those words have a same meaning. Normalization has the same goal too, that is to sychronize words of a data. Since the collected data were from social media, so the possibility of the words in the dataset to be slang is high. Here, normalization tries to transform the slang words to the standard word using [10]. Here is the example of normalization and case folding process.

Input *: Jokowi emang ga pantes jadi presiden. Mending lu mundur aja pak.*
Output *: jokowi memang tidak pantas jadi presiden. mending kamu mundur saja pak*

The next phase is filtering. This phase aims to eliminate not required words in dataset. Those characters are meaningless for classification. Since in the classification process, we only need the words that have a meaning to classify a data so in this process we eliminate characters such as punctuation, URL and RT (retweet). The example of this process is below.

Input : *Jokowi emang ga pantes jadi presiden. Mending lu mundur aja pak https://t.co/CZRyMMWstT*
Output : *jokowi memang tidak pantas jadi presiden mending kamu mundur saja pak*

After that, we will do stopword eliminating. As well as filtering, this phase aims to eliminate meaningless words. Those kind of words are such as *seperti, dan, atau, saja* and etc. Those words are considered as the unnecessary words that have no effect to decide a sentence to be hate speech or not. The example of this process is below.

Input : *Jokowi emang ga pantes jadi presiden. Mending lu mundur aja pak https://t.co/CZRyMMWstT*
Output : *jokowi tidak pantas presiden mundur pak*

The last phase of preprocessing is stemming. After our data is clean from meaningless words and characters, so the next phase is transforming the words to the basic words. There are affixes that might be constructed in a word, and this phase aims to eliminate those affixes. This process tries to synchronize every single word in dataset so it can be easily categorized as the same word on feature extraction process. As the example, word of "*memiliki*", "*dimiliki*", "*kepemilikan*" and "*pemilik*" will be transformed to the same word, which is "*milik*". Those words have the same root word, and this process will define it. If those words are not transformed to be the root word, they might be categorized as different words of each other though they have a same meaning. And that matter will affect to term frequency calculation later.

2.4. Feature extraction

Unigram is a method that would parse words one by one before input them to term frequency calculation. The aim of this process is to seek the most important syllable that can affect a sentence to be labelled as a hate speech. That thing can be discovered by its TF-IDF score resulted by one parsed syllable.

2.5. Classification

Classification process done by using data train in term weighting form as a result from feature extraction process. This process conducted by using machine learning as the tools for classifying sentiment and used data mining classification algorithm as the method to gain a model for predicting sentiment to be labelled as hate speech or not. The result of this process is classification model of algorithm used, that are Naïve Bayes, Decision Tree, Multi Level Perceptron (MLP), Support Vector Machine (SVM) and AdaBoost Classifier.

2.6. SMOTE

This process aims to balance the data, used in this study. Since the data was from social media so it is possible for having a data of a class more than data of other class. The goal is to improve classification model. At last, we conducted a comparison between model that used SMOTE and not. This process was not done by previous study [8]. To handle the imbalanced data on their study, they used under-sampling method which is different with SMOTE that used over sampling technique. At last, we compared the result model by previous study with this study that used SMOTE [8].

2.7. Evaluation

Experiments result conducted before will be evaluated in this process by using cross validation method to gain the best classification model. The evaluation was done by dividing dataset to be data train and data test, then set its fold experiment called k-fold cross validation. Results gained by this process were precision, recall and accuracy. Those scores were the base to define the best classification model. Since the goal of this study is to detect hate speech, so we decide to use recall score as the prime score to be compared.

3. Results and discussion

3.1. Experiment result

This study conducted 2 experiment scenarios which is comparison between model with SMOTE and non SMOTE together with the classification algorithms.

Table 1. Recall performance of algorithms.

Feature	*Recall*					
	AB	*MLP*	*MNB*	*BNB*	*SVM*	*DT*
SMOTE	61.8%	75.9%	75.3%	72.3%	40.8%	74.1%
Non-SMOTE	77.2%	77.5%	93.2%	78.6%	91.1%	77.4%

Table 1 shows the classification result of SMOTE and non SMOTE model of classification algorithm. The highest recall score belonged to model of Multinomial Naïve Bayes algorithm without SMOTE, as much as 93.2%. If we pay attention to this result, model of Multinomial Naïve Bayes algorithm showed consistent performance for any feature used by it. So, we could say that Multinomial Naïve Bayes was a stable recall performance model with any feature used even when we combined with SMOTE or non SMOTE. Following classification model with Support Vector Machine showed a good model without SMOTE, as much as 91.1%. However, if that model used SMOTE will decrease the model performance.

Table 2. Accuracy performance of algorithms.

Feature	Accuracy					
	AB	*MLP*	*MNB*	*BNB*	*SVM*	*DT*
SMOTE	72.0%	83.4%	73.2%	75.0%	49.2%	77.3%
Non-SMOTE	71.2%	69.1%	71.2%	71.2%	69.2%	70.3%

Table 2 shows classification performance of model using SMOTE and without SMOTE and accuracy score as the comparison. Based on Table 2, we could see that Multi Layer Perceptron algorithm with SMOTE had the highest accuracy score among the other, which was 83.4%.

3.2. Discussion

Based on the experiments conducted from this study, we could see that SMOTE did not affect too much to classification model performance. Model with SMOTE method achieved slightly higher accuracy score than model without SMOTE. Inversely with accuracy score, recall score of model without SMOTE showed significantly higher than model with SMOTE. It shows that to gain the best recall model, we need to use model without SMOTE. From the experiments, we also could see that MLP algorithm was the best model to classify a hate speech with accuracy score 83.4% and recall score 75.9% when using SMOTE.

After comparing our result with Mulia, Alfina, Fanany, & Ekanata and using the same evaluation variable which is F-measure, the highest F-measure score of our study achieved by Multinomial Naïve Bayes with score 83.3% when using all features unigram [8]. This score is slightly different with previous study that achieved 93.5% when using Random Forest Decision Tree and unigram-bigram features. Though we could not compare the recall score from previous study since it is not described. For our study, we prefer evaluate a classification model by its recall score.

4. Conclusion

This study conducts a comparison to find the best model to classify a hate speech. This study also results a new dataset that can be used for further study related to hate speech at social media. The dataset used in this study consists of 2776 hate and 1226 non hate speech labelled manually and agreed by annotator. Based on experiments of this study, we could see that MLP achieved the highest accuracy score (843.4%) when using all unigram and SMOTE. Multinomial Naïve Bayes algorithm also showed a good performance if we review by recall score. The highest recall score of MNB reached 93.2% when using all unigram without SMOTE. As the conclusion, we suggest Multinomial Naïve Bayes algorithm by using unigram feature without SMOTE as the best model to classify a hate speech.

References

[1] Agarwal S and Sureka A 2016 But I did not Mean It!- Intent Classification of Racist Posts on Tumblr *European Intelligence and Security Informatics Conference* IEEE

[2] Smith A G, Suedfeld P, and Conway L G 2008 The language of violence: distinguishing terrorist from nonterrorist groups by thematic content analysis *Dynamic Asymmetric Conflict* (pp. 1-10). Routledge.

[3] Nockleby J T 2000 Hate Speech. *Encyclopedia of the American Constitution* (pp. 1277-1279). Macmillan

[4] Warner W and Hirschberg J 2012 Detecting hate speech on the world wide web. *LSM '12 Proceedings of the Second Workshop on Language in Social Media* (pp. 19-26). Montreal, Canada: Association for Computational Linguistics.

[5] Statista 2016 *Number of active Twitter users in leading markets as of May 2016 (in millions)*. From https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/

[6] Waseem Z and Hovy D 2016 Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. *Proceedings of NAACL-HLT 2016* (pp. 88-93). San Diego, California: Association for Computational Linguistics

[7] Kwok I and Wang Y 2013 Locate the Hate: Detecting Tweets against Blacks. *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence* (pp. 1621-1622). Association for the Advancement of Artificial Intelligence

[8] Mulia R, Alfina I, Fanany M I, and Ekanata Y 2017 Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study *The 9th International Conference on Advanced Computer Science and Information Systems (ICACSIS 2017)*

[9] Pratiwi S H 2016 *Detection of Hate Speech against Religion on Tweet in the Indonesian Language Using Naïve Bayes Algorithm and Support Vector Machine* (Universitas Indonesia)

[10] KBBI Online 2017 *Kamus Besar Bahasa Indonesia (KBBI) Kamus versi online/daring (dalam jaringan)*. From https://kbbi.web.id/