ISBN: 978-981-18-1791-5
2021 the 11th International Workshop on Computer Science and Engineering (WCSE 2021)
doi: 10.18178/wcse.2021.02.011

Abusive Language and Hate Speech Detection for Javanese and Sundanese Languages in Tweets: Dataset and Preliminary Study

Shofianina Dwi Ananda Putri [1 +], Muhammad Okky Ibrohim [1] and Indra Budi [1]

[1]Faculty of Computer Science, Universitas Indonesia, Depok 16424, Indonesia

Abstract. Indonesia's demography as an archipelago with lots of tribes and local languages added variances in their communication style. Every region in Indonesia has its own distinct culture, accents, and languages. The demographical condition can influence the characteristic of the language used in social media, such as Twitter. It can be found that Indonesian uses their own local language for communicating and expressing their mind in tweets. Nowadays, research about identifying hate speech and abusive language has become an attractive and developing topic. Moreover, the research related to Indonesian local languages still rarely encountered. This paper analyzes the use of machine learning approaches such as Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest Decision Tree (RFDT) in detecting hate speech and abusive language in Sundanese and Javanese as Indonesian local languages. The classifiers were used with the several term weightings features, such as word n-grams and char n-grams. The experiments are evaluated using the F-measure. It achieves over 60 % for both local languages.

Keywords: abusive, hate speech, twitter, Indonesian Local language, Javanese, Sundanese

1. Introduction

People are now used to express their thoughts, daily activities, and shares pictures, which now cemented Indonesia as one of the highest percentages of social media usages in this modern day. Indonesia demography as an archipelago with lots of tribes and local languages added colours to Indonesian social media users' communication style due to its culture, accents, and languages [1]. Among the many existing social networks, Twitter currently ranks as one of the leading platforms and is one of the most critical data sources for researchers. The attraction is that Twitter data is more accessible than other social media platforms. Twitter is in a "people as sensor" network structure, in which there are interactions between Twitter users who react to external and social events, making it the most suitable medium for studying public opinion [2].

Twitter does not have rules that limit writing styles and usage of languages for a tweet, but the character use limit. However, Indonesia's demographical condition is able to influence the character of the language used in a tweet, which makes Indonesian can use their local language for communicating and expressing their mind. The use of regional languages poses a challenge in detecting abusive language and hate speech due to the limitation of resources related to local languages. Twitter is spotlighted as a social media platform containing a lot of abusive language and hate speech. Abusive language is an expression (both oral and text) containing abusive/dirty words or phrases in the context of jokes, a vulgar sex conversation, or cursing someone [3]. On the other hand, hate speech is defined as any communication that disparages a person or a group on the basis of some characteristic such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or other characteristics [4]. In Indonesia, Komnas HAM (National Commission on Human Rights) identify hate speech as an act, either directly or not based on hatred towards religion, belief, race, skin color, ethnicity, gender, people with disabilities, and sexual orientation, that contains incitement to individual or group through various media [5]. Hate speech is considered dangerous for various reasons such as condescending humanity, raising material losses and fatalities, potentially escalating conflicts, and genocides [5]. Indonesia considers hate speech contrary to state values. Thus, it becomes a concern arises to handle the spread of hate speech.

+ Corresponding author. Tel.: + (62)8567458885
E-mail address: shofianina.dwi91@ui.ac.id.

Nowadays, the research about identifying hate speech and abusive language has become an attractive and developing topic. Some studies are focused on detecting abusive language in monolingual, but the phenomenon of the use of local language becomes a challenge in itself for detection processes. Extracting Twitter information regarding abusive language and hate speech is essential to raising the attempt in managing the hate speech and abusive language spread in social media. This paper builds a valid dataset for hate speech and abusive language detection using Sundanese and Javanese Twitter datasets. We then conduct a preliminary experiment to find the best result using several machine learning approaches, which use several classifiers and features as preliminary performance. This paper has 5 sections and is organized as follows. Section 2 sheds some light on related work about the hate speech and abusive language detection studies and section 3 gives details about the dataset and methodology. Result and analysis are described in section 4. The last, the conclusion of our paper is in the Conclusion & Future Work section.

2. Related Work

Many researches have been done on abusive and hate speech detection in social media using a machine learning approach. This approach is commonly used and proven to perform well with high accuracy for classification tasks [6-10]. The abusive language and hate speech detection in some language has been done by [6-7]. In [7], they evaluate the hate speech detection across Facebook for the Amharic language using Naïve Bayes and Random forest algorithms, while for Thai language was done by [8] using NB, k-Nearest Neighbor (kNN), SVM, RFDT. For Indonesian language, [9] starts the study in detecting hate speech through social media. The research analyzes the Indonesian bullying words on Twitter and did not use any machine learning approaches. They found that *"bangsat"* (bastard) and *"anjing"* (bitch) are the most used in bullying word patterns in Indonesian Twitter. Hate speech detection using a machine learning approach to evaluate the performance was done by [6] and [10]. NB, SVM, Linear Regression (LR), and RFDT were used as the classifier with several word n-gram and character n-gram features. The result shows that the word n-gram has better performance with an $F-Measure$ of 93.5%. Meanwhile, [6] has an $F-Measure$ of 80.71% using LR with hate code binary and hate code dictionary feature.

The previous research has been done a preliminaries research on abusive language identification in Twitter [1]. The studies were performed using a machine learning approach: NB, SVM, and RFDT with word n-grams and character n-grams features. The result shows that all classifiers show above 80% of $F-Measure$. For this study, the retrieved data from Twitter is still in Bahasa Indonesia. Our study's conditions resulted in improved research regarding hate speech detection for texts containing the local Indonesian language. The diversity of culture and language in Indonesia enables social media user to use their distinctive local language to form sentences which contain abusive and cursing words. It affected the process of detecting offensive and abusive language negatively, making it harder to do. In this research, we evaluate the use of Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest Decision Tree (RFDT) with word n-gram and character n-gram features in detecting hate speech and abusive language in local Indonesian languages [3]. The Javanese and Sundanese were chosen as the local Indonesian language that was evaluated in this study. The use of these languages due to a large number of speakers in Indonesia [11]. Based on our research, there are limited resources available that provide local Indonesian language datasets for hate speech and abusive language detection. This paper aims to build a valid dataset that could serve as an initial source for future development on abusive language and hate speech research in Indonesia. The dataset is available on GitHub[1].

3. Dataset and Methodology

3.1. Dataset

The Indonesian local language dataset collection was conducted using Twitter search API[2] to collect the tweets and then implemented using Tweepy Library[3]. The tweets were collected using queries from the list of abusive words in Indonesian tweets, which have been done by [3]. The abusive words were translated into

[1] https://github.com/Shofianina/local-indonesian-abusive-hate-speech-dataset
[2] https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets
[3] https://www.tweepy.org/

local Indonesian languages, which are Javanese and Sundanese. The translated words are then used as queries to collect tweets containing Indonesian and local languages. The translation process involved native speakers for each local language. The crawling process has collected a total of more than 5000 tweets. Then, the crawled data were filtered to get tweets that contain local's vocabulary and/or sentences in Javanese and Sundanese. Next, after the filtering process, the data will be labeled whether the tweets are labeled as hate speech and abusive language or not. Some examples of tweets from the dataset can be seen below:

T1: *"@USER gua juga kaget asu"*

Translation: "@USER I was also shocked asu" (asu is Javanese which means dog)

T2: "@USER @USER *Ngajak gelut manehna bagong"*

Translation: "@USER @USER you're asking for fight, Bagong" (bagong is Sundanese which means pig)

T3: "@USER @USER *Nambahin sedikit, punakawan yg paling kita kenal adalah Semar, Gareng, Petruk, Bagong*"

Translation: "@USER @USER Added a little, the puppets we know the most are Semar, Gareng, Petruk, Bagong". (Bagong is the name of puppet character from Java Traditional art)

From the examples above, it is recognized that translated abusive words do not necessarily contain abusive tweets. The **T1** used the abusive word to strengthen the tweet about their incredible feeling toward something that has happened. The **T3** tweet talked about a puppet character who coincidentally has the same name as an animal, often used as an abusive word. On the examples, it shows that local language can be a tool to express hate speech and abusive language, either to disguise the speech intended or to make it easier for the interlocutors to understand the meaning of their tweet.

Based on the example above, it can be seen that each region has its characteristics and uniqueness in language. Thus, in the annotation process, people who have a good understanding of the local language are needed. The Javanese and Sundanese were annotated manually by annotator from each region. The annotation process involved multiple-step processes. It was carried by two annotators for each language, after an initial step where the guidelines were discussed and refined to reach unanimous comprehension. The annotation process gives 3449 and 2207 tweets for Javanese and Sundanese dataset respectively with 100% agreement. To be more specific, we measure the agreement coefficient using Cohen's Kappa. It achieves 0.44 and 0.46 of Cohen's Kappa coefficients for Javanese and Sundanese respectively. According to [12], if the mean value of Cohen Kappa for each annotation label is greater than 0.20 then the annotation result is valid, where Cohen's Kappa's highest score is acquired from the Sundanese dataset. The obtained Cohen's Kappa value indicates that the dataset is valid and reliable to research.

3.2. Methodology

As our extensive research, we use a machine learning approach to find which algorithms show the best performance. The methodology consists of preprocessing, feature extraction, classification, and evaluation. After the annotation process, we do preprocess to clean the labeled data. We adopt some preprocessing methods, which are case-folding to change all the text in low case letters. We remove several attributes, such as unnecessary characters, username, re-tweet (RT), emoticon, punctuation, hashtag, and uniform resource locator (URL). In feature extraction process, we use word n-gram and the combination of word n-grams, which are Unigram-Bigram, Unigram-Bigram-Trigram, and Bigram-Trigram. We also use character n-grams features, where n vary from 3 (Trigram) to 4 (Quadgram), and the combination of Trigram and Quadgram.

In this experiment, we use Naïve Bayes, Support Vector Machine, and Random Forest Decision Tree classifiers for the classification task. These three algorithms are widely used as a baseline in the text classification process. All three have the advantages of being easy to implement and show good performance in several studies related to identifying abusive language and hate speech topics in various languages. In this paper, we classify the tweet into two labels, which are abusive language and hate speech. The use of two labels is based on previous research findings that abusive words do not necessarily mean hate speech and vice versa. As for the multi-label data conversion, we use the label power-set (LP) method. It transforms the multi label data into unique label multi class classification problems.

To find the best model and feature combination, we apply three different classifiers that are NB, SVM, and RFDT. The Naive Bayes algorithm is a classification algorithm based on the Bayes Theorem, which assumes that each feature is independent and calculates each class's probability, where the highest probability result as the most likely classification. SVM works to find the best hyperplane that can separate different classes in input space [13], and RFDT is an ensemble method that combines several decision trees and use the majority voting to determine the decision. The classification was evaluated using the 5-fold-cross-validation[4]. This method will divided the data into five parts: the four parts of the data will act as the training data, and the rest part of the data is for the testing data. The process will be done in five times. So, each data will be the training and testing data at the same time. In the end, the performance was evaluated by calculating the $F - Measure$.

4. Result and Discussion

Tabel 1: $F - Measure$ for Javanese and Sundanese Dataset Evaluation

	Javanese			Sundanese		
	NB	SVM	RFDT	NB	SVM	RFDT
Word Unigram	0.752	0.778	0.762	0.794	0.820	0.819
Word Bigram	0.627	0.709	0.680	0.802	0.807	0.807
Word Trigram	0.627	0.641	0.628	0.802	0.807	0.807
Word Unigram+Bigram	0.750	0.780	0.771	0.800	0.816	0.816
Word Bigrams+Trigram	0.624	0.675	0.665	0.799	0.807	0.807
Word Unigram+Bigram+Trigram	0.750	0.780	0.755	0.800	0.816	0.816
Char Trigram	0.726	0.709	0.711	0.807	0.819	0.819
Char Quadgram	0.743	0.752	0.758	0.799	0.818	0.820
Char Trigram+Quadgram	0.708	0.660	0.702	0.799	0.819	0.819

The results were shown in Table 1 for the Javanese and Sundanese dataset as the performance evaluation matrix. For the Javanese dataset, the evaluation results show that the best $F - Measure$ value is obtained with the word n-gram feature, with the best feature combination using Unigram+Bigram and Unigram+Bigram+Trigram. The highest $F - Measure$ value is obtained at 0.780 with SVM classifier. Meanwhile, the character n-gram and its combination show competitive value of $F - Measure$. Based on the results in Table 1, the value of $F - Measure$ can increase along with the increasing n in character n-gram features. However, the combination of the two causes a decrease in the value of $F - Measure$ for all classifier. Therefore, the use of the combination of the n-gram characters is not recommended in this dataset.

As for the Sundanese dataset, the word n-gram and character n-gram feature extractions indicate slight difference on $F - Measure$ values. The combination for each word n-gram and character n grams did not show a significant impact on increasing the F-Measure. The highest $F - Measure$ value is achieved at 0.82. This value can be found in the word unigram with SVM classifier and character Quadgram feature with RFDT classifier. For all classifiers, it can be found that the use of Bigram+Trigram produced lower $F - Measure$ than other combination in word n-gram features. In other hand, the use of Unigram+Bigram and the combination of all achieved higher value of $F - Measure$.

According to the $F - Measure$, SVM works well for Javanese dataset. Meanwhile, for Sundanese dataset, it is shown that SVM perform as well as RFDT. These results are consistent with several previous studies that used SVM and RFDT as an outperformed classifier compared to Naïve Bayes in classification task [14] [15]. For future works, we suggest the use of SVM and consider the use of RFDT with word n-gram and character n-gram features to detect abusive language and hate speech in Indonesian local languages dataset.

5. Conclusion and Future Work

In this paper, we try to collect twitter datasets that used Sundanese and Javanese as Indonesian local languages to detect abusive language and hate speech. To evaluate performances, we use word n-gram combination and character n-gram as feature extractions, with NB, SVM, and RFDT as the classifiers. The

[4] https://scikit-learn.org/

result shows that SVM for Javanese dataset with word n-gram features achieves better performance than others. As for Sundanese dataset, the word and character n-gram features show good performance with SVM and RFDT classifiers. The dataset is a valid and can be used for research with an approval of Cohen's Kappa value more than 0.4. This study tries to use different methods to find a model that can detect abusive language and hate speech in local language. For future work, this may be done using classifiers and other feature extraction techniques to improve the detection process's performance, such as deep learning approach. However, since deep learning approach requires a more considerable amount of data, so the dataset needs to be improved and consider codemixed issues in the text.

6. Acknowledgements

This work was supported by the PUTI Prosiding research grant NKB-3486/UN2.RST/HKP.05.00/2020 from Directorate Research and Community Services, Universitas Indonesia

7. References

[1] M. Hayaty, S. Adi, dan A. D. Hartanto, "Lexicon-Based Indonesian Local Language Abusive Words Dictionary to Detect Hate Speech in Social Media," *J. Inf. Syst. Eng. Bus. Intell.*, vol. 6, no. 1, hal. 9, 2020.

[2] K. Philander dan Y. Y. Zhong, "Twitter sentiment analysis: Capturing sentiment from integrated resort tweets," *Int. J. Hosp. Manag.*, vol. 55, no. May, hal. 16–24, 2016.

[3] M. O. Ibrohim dan I. Budi, "A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media," *Procedia Comput. Sci.*, vol. 135, hal. 222–229, 2018.

[4] A. Schmidt dan M. Wiegand, "A Survey on Hate Speech Detection using Natural Language Processing," in *Proceedings ofthe Fifth International Workshop on Natural Language Processing for Social Media*, 2017, no. 2012, hal. 1–10.

[5] Komnas HAM, *Buku Saku Ujaran Kebencian (Hate Speech)*. Komisi Nasional Hak Asasi Manusia, Republik Indonesia.

[6] N. I. Pratiwit, I. Budi, dan M. A. Jiwanggi, "Hate speech identification using the hate codes for Indonesian tweets," in *2nd International Conference on Data Science and Information Technology, DSIT 2019*, 2019, hal. 128–133.

[7] Z. Mossie dan J.-H. Wang, "Social Network Hate Speech Detection for Amharic Language," *Comput. Sci. Inf. Technol.*, hal. 41–55, 2018.

[8] S. Tuarob dan J. L. Mitrpanont, "Automatic Discovery of Abusive Thai Language Usages in Social Networks," in *International Conference on Asian Digital Libraries*, 2017, hal. 267–278.

[9] H. Margono, X. Yi, dan G. Raikundalia, "Mining Indonesian cyber bullying patterns in social networks," in *Proceedings of the Thirty-Seventh Australasian Computer Science Conference*, 2014.

[10] I. A. Ekanata, R. Mulia, M. I. Fanany, dan Yudo, "Hate speech detection in the Indonesian language: A dataset and preliminary study," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, hal. 233–238.

[11] M. S. Saputri dan M. Adriani, "Identifying Indonesian local languages on spontaneous speech data," in *2019 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2019*, 2019, hal. 247–254.

[12] A. J. Viera dan J. M. Garrett, "Anthony J. Viera, MD; Joanne M. Garrett, PhD (2005). Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360-63.," *Fam. Med.*, vol. 37, no. 5, hal. 360–3, 2005.

[13] E. K. Andana, M. Othman, dan R. Ibrahim, "Comparative analysis of text classification using naive bayes and support vector machine in detecting negative content in Indonesian twitter," *Int. J. Adv. Trends Comput. Sci. Eng.*, vol. 8, no. 1.3 S1, hal. 356–362, 2019.

[14] I. Alfina, R. Mulia, M. I. Fanany, dan Y. Ekanata, "Hate speech detection in the Indonesian language: A dataset and preliminary study," *2017 Int. Conf. Adv. Comput. Sci. Inf. Syst. ICACSIS 2017*, vol. 2018-Janua, no. October, hal. 233–237, 2018.

[15] T. Pranckevičius dan V. Marcinkevičius, "Comparison of Naive Bayes, Random Forest, Decision Tree, Support Vector Machines, and Logistic Regression Classifiers for Text Reviews Classification," *Balt. J. Mod. Comput.*, vol. 5, no. 2, 2017.