Multi-label Hate Speech and Abusive Language Detection in Indonesian Twitter

Muhammad Okky Ibrohim
Faculty of Computer Science
Universitas Indonesia
Kampus UI, Depok, 16424, Indonesia
okkyibrohim@cs.ui.ac.id

Indra Budi
Faculty of Computer Science
Universitas Indonesia
Kampus UI, Depok, 16424, Indonesia
indra@cs.ui.ac.id

Abstract

Hate speech and abusive language spreading on social media need to be detected automatically to avoid conflicts between citizens. Moreover, hate speech has a target, category, and level that also need to be detected to help the authority in prioritizing which hate speech must be addressed immediately. This research discusses multi-label text classification for abusive language and hate speech detection including detecting the target, category, and level of hate speech in Indonesian Twitter using machine learning approaches with Support Vector Machine (SVM), Naive Bayes (NB), and Random Forest Decision Tree (RFDT) classifier and Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) as the data transformation method. We used several kinds of feature extractions which are term frequency, orthography, and lexicon features. Our experiment results show that in general the RFDT classifier using LP as the transformation method gives the best accuracy with fast computational time.

1 Introduction

Hate speech is a direct or indirect speech toward a person or group containing hatred based on something inherent to that person or group (Komnas HAM, 2015)[1]. Factors that are often used as bases of hatred include ethnicity, religion, disability, gender, and sexual orientation. Hate speech spreading is a very dangerous action which can have some negative effects such as discrimination, social conflict, and even human genocide (Komnas HAM, 2015). One of the most horrific genocides caused by the act of spreading hate speech was the Tutsi ethnic genocide in Rwanda in 1994 (Stanton, 2009). The cause of the tragedy was hate speech propagated by some groups, claiming that the cause of increasing pressure in politics, economic and social was the Tutsi ethnic.

In everyday life, especially in social media, the hate speech spreading is often accompanied with abusive language (Davidson et al., 2017). Abusive language is an utterance that contains abusive words/phrases that is conveyed to the interlocutor (individuals or groups), both verbally and in writing. Hate speech that contains abusive words/phrases often accelerates the occurrence of social conflict because of the use of the abusive words/phrases that triggers emotions. In Indonesia, abusive words are usually derived from an unpleasant condition such as mental disorder, sexual deviation, physical disability, lack of modernization, a condition where someone does not have etiquette, conditions that is not allowed by religion, and other conditions related to unfortunate circumstances; animals that have a bad characteristic, disgusting, and forbidden in certain religion; astral beings that often interfere with human life; a dirty and bad smell object; a part of the body and an activity that related to sexual activity; and low-class profession that is forbidden by religion (Wijana and Rohmadi., 2010; Ibrohim and Budi, 2018). In general, the use of abusive words aimed to curse someone (spreading hate speech) in Indonesia is divided into three types that are words, phrases, and clauses (Wijana and Rohmadi., 2010). The spread of hate speech that is accompanied with abusive language often accelerates the occurrence of social conflict because of the use of the abusive words/phrases that triggers emotions. Although abusive language are sometimes just being used as jokes (not to offend someone), the use of abusive language in social media still can lead to conflict because of misunderstand-

[1]Komisi Nasional Hak Asasi Manusia (Komnas HAM) is an independent institution that functions to carry out studies, research, counseling, monitoring, and mediation of human rights in Indonesia. See https://www.komnasham.go.id/index.php/about/1/tentang-komnas-ham.html

Proceedings of the Third Workshop on Abusive Language Online, pages 46–57
Florence, Italy, August 1, 2019.

ings among netizens (Yenala et al., 2017). Moreover, children could be exposed to language inappropriate for their age from those abusive language scattered in their social media (Chen et al., 2012).

The hate speech and abusive language on social media must be detected to avoid conflicts between citizens and children learning the hate speech and inappropriate language from the social media they use (Komnas HAM, 2015; Chen et al., 2012). In recent years, many researchers have done research in hate speech detection (Waseem and Hovy, 2016; Alfina et al., 2017, 2018; Putri, 2018; Vigna et al., 2017) and abusive language detection (Turaob and Mitrpanont, 2017; Chen et al., 2012; Nobata et al., 2016; Ibrohim and Budi, 2018; Ibrohim et al., 2018) in various social media genres and languages.

According to (Komnas HAM, 2015), a hate speech has a certain target, category, and level. Hate speeches can belong to a certain category such as ethnicity, religion, race, sexual orientation, etc. that are targeted to a particular individual or group with a certain level of hatred. However, based on our literature study, there has been no research on abusive language and hate speech detection including the detection of hate speech target, category, and level conducted simultaneously. Many research in hate speech detection (Waseem and Hovy, 2016; Alfina et al., 2017, 2018; Putri, 2018) just identifying whether a text is hate speech or not. In 2017, (Vigna et al., 2017) performed research on hate speech level detection. Their research was done to classifying Italian Facebook post and comment into three labels which are *no hate speech*, *weak hate speech*, and *strong hate speech*. However, (Vigna et al., 2017) did not classifying the target and category of hate speech. Similar to research in hate speech detection, many studies in abusive language detection (Turaob and Mitrpanont, 2017; Chen et al., 2012; Nobata et al., 2016) also just identify whether a text is abusive language or not. In 2018, (Ibrohim and Budi, 2018) conducted research on hate speech and abusive language detection. Their research was done to classify Indonesian tweet into three labels that are *no hate speech*, *abusive but no hate speech*, and *abusive and hate speech*. However, same as other studeis on hate speech and abusive language detection, (Ibrohim and Budi, 2018) did not classify the target and category of hate speech.

Depending on (Hernanto and Jeihan, 2018)[2][3], detection of the hate speech target, category, and level is important to help authorities prioritize cases of hate speech that must be handled immediately. In this work, we do research on hate speech and abusive language detection in Indonesian Twitter. We chose Twitter as our dataset because Twitter is one of the social media platforms in Indonesia that is often used to spread the hate speech and abusive language (Alfina et al., 2017, 2018; Putri, 2018; Ibrohim and Budi, 2018; Ibrohim et al., 2018). This problem is a multi-label text classification problem, where a tweet can be *no hate speech, no hate speech but abusive, hate speech but no abusive,* and *hate speech and abusive*. Furthermore, hate speech also has a certain target, category, and level.

In doing multi-label hate speech and abusive language detection, we use machine learning approach with several classifiers. The classifiers that we use include Support Vector Machine (SVM), Nave Bayes (NB), and Random Forest Decision Tree (RFDT) using problem transformation methods including Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC). Based on several previous works, these three classifiers are algorithms that can produce pretty good performance for hate speech and abusive language detection in Indonesian (Alfina et al., 2017, 2018; Putri, 2018; Ibrohim and Budi, 2018; Ibrohim et al., 2018). We used several kinds of text classification features including term frequency (word n-grams and character n-grams), orthography (exclamation mark, question mark, uppercase, and lowercase), and sentiment lexicon (negative, positive, and abusive). We use accuracy for evaluating our proposed approach (Kafrawy et al., 2015). To validate our experiment results, we use 10-fold cross validation technique (Kohavi, 1995).

In this paper, we built an Indonesian Twitter dataset for abusive language and hate speech detection including detecting the target, category, and level of hate speech. In general, the contributions of this research are:

- Analyzing the target, category, and level of

[2]Staff of Direktorat Tindak Pidana Siber Bareskrim Polri

[3]Direktorat Tindak Pidana Siber Badan Reserse Kriminal Kepolisian Negara Republik Indonesia (Bareskrim Polri) is a directorate of the Indonesian national police that charge of fostering and carrying out the function of investigating and investigating cyber crimes in Indonesia. See https://humas.polri.go.id/category/satker/cyber-crime-bareskrim-polri/

hate speech to make an annotator guide and gold standard annotation for building Indonesian hate speech and abusive language dataset. Our annotator is arranged based on (Komnas HAM, 2015) and the results of interviews and discussions with the staff of Direktorat Tindak Pidana Siber Bareskrim Polri (Hernanto and Jeihan, 2018) and a linguistic expert (Nurasijah, 2018).

- Building a dataset for abusive language and hate speech detection including detecting the target, category, and level of hate speech in Indonesian Twitter. We provide this research dataset for public[4] so that it can be used by other researchers who are interested in doing future work of this paper.

- Conducting preliminaries experiments on multi-label abusive language and hate speech detection (including hate speech target, category, and level detection) in Indonesian Twitter using machine learning approaches.

This paper is organized as follows. We discuss hate speech target, category, and level in Indonesia in Section 2. Our data collection and annotation process is described in Section 3. Section 4 presenting our experiment results and discussion. Finally, the conclusions and future work of our research are presented in Section 5.

2 Hate Speech Target, Categories, and Level in Indonesia

In this research, we conducted Focus Group Discussion (FGD) with the staff of Direktorat Tindak Pidana Siber Badan Reserse Kriminal Kepolisian Negara Republik Indonesia (Bareskrim Polri), which is the agency responsible for investigating cybercrimes in Indonesia. This is done in order to get a valid definition of hate speech, including the characterization. From the FGD with staff of Bareskrim Polri (Hernanto and Jeihan, 2018), it was obtained that hate speech has a particular target, categories, and level.

Every hate speech is aimed at a particular target. In general, the target of hate speech is divided into two kinds, which are *individual* and *group*. Hate speech with individual target is hate speech that aimed at someone (an individual person), while hate speech with group target is hate speech that aimed at a particular groups, associations, or communities. These groups, associations, and communities can be in the form of religious groups, races, politics, fan clubs, hobby communities, etc.

Both aimed at individual or group, hate speech has a particular category as the basis of hate. According to FGD results, in general, hate speech categories are as follows:

1. *Religion/creed*, which is hate speech based on a religion (Islam, Christian, Catholic, etc.), religious organization/stream, or a particular creed;

2. *Race/ethnicity*, which is hate speech based on a human race (human groups based on physical characteristics such as face shape, height, skin color, and others) or ethnicity (human groups based on general citizenship or shared cultural traditions in a geographical area);

3. *Physical/disability*, which is hate speech based on physical deficiencies/differences (e.g. shape of face, eye, and other body parts) or disability (e.g. autism, idiot, blind, deaf, etc.), either just cursing someone (or a group) with those words related to physical/disability or those that are truly experienced by those who are the target of the hate speech;

4. *Gender/sexual orientation*, which is hate speech based on gender (male and female), cursing someone (or a group) using words that are degrading to gender (e.g.: gigolo, bitch, etc.), or deviant sexual orientation (e.g.: homosexual, lesbian, etc.);

5. *Other invective/slander*, which is hate speech in the form of swearing/ridicule using crude words/phrases or other slanders/incitement which are not related to the four groups previously explained.

Notice that a hate speech can be categorized in several categories at once except *other invective/slander* category. In other words, a hate speech under category *religion/creed, race/ethnicity, physical/disability,* and *gender/sexual orientation* can not be categorized as *other invective/slander* category, and vice versa.

[4] https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection

Besides having targets and categories, hate speech also has a certain level. Based on the FGD results, we divide hate speech into three levels, which are *weak, moderate,* and *strong*. The explanation for every level of hate speech are as follows:

1. *Weak hate speech*, which is hate speech in the form of swearing/slanders that aimed at individuals without including incitement/provocation to bring open conflict. In Indonesia, hate speech in this form categorized as *weak hate speech* because it is a personal problem. It means, if the target of hate speech does not report to the authorities (feeling ordinary and forgiving people who spread the hate speech towards him) then that hate speech is not too prioritized to be resolved by the authorities.

2. *Moderate hate speech*, which is hate speech in the form of swearing/blasphemy/stereotyping/labeling aimed at groups without including incitement/provocation to bring open conflict. Although it can invite conflict between groups, this kind of hate speech is belonging to moderate hate speech because the conflict that will occur is estimated to be limited to conflict on social media.

3. *Strong hate speech*, which is hate speech in the form of swearing/slanders/blasphemy/stereotyping/labeling aimed at individual or group including incitement/provocation to bring open conflict. This kind of hate speech is belonging to strong hate speech, because it is a hate speech that needs to be prioritized to be resolved soon because it can invite conflicts that are widespread and can lead to conflicts/physical destruction in the real world.

3 Data Collection and Annotation

In this research, we used hate speech and abusive language Twitter dataset from several previous researches consisting of (Alfina et al., 2017, 2018), (Putri, 2018), and (Ibrohim and Budi, 2018). Besides using Twitter dataset from previous researches, we also crawled tweets in order to enrich dataset such that it can include the kinds of writing of hate speech and abusive language that may not yet exist in the data from previous researches. We crawled Twitter data using Twitter Search API[5] which is implemented using Tweepy Library[6]. The queries which we used for crawling Twitter data are words/phrases that often used by netizens when spreading hate speech and abusive language in Indonesian social media, that can be seen in Appendix 1[7]. We crawled the twitter data for about 7 months, from March 20th, 2018 until September 10th, 2018. The purpose of crawling with a long time is to get more tweet writing patterns.

In this research, we used crowdsourcing with a paid mechanism (Sabou et al., 2014) for the annotation process. Since the tweets that we want to annotate has many labels, we decided to conduct two phases of annotation process. This is because annotators who are not linguistic experts should not annotate data with too many labels (Sabou et al., 2014). The first phase annotation process was done to annotate the Twitter data whether tweets are hate speech and abusive language or not, while the second phase annotation process was done to annotate the hate speech target, categories, and level. For tweets from (Alfina et al., 2017, 2018) and (Putri, 2018), tweets were just annotated to determine whether the tweet is an abusive language or not in the first phase annotation process, since the hate speech label is already obtained. Meanwhile, tweets from (Ibrohim and Budi, 2018) can be annotated directly in the second phase since their dataset was annotated for hate speech and abusive labels.

For the annotation process, we built a web based annotation system in order to make it easy for the annotators to annotate data so that it can speed up the annotation process and minimize annotation errors. We conducted an annotator guideline to give the task definition and example for helping the annotators in understanding the annotation task. We also conducted a gold standard annotation for testing whether the annotators already understand the task or not. In this research, we are doing a discussion and consultation with an expert

[5] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

[6] http://www.tweepy.org/

[7] For complete list of queries, see https://github.com/okkyibrohim/id-multi-label-hate-speech-and-abusive-language-detection

linguistic (Nurasijah, 2018)[8] in order to get a valid annotation guideline and gold standard annotation. Twitter data that used for the gold standard came from previous research (Alfina et al., 2017; Ibrohim and Budi, 2018) and hate speech handbook (Komnas HAM, 2015).

After the annotation system was built and tested well, the next process is the annotator recruitment process. In this research, annotators came from different religious, racial/ethnic, and residential backgrounds. This is done to reduce bias because the annotation of hate speech is quite subjective (Alfina et al., 2017). The selected annotators' criteria are as follows: (a) have the age of 20-30 years old (since most Twitter users in Indonesia came from that age (APJII., 2017)); (b) native in the Indonesian language (Bahasa Indonesia); (c) experienced using Twitter; (d) not members of any political party/organization (this is done to reduce annotation bias, especially when the annotators annotate tweets that are related to politics).

In this research, we use 30 annotators to annotate our dataset from various demographic background. The annotators consist of 14 males and 16 females that have various age (25 annotators aged 20-24 years and 5 annotators aged 25-32 years) and last education background (12 annotators have bachelor degree for last education and 18 annotators have senior high school degree for last education). Furthermore, annotators also come from various jobs, ethnicities, and religions. The kind of annotators' jobs consists of bachelor students (12 annotators), master students (3 annotators), civil servants (1 annotator), honorary employees (1 annotator), teacher/tutor/teaching assistant (5 annotators), and private employees (8 annotators); the annotators' origin ethnic consists of Java (11 annotators), Bali (4 annotators), Tionghoa (4 annotators), Betawi (3 annotators), Batak (2 annotators), and others (6 annotators, came from Melayu, Minang, Sunda, Cirebon, Ambon, Toraja); and the annotator's religion consists of Islam (15 annotators), Christian (5 annotators), Catholic (5 annotators), Hindu (3 annotators), and Buddha (2 annotators).

In the first annotation phase, we collect 16,500 tweets from the crawling process and previous researches (Alfina et al., 2017, 2018; Putri, 2018) to be annotated by those 30 annotators. Every tweet was annotated by 3 annotators and the final label was decided using 100% agreement technique. From this phase, we get 11,292 (68.44% total tweets that were annotated in the first phase) consisting of 6,187 not hate speech tweets and 5,105 hate speech tweets that have 100% agreement (reliable dataset). According to (McHugh, 2012), this percentage amount of reliable dataset (data can be used for research experiment) shows that the annotation result has a good level of agreement.

Next, in the second annotation phase, we annotated 5,700 hate speech tweets (5,105 tweets from the first phase annotation and 595 tweets from (Ibrohim and Budi, 2018)). In this phase, we use the best three annotators from the first annotation phase to annotate the target, categories, and level of hate speech. The final label in this phase was decided using majority voting. Since we use 3 annotators, each tweet label must have a minimum agreement from two annotators. If there is no agreement among the annotators in giving the label, then the tweet is deleted. From the second phase annotations results, there were 139 tweets that were deleted because there was no agreement in hate speech categories or hate speech level labels. Therefore, we get 5,561 reliable data (97.56% from total tweets that annotated in the second phase) that can be used for the research experiment. According to (McHugh, 2012), this percentage amount of reliable dataset shows that the annotation result has a almost perfect level of agreement.

From these two phase annotation process, we get 13,169 tweets already used for research experiments that consist of 7,608 not hate speech tweets (6,187 tweets from the first phase annotation and 1,421 tweets from (Ibrohim and Budi, 2018)) and 5,561 hate speech tweets. The distribution of abusive language towards not hate speech tweets and hate speech tweets from the collected tweets can be seen in Figure 1. From Figure 1, we can see that not all hate speech is abusive language. On the contrary, an abusive language also not necessarily a hate speech.

From the total 5,561 hate speech tweets we have, most of that hate speech tweets are directed at individuals (3,575 tweets targeted to an individual and 1,986 tweets targeted to a group). Those hate speech tweets consist of several hate speech categories which are 793 tweets related to religion/creed, 566 tweets re-

[8]Master in sociolinguistics

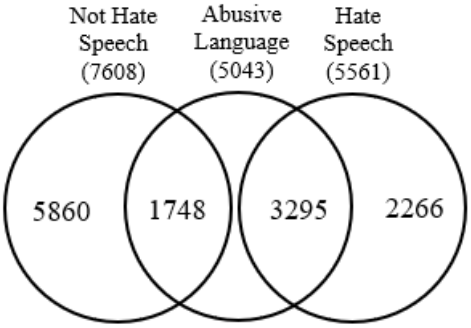

Figure 1: Distribution of abusive language towards not hate speech tweets and hate speech tweets

lated to race/ethnicity, 323 tweets related to physical/disability, 306 tweets related to gender/sexual orientation, and 3,740 tweets related to other invective/slander. Notice that the hate speech categories of religion/creed, race/ethnicity, physical/disability, and gender/sexual orientation are multi-label. It means, a tweet of hate speech can be related in several categories. Meanwhile, for the hate speech level labels, our hate speech dataset consists of 3,383 weak hate speech, 1,705 moderate hate speech, and 473 strong hate speech.

4 Experiments and Discussions

We conduct two scenarios for the experiment. The first experiment scenario uses multi-label classification to identify abusive language and hate speech including the target, categories, and level that contained in a tweet. Meanwhile, the second scenario uses multi-label classification to identify abusive language and hate speech that contained in a tweet without identifying the target, categories, and level of hate speech. Both of these scenarios are performed to find out the best classifier, transformation method, and features for each scenario.

In general, both the first scenario and the second scenario have the same flow that can be seen in Figure 2.

First, we do data preprocessing in order to make classification process more efficient and gives better results. We do five processes in data preprocessing consists of case folding, data cleaning, text normalization, stemming, and stop words removal. Case folding was done to make all character in lower case in order to standardize character case. Next, data cleaning was done to remove unnecessary characters such as re-tweet symbol (RT), username, URL, and punctuation. Since we do not use emoticon for feature extraction, we also

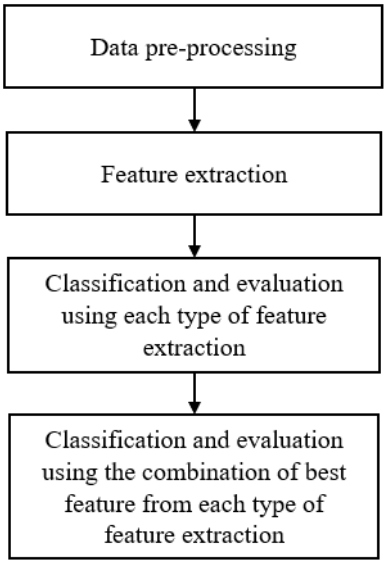

Figure 2: The experiment flowchart

remove emoticon in data cleaning process. After that, we do text normalization, which is changing non-formal words into formal ones. In this research, we do text normalization simply using dictionary obtained from the combination dictionaries from several previous works (Alfina et al., 2017; Ibrohim and Budi, 2018; Salsabila et al., 2018) and the dictionary that we build based on our dataset. Next, we do stemming to lemmatize words in every tweet. In this paper, stemming was done using Nazief-Adriani Algorithm (Adriani et al., 2007) that implemented using Sastrawi Library[9]. For stop words removal, we used stop word list given by (Tala, 2003).

The next step after data preprocessing is feature extraction. In this research, we used several kinds of feature extractions which are *term frequency, orthography* and *lexicon* features. Term frequency features that we used in our experiments consist of *word n-grams* (unigram, bigrams, trigrams, and the combination of word unigram, bigrams, and trigrams) and *character n-grams* (trigrams, quadgrams, and the combination of character trigrams and quadgrams). For the orthography feature, we used the number of *exclamation mark, question mark, uppercase* and *lowercase*. Meanwhile, for the lexicon features, we used *sentiment lexicon* (negative and positive sentiment) given by (Koto and Rahmaningtyas, 2017) and *abusive lexicon* that we built ourselves compiled from abusive words that used as queries when crawling Twit-

[9] https://github.com/har07/PySastrawi

ter data. After the feature extraction process was done, the dataset is ready for classification process.

For the classifier, we used three machine learning classification algorithms which are Naive Bayes (NB), Support Vector Machine (SVM), and Random Forest Decision Tree (RFDT). Based on the previous works (Alfina et al., 2017, 2018; Putri, 2018; Ibrohim and Budi, 2018), those three algorithms can give a pretty good performance in doing hate speech and abusive language detection in Bahasa Indonesia (Indonesian language). Notice that these three classifiers are single label output classifiers. It means, those three classifiers cannot solve multi-label text classification directly. To overcome this problem, we applied data transformation method (Kafrawy et al., 2015) such that the classifiers that we use can solve multi-label text classification problem. We used three data transformation methods that are Binary Relevance (BR), Label Power-set (LP), and Classifier Chains (CC) (Kafrawy et al., 2015). In doing classification, we do classification using each type of feature extraction first. After that, we do classification using the combination of best feature from each type of feature extraction. For the evaluation, we used 10-fold cross-validation technique (Kohavi, 1995) with accuracy as the metric evaluation. Accuracy in this research is calculated using formula as follow (Kafrawy et al., 2015):

$$Accuracy = \left(\frac{1}{D} \sum_{i=1}^{D} \left| \frac{\hat{L}^{(i)} \wedge L^{(i)}}{\hat{L}^{(i)} \vee L^{(i)}} \right| \right) \times 100\% \quad (1)$$

where D is total document in corpus (dataset), $\hat{L}^{(i)}$ is the prediction result of i^{th} document, and $L^{(i)}$ is the actual label of i^{th} document.

4.1 First Scenario Experiment Result

In this research, the first experiment scenario was done to know the combination of features, classifier, and data transformation method that we used that can give the best accuracy in identifying abusive language and hate speech including the target, categories, and level that was contained in a tweet. To obtain that, we do experiments using every type of feature extractions first. The experiment results for the best type of feature based on average accuracy using all classifiers and data transformation methods for the first scenario is in Table 1.

Based on average accuracy when doing experiments using every type of feature extractions (Table 1), we observe that for the first experiment scenario, the combination of word unigram, bigrams, and trigrams is the best *word n-grams* feature, character quadgrams is the best *character n-grams* feature, question mark is the best *orthography* feature, and negative sentiment is the best *lexicon* feature. However, if viewed individually, RFDT classifier with LP data transformation method when using word unigram feature gives the best performance with 66.12% of accuracy.

Table 1: Best of each type of feature extraction based on average accuracy for the first scenario

Type of Feature	Best Feature Based on Average Accuracy	Average Accuracy (%)
word n-gram	word unigram + bigram + trigram	59.44
character n-gram	character quadgrams	52.55
ortography	question mark	44.44
lexicon	negative sentiment	44.45

After obtaining the best features of each type of features, we do experiments using the combination of best features from each type of features. Based on experiments using the combination of best features from each type of features, we obtain that the combination of best features in this first experiment scenario does not give a significant result on the classification accuracy results. The best performance in experiments using the combination of best features is obtained when using RFDT classifier with LP data transformation method using the combination of character quadgrams, question mark, and negative sentiment just gives 65.73% of accuracy, still cannot exceed the accuracy given by the RFDT classifier with LP data transformation method using word unigram feature that can give 66.12% of accuracy.

Based on classification results and data analysis, we observe that the word unigram features gives the best accuracy may be because they represent the characteristics of each label. In each classification label, there are words that characterize the label. For example, in hate speech label, each tweet that labeled as hate speech contain hate words such as abusive words that demean an individual or group (e.g. *jelek* (ugly), *murahan* (gimrack), etc.), hate words related to politics in Indonesia (e.g. *antek* (henchman), *komunis* (communist), etc.), and

threatening/provoking words (e.g. *bakar* (burn), *bunuh* (kill), etc.). Next, for classifiers analysis, the ensemble method on RFDT relatively can give better accuracy compared to NB and SVM. For data transformation methods, LP can give the best accuracy because each unique label formed from the power-set process will have a correlation between labels so that it can reduce classification error (Kafrawy et al., 2015).

4.2 Second Scenario Experiment Result

The second experiment scenario in this research was done to know the combination of features, classifiers, and data transformation methods that can give the best accuracy in identifying abusive language and hate speech in a tweet without identifying the target, categories, and level of hate speech. Same as the first experiment scenario, we do experiments using every type of feature extractions first. The experiment results for best each type of feature based on average accuracy using all classifier and data transformation method for the second scenario can be seen in Table 2.

Table 2: Best of each type of feature extraction based on average accuracy for the second scenario

Type of Feature	Best Feature Based on Average Accuracy	Average Accuracy (%)
word n-gram	word unigram	73.53
character n-gram	character quadgrams	72.44
ortography	exclamation mark	45.27
lexicon	positive sentiment + abusive lexicon	52.10

Based on average accuracy when doing experiment using every type of feature extractions (Table 2), we obtain that for the second experiment scenario, word unigram feature is the best *word n-grams* feature, character quadgrams is the best *character quadgrams* feature, exclamation mark is the best *orthography* feature, and the combination of positive sentiment and abusive lexicon is the best *lexicon* feature. If viewed individually, RFDT classifier with LP data transformation method when using word unigram feature gives the best performance with 76.16% of accuracy.

After obtaining the best features of each type of features, we do experiment using the combination of best features from each type of feature. Based on the experiment using the combination of best features from each type of features, we obtain that the combination of best features in this second experiment scenario can give slightly better performance compared to when we do not combine the best feature. RFDT classifier with LP data transformation method when using the combination of word unigram, character quadgrams, positive sentiment, and abusive lexicon features can gives the best performance with 77.36% of accuracy.

4.3 Discussions

Based on the first and second experiment scenario results, we obtained that word unigram, RFDT, and LP is the best combination of feature, classifier, and data transformation method for both scenarios. From the second experiment scenario, our approach can reach a good enough performance in doing multi-label text classification to identify abusive language and hate speech without identifying the target, categories, and level of hate speech with 77.36% of accuracy when using RFDT classifier with LP data transformation method and word unigram feature extraction. However, when doing multi-label text classification to identify abusive language and hate speech including its target, categories, and level in the first scenario, the best performance from all our approaches that we use still does not give a good enough performance (only 66.12% of accuracy).

From our error analysis using confusion matrix (Fawcett, 2006) on each classification labels, the most common type of error is false negative. This misclassification is likely due to a large amount of unbalanced data in our dataset. According to (Ganganwar, 2012), unbalanced dataset can give negative results on classification performance because the unbalanced number of dataset between the majority and minority classes tends to make the classification performance on majority class better than classification performance on the minority class, such that it is necessary to balance the dataset. The balancing dataset process can be done by collecting new data and doing the annotation process with a focus on minority labeled data. However, this method needs to consider the data labeling process may be more expensive (Sabou et al., 2014). Some other methods that can be done to balance the dataset are data resampling (Chawla et al., 2002) and data augmentation (Wang and Yang, 2015; Kobayashi, 2018).

Notice that balancing dataset on multi-label problems is a quite difficult process because of the relationship between labels (Giraldo-Forero et al., 2013). To overcome this problem, several techniques can be used, one of which is the hierarchical multi-label classification (Madjarov et al., 2014). In this paper, the multi-label classification problem can be seen as hierarchical multi-label classification problem that can be done by identifying hate speech and abusive language first, and then reclassifying the tweets identified as hate speech to identify the target, categories, and level of hate speech separately. This approach can make the process of dataset balancing easier as classification is done separately for each label type (Feng and Zheng, 2017).

5 Conclusions and Future Works

In this paper, we discussed hate speech and abusive language detection in Indonesian Twitter. We conducted Focus Group Discussion (FGD) with staffs of Direktorat Tindak Pidana Siber Bareskrim Polri as the agency responsible for investigating cyber crimes in Indonesia in order to get a valid definition of hate speech, including the hate speech characterization. The results of the FGD are then poured into annotation guidelines for the purposes of annotating hate speeches. Besides conducted FGD with with staffs of Direktorat Tindak Pidana Siber Bareskrim Polri, we also conducted discussions with an expert linguist in order to make sure that the annotator guidelines we built valid and easy to understand by an annotator who is not a linguistic expert. Moreover, we also built gold standard annotations for testing whether a prospective annotator has read and understood the annotations guide or not. We then built a dataset for abusive language and hate speech identification (including identification of targets, categories, and level hate speech) using annotation guidelines and gold standard annotations that have been made. Our dataset including the annotation guidelines and gold standard annotations are open for public such that other researchers who are interested in doing research in hate speech and abusive language identification in Indonesian social media can use it.

After building the dataset, we did two experiment scenarios. Our experiment results show that word unigram, RFDT, and LP is the best combination of feature, classifier, and data transformation method for all scenarios we did. However, although our approach can reach a good enough performance in doing multi-label classification to identify abusive language and hate speech without identifying the target, categories, and level of hate speech (77.36% of accuracy), all the approaches we used still does not give a good enough performance when doing multi-label classification to identify abusive language and hate speech including identify the target, categories, and level of hate speech (only 66.12% of accuracy).

For future work, we suggest using hierarchical multi-label classification approach (Madjarov et al., 2014) for abusive language and hate speech identification including identify the target, categories, and level of hate speech. Our error analysis shows that a lot of false negative errors is probably caused by the unbalanced dataset (Ganganwar, 2012) such that it is necessary to balance the dataset. This hierarchical multi-label classification approach can make the process of dataset balancing easier because the classification is done separately on each label type (Feng and Zheng, 2017).

Besides doing hierarchical multi-label classification and dataset balancing, another thing that needs to be tried to improve the accuracy of this research is to add a semantic feature, namely *word embedding* (Mikolov et al., 2013) in the feature extraction process. In some text classification experiments in the Indonesian language (Saputri et al., 2018; Jannati et al., 2018), adding *word embedding* features to basic features such as *word n-grams* is shown to improve classification performance because the word embedding feature can recognize word meaning that cannot be captured by features such as frequency term, orthography and lexicon features.

From the FGD results, we obtained that handling hate speech problem in social media is not just about identifying whether a text/document is hate speech or not. There are several other tasks which needs to done to help the authorities in handling hate speech problems such as the identification of buzzers, thread starters, and fake account spreaders of hate speech.

Acknowledgments

The authors acknowledge the PITTA A research grant NKB-0350/UN2.R3.1/HKP.05.00/2019 from Directorate Research and Community Services, Universitas Indonesia.

References

Mirna Adriani, Jelita Asian, Bobby Nazief, S. M.M. Tahaghoghi, and Hugh E. Williams. 2007. Stemming indonesian: A confix-stripping approach. *ACM Transactions on Asian Language Information Processing (TALIP)*, 6(4):1–33.

Ika Alfina, Rio Mulia, Mohamad Ivan Fanany, and Yudo Ekananta. 2017. Hate speech detection in the indonesian language: A dataset and preliminary study. In *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, pages 233–238.

Ika Alfina, Siti Hadiyan Pratiwi, Indra Budi, Rio Mulia, and Yudo Ekanata. 2018. Detecting hate speech against religion in the indonesian language. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*.

APJII. 2017. *Infografis Penetrasi dan Pengguna Internet Indonesia Survey 2017*. Pustaka PelajarAsosiasi Penyelenggara Jasa Internet Indonesia, Jakarta.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Ying Chen, Yilu Zhou, Sencun Zhu, and Heng Xu. 2012. Detecting offensive language in social media to protect adolescent online safety. In *Proceedings of the 2012 ASE/IEEE International Conference on Social Computing and 2012 ASE/IEEE International Conference on Privacy, Security, Risk and Trust*, SOCIALCOM-PASSAT '12, pages 71–80, Washington, DC, USA. IEEE Computer Society.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media (ICWSM)*, pages 512–515.

T Fawcett. 2006. An introduction to roc analysis. *Pattern Recognition Letters*, 27:861–874.

Fu P. Feng, S. and W. Zheng. 2017. A hierarchical multi-label classification algorithm for gene function prediction. *Algorithms*, 10(4):1–14.

V Ganganwar. 2012. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, 2(4):42–47.

A. F. Giraldo-Forero, J. A. Jaramillo-Garzon, J. F. Ruiz-Munoz, and C. G. Castellanos-Dominguez. 2013. Managing imbalanced data setsin multi-label problems: A case study with the smote algorithm. In *Proceedings of the 18th Iberoamerican Congress on Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, pages 334–342.

Bayu Hernanto and Jeihan. 2018. Personal communication.

Muhammad Okky Ibrohim and Indra Budi. 2018. A dataset and preliminaries study for abusive language detection in indonesian social media. *Procedia Computer Science*, 135:222 – 229.

Muhammad Okky Ibrohim, Erryan Sazany, and Indra Budi. 2018. Identify abusive and offensive language inindonesian twitter using deep learning approach. *Journal of Physics: Conference Series*.

R. Jannati, R. Mahendra, C. W. Wardhana, and M. Adriani. 2018. Stance classification towards political figures on blog writing. In *2018 International Conference on Asian Language Processing (IALP)*, pages 96–101.

Passent El Kafrawy, Amr Mausad, and Heba Esmail. 2015. Article: Experimental comparison of methods for multi-label classification in different application domains. *International Journal of Computer Applications*, 114(19):1–9.

S. Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. In *Proceedings of NAACL-HLT 2018*, pages 452–457.

Ron Kohavi. 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence - Volume 2*, IJCAI'95, pages 1137–1143, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Komnas HAM. 2015. *Buku Saku Penanganan Ujaran Kebencian (Hate Speech)*. Komisi Nasional Hak Asasi Manusia, Jakarta.

F. Koto and G. Y. Rahmaningtyas. 2017. Inset lexicon: Evaluation of a word list for indonesian sentiment analysis in microblogs. In *2017 International Conference on Asian Language Processing (IALP)*, pages 391–394.

G. Madjarov, I. Dimitrovsk, D. Gjorgjevikj, and S. Dzerosk. 2014. Evaluation of different data-derived label hierarchies in multi-label classification. In *Proceedings of the 3rd International Conference on New Frontiers in Mining Complex Patterns*, pages 19–37.

Mary L. McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y Chang. 2016. Abusive language detection in online user content. In *International World Wide Web Conference Committee (IW3C2)*, pages 145–153.

Muzainah Nurasijah. 2018. Personal communication.

Tansa Trisna Astono Putri. 2018. Analisis dan deteksi hate speech pada sosial twitter berbahasa indonesia. Master's thesis, Faculty of Computer Science, Universitas Indonesia.

Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. 2014. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*. European Language Resources Association (ELRA).

Nikmatun Aliyah Salsabila, Yosef Ardhito Winatmoko, and Ali Akbar Septiandri. 2018. Colloquial indonesian lexicon. In *2018 International Conference on Asian Language Processing (IALP)*, pages 236–239.

M. S. Saputri, R. Mahendra, and M. Adriani. 2018. Emotion classification on indonesian twitter dataset. In *2018 International Conference on Asian Language Processing (IALP)*, pages 90–95.

Gregory H. Stanton. 2009. The rwandan genocide: Why early warning failed. *Journal of African Conflicts and Peace Studies*, 1(2):6–25.

F. Z. Tala. 2003. A study of stemming effects on information retrieval in bahasa indonesia. Master's thesis, Universiteti van Amsterdam The Netherlands.

S. Turaob and J.L Mitrpanont. 2017. Automatic discovery of abusive thai language. In *International Conference on Asia-Pacific Digital Libraries*, pages 267–278.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Esconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

W. Y. Wang and D. Yang. 2015. Thats so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using #petpeeve tweets. In *EMNLP*, pages 2557–2563.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

I Dewa Putu Wijana and Muhammad Rohmadi. 2010. *Sosiolinguistik: Kajian, Teori, dan Analisis*. Pustaka Pelajar, Yogyakarta.

Harish Yenala, Ashish Jhanwar, Manoj K. Chinnakotla, and Jay Goyal. 2017. Deep learning for detecting inappropriate content in text. In *International Journal of Data Science and Analytics*.

Appendix 1: Example of query used for crawling Twitter data

Query	Description	Citation
keparat	Abusive word (other)	(Wijana and Rohmadi., 2010)
anjing	Abusive word (other)	(Wijana and Rohmadi., 2010)
asu	Abusive word (other), other form of *anjing*	(Ibrohim and Budi, 2018)
banci	Abusive word related to gender/sexual orientation	(Wijana and Rohmadi., 2010)
bangsat	Abusive word (other)	(Wijana and Rohmadi., 2010)
bencong	Abusive word related to gender/sexual orientation, another form of *banci*	(Wijana and Rohmadi., 2010)
jancuk	Abusive word related to gender/sexual orientation	(Wijana and Rohmadi., 2010)
budek	Abusive word related to physical/disability	(Wijana and Rohmadi., 2010)
burik	Abusive word related to physical/disability	(Wijana and Rohmadi., 2010)
cocot	Abusive word (other)	(Ibrohim and Budi, 2018)
ngewe	Abusive word related to gender/sexual orientation	(Ibrohim and Budi, 2018)
kafir	Abusive word related to religion/creed	(Wijana and Rohmadi., 2010)
kapir	Abusive word related to religion/creed, another form of *kafir*	(Ibrohim and Budi, 2018)
sinting	Abusive word related to physical/disability	(Wijana and Rohmadi., 2010)
antek	Word related to hate speech issue in politics	(Hernanto and Jeihan, 2018)
asing	Word related to hate speech issue in politics	(Hernanto and Jeihan, 2018)
aseng	Word related to hate speech issue in politics, another form of *asing*	(Hernanto and Jeihan, 2018)
ateis	Abusive word related to religion/creed	(Wijana and Rohmadi., 2010)
sitip	Abusive word related to race/ethnicity	(Wijana and Rohmadi., 2010)
autis	Abusive word related to physical/disability	(Wijana and Rohmadi., 2010)
picek	Abusive word related to physical/disability	(Ibrohim and Budi, 2018)
ayam kampus	Abusive phrase related to gender/sexual orientation	(Ibrohim and Budi, 2018)
bani kotak	Phrase related to hate speech issue in politics	(Hernanto and Jeihan, 2018)
cebong	Word related to hate speech issue in politics	(Hernanto and Jeihan, 2018)
cina	Word related to hate speech issue in politics and race/ethnicity	(Hernanto and Jeihan, 2018)
china	Word related to hate speech issue in politics and race/ethnicity, other form of cina	(Hernanto and Jeihan, 2018)
hindu	Word related to hate speech issue in religion/creed	(Hernanto and Jeihan, 2018)
katolik	Word related to hate speech issue in religion/creed	(Hernanto and Jeihan, 2018)
katholik	Word related to hate speech issue in religion/creed, another form of *katolik*	(Hernanto and Jeihan, 2018)
komunis	Word related to hate speech issue in politics and race/ethnicity	(Hernanto and Jeihan, 2018)
kristen	Word related to hate speech issue in religion/creed	(Hernanto and Jeihan, 2018)
onta	Word related to hate speech issue in politics	(Hernanto and Jeihan, 2018)
pasukan nasi	Phrase related to hate speech issue in politics	(Hernanto and Jeihan, 2018)
tionghoa	Word related to hate speech issue in politics and race/ethnicity	(Hernanto and Jeihan, 2018)