Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 135 (2018) 222–229

Procedia Computer Science

www.elsevier.com/locate/procedia

3rd International Conference on Computer Science and Computational Intelligence 2018

A Dataset and Preliminaries Study for Abusive Language Detection in Indonesian Social Media

Muhammad Okky Ibrohim[*,a], Indra Budi[**,a]

[a]*Faculty of Computer Science, Universitas Indonesia, Kampus UI Depok 16424, Indonesia*

Abstract

Abusive language is an expression (both oral or text) that contains abusive/dirty words or phrases both in the context of jokes, a vulgar sex conservation or to cursing someone. Nowadays many people on the internet (netizens) write and post an abusive language in the social media such as Facebook, Line, Twitter, etc. Detecting an abusive language in social media is a difficult problem to resolve because this problem can not be resolved just use word matching. This paper discusses a preliminaries study for abusive language detection in Indonesian social media and the challenge in developing a system for Indonesian abusive language detection, especially in social media. We also built reported an experiment for abusive language detection on Indonesian tweet using machine learning approach with a simple word n-gram and char n-gram features. We use Naive Bayes, Support Vector Machine, and Random Forest Decision Tree classifier to identify the tweet whether the tweet is a *not abusive language, abusive but not offensive,* or *offensive language*. The experiment results show that the Naive Bayes classifier with the combination of word unigram + bigrams features gives the best result i.e. 70.06% of $F_1 - Score$. However, if we classifying the tweet into two labels only (*not abusive language* and *abusive language*), all classifier that we used gives a higher result (more than 83% of $F_1 - Score$ for every classifier). The dataset in this experiment is available for other researchers that interest to improved this study.

Keywords: abusive language; twitter; machine learning.

* Corresponding author 1.
E-mail address: muhammad.okky71@ui.ac.id

** Corresponding author 2.
E-mail address: indra@cs.ui.ac.id

1877-0509
10.1016/j.procs.2018.08.169

1. Introduction

Indonesia is one of the countries with the most social media users[1]. Indonesian people often use social media for various purposes, such as finding and sharing information, establish communication, advertising media, or to just release feelings of the heart (vent). A large number of social media users often leads to uncontrolled communication and many netizens who communicate with an abusive language[1]. Abusive language is an expression that contains abusive/dirty words or phrases, both oral or text[1]. According to[1], the causes of uncontrolled the use of abusive words in social media are the absence of effective tools to filter abusive language in social media, lack of empathy among citizens, and lack of parental guidance. Abusive language in social media needs to be filtered so that there are no children and adolescents who learn abusive language from the social media that they used[2]. However, it is almost impossible to filter abusive language in social media manually because of a large number of people who write the abusive language. Thus, the abusive language in social media needs to be automatically detected.

Detecting an abusive language in social media is a problem difficult to resolve. Nobata et. al.[3] said that detecting an abusive language in social media can not just use word matching. Moreover, the spelling and grammar from netizens when speech abusive language in social media is very informal. Especially in short text data, classifying short text data to detect an abusive language is more difficult to resolve. For example in Twitter data, there are a lot of netizens posting a tweet using abbreviations because of the word limit of a tweet. Hanafiah et. al.[4] said that some of non-formal words often used by Indonesians are: words that show feelings, character repetition to emphasize meaning, using slang words, and changing vowels to numbers.

Research about abusive language detection in social media has been done in recent years with the various approach. Turaob et. al.[1] discuss abusive language detection in the Thai language. The dataset which they used is the Facebook post and comment data. They used several classifier which are Naive Bayes (NB), k-Nearest Neighbor (kNN), Support Vector Machine (SVM), Random Forest Decision Tree (RFDT), etc. with several term weighting features such as word n-grams (word unigram and word bigrams) and Term Frequency – Inverse Document Frequency (TFIDF).

Chen et. al.[2] explain the importance of research about abusive language detection in social media. We need to classifying abusive language on social media so that we can create better social media for kids and teens. Their research was done using English Youtube comment board as the dataset with NB and SVM as the classifier. The features used in their research are word n-grams, appraisal approach, and lexical syntactical feature.

Based on the author's knowledge, study about abusive language detection in Indonesian social media has never been done before. However, there are some researchers about text classification on Indonesian social media. Alfina et. al.[5] have done a research about hate speech detection in the Indonesian language. They used Indonesian tweet as the dataset which crawled using Twitter API[2]. Each tweet data was labeled manually by three annotators which are a college student from different age an religion background. The agreement level between the annotators on every tweet is 100%. It means that if theres a tweet with different annotation from the annotators, then that tweet will be deleted (not used in the dataset). This is done to prevent some ambiguity in the classification process. They compare several algorithms which are NB, SVM, Binary Logistic Regression (BLR), and RFDT to classify whether a tweet is a hate speech or not. Features used in this research are word n-gram and char n-gram features.

Besides choosing which classifier and features to use in the classification process, there is another important to do before the classification process, which is balancing the dataset number to each of data label class. Ganganwar[6] said that an unbalanced dataset can give a negative result to a classification. This is because an unbalanced number of dataset between major and minor class tends to make the major class have a better performance than the minor one. This problem can be solved with data re-sampling technique. Data re-sampling is a technique to balance a dataset by deleting some of the major class data or duplicate some of the minor class data so that the number of the dataset on each class become more balance.

In this research, we built a new Twitter dataset for Indonesian abusive language detection. In general, the contributions of this research are:

[1] https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/
[2] https://apps.twitter.com/

- Analyzing the types of Indonesian abusive language and give the examples of every type, and also giving analyze results about the challenge in detecting abusive language in Indonesian social media;
- Building a new Twitter dataset for research in Indonesian abusive language detection. Our dataset was opened for public such that other researcher who interested doing research in this scope can use this data[3];
- Presenting a preliminaries performance in Indonesian abusive language detection using several classifiers with word n-gram and char n-gram features.

This paper is organized as follows. Section 2 discusses the types of Indonesian abusive language including the examples of every type. Section 3 explains about the challenge on abusive language detection in Indonesian. The dataset and method that we used in this research are explained in Section 4. Our experiment results are presented in Section 5. The conclusions and futures work from this paper are presented in Section 6.

2. Indonesian Abusive Language

In the Indonesian language, the abusive/dirty words usually come from a condition, animals, astral beings, an object, a part of a body, family member, activity and a profession[7,8]. Below is further explanations about the types of abusive/dirty word references in Indonesian language[4].

- **Condition**. Words that expressed unpleasant condition in a conversation usually used as abusive words. In general, there are three things that could or might be connected with these unpleasant condition, which are mental disorder (e.g.: *gila, bego, goblok, idiot, sinting, bodoh, tolol, sontoloyo, geblek,sarap*), sexual deviation (e.g.: *lesbian, homo, banci*), lack of modernization (e.g.: *kampungan, udik, alay*), physical disability (e.g.: *buta, budek, bolot, bisu*), condition where someone doesn't have etiquette (e.g.: *berengsek, bejat*) conditions that are not sanctioned by god or religion (e.g.: *keparat, jahanam, terkutuk, kafir, najis*), and condition that related to unfortunate circumstances (e.g.: *celaka, mati, modar, sialan, mampus*).
- **Animals**. Not all animals can be used as abusive words. Animals that are used as offensive words usually refer to its certain bad characteristic, which are disgusting for some people (e.g.: *anjing*), disgusting and forbidden in certain religion (e.g.: *babi*), annoying (e.g.: *bangsat, monyet, kunyuk*), parasitic (e.g.: *lintah*), lusty (e.g.: *buaya, bandot*), and noisy (e.g.: *beo*).
- **Astral beings**. The examples of astral beings that usually used as abusive words are *setan, setan alas, iblis, tuyul,* and *kunti*. They are all astral beings that often interfere with human life.
- **An object**. Same as animals and astral beings, the objects that usually used as abusive words are based on their bad characteristic, such as bad smell (e.g.: *tai, tai kucing, bangkai*), dirty and worn out (e.g.: *gembel, gombal*), and disturbing sound (e.g.: *sompret*).
- **A part of a body**. Body parts that used as abusive words usually closely related to sex activity such as *kontol, memek, tempik,* and *jembut*. Another body part often used in cursing is eye (*mata* in Indonesian) in the form of *matamu* (your eye) which means one cursing the other for not using their eyes properly and making mistakes because of it. The other phrases are *hidung belang* and *mata duitan* which are used figuratively to curse a pervert man and a person who choose money over everything, respectively.
- **Family member**. Indonesian people usually adds suffix *-mu* on words refer to relatives as a curse, such as *ibumu* (your mother), *bapakmu* (your father), *kakekmu* (your grandfather), and *nenekmu* (your grandmother).
- **Activity**. Abusive words on the activity are usually more indented towards sexual, such as *ngentot,* and *ngewe*.
- **Profession**. One's occupation, especially low-class occupation forbidden by religion, often used by Indonesian people as abusive words. Those occupations include *maling, sundal, bajingan, copet, lonte, cecenguk, kacung, pelacur, pecun, jablay* and *perek*.

[3] https://github.com/okkyibrohim/id-abusive-language-detection
[4] To find out the meaning of each word, the reader can look it up on Kamus Besar Bahasa Indonesia (KBBI): https://kbbi.kemdikbud.go.id

From the list of abusive words already described, how the use of abusive words to curse someone in the Indonesian language can be divided into three types. Abusive language in the context to curse someone also called as offensive language[9]. The three types of offensive language are offensive language in the form of words, phrases, and clauses[7,8], which can be seen in Table 1.

Table 1. Examples of offensive languages forms

Types	Examples
words	*Anjing!* (You are dog!); *Ngentot!* (Fuck You!); etc.
phrases	*Dasar bodoh!* (You stupid!); *Dasar cowo hidung belang!* (You perverted man!)
clauses	*Goblok lu, gitu aja ga bisa!?* (Idiot, you can't even do something like this!?)

Note that not every abusive/dirty word or phrases are used to curse someone (used as offensive language). In a modern conversation in Indonesian, abusive words can be used to express astonishment, amazement, wonder, etc[7]. For example, the clause *'Gila, ganteng banget!'* is not to curse that someone is crazy. Here, the abusive word *gila* is used to express 'wow' as a form of amazement. Therefore, the clause *'Gila, ganteng banget!'* is meant to 'Wow, how handsome He is!'. For more example, the clause *'Anjing! Kocak banget lu* :D' does not mean cursing someone with the word *anjing*, but a form of admiration for the joke given by someone. Therefore, the clause *'Anjing! Kocak banget lu* :D' is meant to 'Damn! you are so funny :D'. These clause appertains as abusive language, but not offensive.

3. Challenge in Detecting Abusive Language in Indonesian Social Media

As explained by[3], detecting an abusive language in social media is a difficult problem. Especially in Twitter data, the habit of netizens who post tweets with informal language makes the researchers have to perform special techniques in normalizing the data[4]. Detecting an abusive language in Indonesian social media become more difficult because many netizens in Indonesia use abusive words in a foreign language in their conversations[10], both in the context of jokes (abusive language but not offensive) or to curse someone (offensive language). Examples of abusive words in a foreign language that often used by Indonesian netizens is *fuck, shit, bitch, motherfucker* (English) and *cyka blyat* (Russian language). The use of abusive words in a foreign language is not only in the formal form, but informal ones[11]. For example, many Indonesian netizens type *'Fak yu!'* to say *'Fuck you!'*.

Not only using abusive words in foreign language, many Indonesian netizens also usually use abusive words in their local language[12]. The examples of abusive words in Indonesian local language that often used by Indonesian netizens in their conversation are *asu* (Javanese language, means *dog*), *kimak* (North Sumatera's language, means *pussy*), *kampang* (West Sumatera's language, means *pussy*), *jancuk* (East Javanese language, means *fuck*), etc. Based on the author's analysis of some social media that often used by Indonesian netizens such as Facebook, Line, and Twitter, there are some abusive word writing patterns (both formal and informal) in Indonesian social media which need to be considered in doing abusive language detection. Below are explanations about the abusive word writing patterns in Indonesian social media.

- **Using informal form of abusive words**. Indonesian netizens usually make an informal form of an abusive word for extend the abusive word. The informal form is made by making a new vocabulary that the pronunciation is similar to the original abusive word. For examples, many netizens type *meki* for saying *memek* and type *kintil* for saying *kontol*.
- **Using foreign and local language**. As explained before, many Indonesian netizens saying abusive words using a foreign language (e.g.: *fuck, shit, bitch, motherfucker, cyka blyat,* etc.) or local language (e.g.:*asu, kimak, kampang, jancuk,* etc.). Not only in formal form, they also usually type in informal form. Some netizens usually type an abusive language in one language, whether pure Indonesian language, pure foreign language, or pure local language. However, there are also netizens who type an abusive language with mixed language.
- **Erase the vowel**. In many social media especially in Twitter which limits the number of characters in a post, often found netizens who type an abusive word by removing the vowels. For examples, they are type *bgst* for saying *bangsat* and type *anjg* for saying *anjing*.

- **Character repetition**. In very angry circumstances, netizens sometimes type abusive words by repeating some characters to show their anger. The examples for this patterns are *baaaanggsaaaattt* (*bangsat*), *taaiiiii* (*tai*), *annjiiiiiingg* (*anjing*), etc.
- **Character substitution**. Many netizens in Indonesia substitute some character in abusive words when speaking an abusive language for some reason. To show their anger, some netizens changing *t* with *d*, e.g.: *bangsad* (*bangsat*), *jembud* (*jembut*), *bejad* (*bejat*), *ngentod* (*ngentot*), etc. In the context of jokes, netizens usually changing *s* with *c*, e.g.: *bangcat* (*bangsat*), *acu* (*asu*), etc. Some other patterns are changing *k* with *q* (e.g.: *qontol* (*kontol*), *qimak* (*kimak*)), changing *j* with *dj* (e.g.: *djembut* (*jembut*), *djancuk* (*jancuk*)), changing *u* with *oe* (e.g.: *jemboet* (*jembut*), *djancoek* (*jancuk*)), and changing vowel with a number (e.g.: *b4ngs4t* (*bangsat*), *b3g0* (*bego*), *j3mbut* (*jembut*)). Besides, we often find netizens changing a character in an abusive words with certain characters to pretend that they have sensored the abusive words, e.g.: *mem*k* (*memek*), *g*blok* (*goblok*), *b#ngsat* (*bangsat*), etc.

4. Dataset and Method

In this research, we built a new Twitter dataset for abusive language detection in Indonesian tweet. Generally, the research flowchart for this experiments in this paper can be seen in Figure 1.

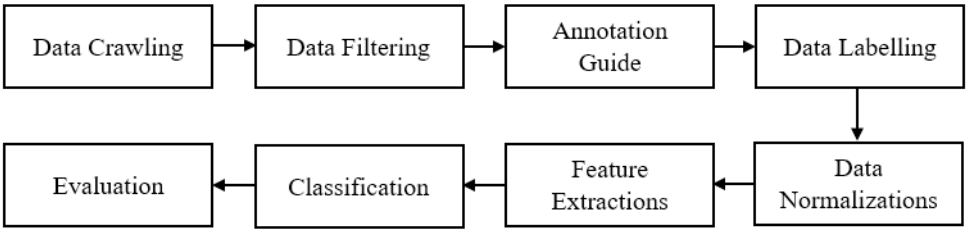

Fig. 1. The research experimments flowchart

The first step of this research experiments is to crawl the Twitter data. Twitter data crawling process was done using Twitter API and Tweepy Library[5]. In this preliminaries experiment for Indonesian abusive language detection, we crawl the tweet data using abusive words references that described in Section 2 and only several informal forms of abusive words that described in Section 3 as the query. The several informal forms of abusive words that we used as query when crawling the Twitter data are *bejad* (*bejat*), *brengsek* (*berengsek*), *asu* (*anjing*), *asw* (*anjing*), *bangke* (*bangkai*), and *goblog* (*goblok*). This is because those informal abusive words form often found in Indonesian social media. We collect about 60,000 tweet data. From the data was collected, we filter the data by erasing the duplicated tweets and the tweets that using foreign or local language. Filtering data process give 2,500 data that is ready to be labeled.

After filtering the tweet data, the dataset was labeled. In this research, the dataset labeled into three labels namely *not abusive language, abusive but not offensive,* and *offensive language*. The dataset was labeled by 20 volunteer annotators. To ensure that annotators understand for they task, we make an annotation guide and explaining that annotation guide to the annotators. Each tweet in the dataset was annotated by three annotators. We used 100% annotators agreement, so the tweet that has a different label removed from the dataset. The labeling process gives 2,016 data that have 100% annotators agreement.

For text normalization process, first, we erased several unnecessary attributes such as username, re-tweet (RT), and uniform resource locator (URL) address. We also delete punctuation, hashtag, and emoticons from the dataset for experiments in this paper. After the deletion, the next step is changing nonformal words into formal ones using a dictionary. In this paper, we build a small slang word dictionary to normalize the nonformal word in our dataset.

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

The next process is feature extractions. Features that will be used in this research are word n-gram features[1–5], and char n-gram[5]. After the feature extractions process, the data will be ready for classification. In this research, we use several classifiers such as Naive Bayes (NB)[13], Support Vector Machine (SVM)[14], and Random Forest Decision Tree (RFDT)[15]. The implementation of those classifiers was done using a library from Scikit-Learn[6]. Scikit-Learn is a library providing some popular classification algorithm in Python language and often used on researches about text classification, one of them is[16] which use SVM library from Scikit-Learn.

To evaluate whether the classification result is good or not and to find the best features combination, we will use 10-fold cross-validation technique. In this technique, a dataset is divided into 10 parts where 9/10 of it will become the training data, and the 1/10 will become the testing data. The training-testing process using this technique will be done 10 times so that every data will become both training data and testing data simultaneously. To avoid the overfitting result because of the unbalanced data, we use SMOTE[17] library for handling the unbalanced data problem. In this paper, we used $F_1 - Score$ (also called $F - Measure$)[18] for the metric evaluation.

5. Experiments and Discussions

In this paper, we conducted two experiment scenarios. In the first scenario, we classify the tweet into three labels which are *non abusive language, abusive but not offensive* and *offensive language*. This is done in accordance with the explanation of[7] that an abusive language not necessarily offensive language. For the second scenario, we just classify the dataset into two labels which are *non abusive language* and *abusive language* in order to know whether classification results would be better if only classify two labels. In this second scenario, tweet that have label *abusive but not offensive* and *offensive language* will labeled as *abusive language*.

For both scenario, we used word n-gram and char n-gram features with NB, SVM and RFDT as classifier. The word n-gram that we used are word unigram, word bigrams, word trigrams and a combination of them all, while char n-gram that we used are char trigrams, char quadgrams, and also combination of char trigrams and char quadgrams. Table 2 and Table 3 show the $F_1 - Score$ for every scenario in %.

Based on Table 2, we can see that for the first scenario NB with word unigram + bigrams feature gives the best result with 70.06% of $F_1 - Score$, followed by NB with word unigram + bigrams + trigrams feature (69.64%) and NB with char quadgrams feature (69.55%). While from Table 3 for the second scenario, we can see that NB with word unigram gives the best result with 86.43% of $F_1 - Score$, followed by NB with char trigrams + quadgrams feature (86.17%) and NB with word unigram + bigrams (86.12%). From both scenarios, it can be seen that NB is better than SVM and RFDT for classifying the tweet in these experiments. Also, word unigram and the combination of word n-grams give better a result than the other features for every classifier that we used. Therefore, for the future works on abusive language detection in Indonesian social media, we suggest that using NB with word unigram and the combination of word n-grams for the baseline classifier and feature extractions.

Table 2. $F_1 - Score$ for three class labels

	NB	SVM	RFDT
word unigram	69.50	67.48	65.25
word bigrams	50.75	53.15	27.27
word trigrams	27.33	22.11	47.26
word unigram + bigrams	**70.06**	66.67	61.14
word unigram + bigrams + trigrams	69.64	67.26	61.03
char trigrams	66.67	64.97	59.63
char quadgrams	69.55	66.18	62.82
char trigrams + quadgrams	69.13	64.38	61.03

From both scenario, it also can be seen that classifying tweet into three labels is more difficult than just classifying whether the tweet is a *not abusive language* or an *abusive language*. Here, all classifier with all features that we used

[6] http://scikit-learn.org

Table 3. $F_1 - Score$ for two class labels

	NB	SVM	RFDT
word unigram	**86.43**	83.94	83.42
word bigrams	55.27	81.18	39.17
word trigrams	16.59	79.72	72.16
word unigram + bigrams	86.12	83.20	83.08
word unigram + bigrams + trigrams	84.90	83.53	81.98
char trigrams	85.20	80.57	82.27
char quadgrams	85.73	81.47	82.73
char trigrams + quadgrams	86.17	80.74	82.39

difficulties to differentiate whether the tweet is an *abusive but not offensive* or an *offensive language*. To solve this problem, future research can use other classifier or add certain features. From our analysis of the dataset, there are certain patterns that differentiate whether the tweet is an *abusive but not offensive language* or an *offensive language*. Many netizens usually type an offensive language using uppercase or using the exclamation point (*!*), while saying an abusive but not offensive language netizens usually add the word that meaning to laugh at something such as *wkwk, haha, hihi, hehe,* etc. Hence, for future work, we suggested to try to use uppercase, punctuation, and laugh word for features extractions.

6. Conclusions and Future Works

In this paper, we discussed the Indonesian abusive language in social media which commonly come from an unpleasant condition or something that is disgusting and forbidden by a religion. We also discussed challenges in detecting abusive language in Indonesian social media included the abusive words writing patterns in Indonesian social media. Here, we build a new dataset and do an experiment for abusive language detection in the Indonesian language.

The experiment results show that NB is better than SVM and RFDT for classifying abusive language using our dataset in all scenarios. For the features extractions, word unigram and the combination of word n-gram gives better results than the other features, either using NB, SVM or RFDT. The experiment results also show that classifying the tweet into three labels (*non abusive language, abusive but not offensive,* and *offensive language*) is more difficult than just classifying whether the tweet is a *not abusive language* or an *abusive language*. Here, the classifier we used difficulties to differentiate whether the tweet is an *abusive but not offensive* or an *offensive language*.

For future works, we suggested using NB with word unigram and the combination of word n-gram as a baseline performance. We also suggested using uppercase, punctuation, and laugh words dictionary for features extractions in order to improve the classification results in differentiating whether the tweet is an *abusive but not offensive language* or an *offensive language*. In Indonesia, offensive language is a part of hate speech[7]. Hence, the researcher that interest researching in hate speech detection can use our dataset.

Acknowledgements

The authors acknowledge the PITTA research grant 1882/UN2.R3.1/HKP.05.00/2018 from Directorate Research and Community Services, Universitas Indonesia.

References

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

[7] *Surat Edaran Kapolri Nomor: SE/6/X/2015 tentang Ujaran Kebencian*

2. Chen, Y.. Detecting offensive language in social media to protect adolescent online safety. *Expert Systems with Applications* 2012;**36**:71–80.
3. Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y., Chang, Y.. Abusive language detection in online user content. In: *International World Wide Web Conference Committee (IW3C2)*. 2016, p. 145–153.
4. Hanafiah, N., Kevin, A., Sutanto, C., Fiona, , Arifin, Y., Hartanto, J.. Text normalization algorithm on twitter in complaint category. In: *Procedia Computer Science*; vol. 116. 2017, p. 20–26.
5. Alfina, I., Mulia, R., Fanany, M., Ekananta, Y.. Hate speech detection in the indonesian language: A dataset and preliminary study. In: *International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. 2017, .
6. Ganganwar, V.. An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering* 2012;**2(4)**:42–47.
7. Wijana, I.D.P., Rohmad., M.. *Sosiolinguistik: Kajian, Teori, dan Analisis*. Pustaka Pelajar, Yogyakarta; 2010.
8. Triadi, R.. Penggunaan makian bahasa indonesia pada media sosial (kajian sosiolinguisik). *Jurnal Sasindo Unpam* 2017;**5(2)**:1–26.
9. Davidson, T., Warmsley, D., Macy, M.W., Weber, I.. Automated hate speech detection and the problem of offensive language. In: *International AAAI Conference on Web and Social Media (ICWSM)*. 2017, p. 512–515.
10. Amrullah, L.. English swear words by indonesian learners. *Journal of English Language Teaching and Linguistics* 2016;**1(1)**:1–12.
11. Wijana, I.D.P.. The use of english in adolescent's slang. *Humaniora* 2012;**23(3)**:315–323.
12. Wijana, I.D.P.. Kata-kata kasar dalam bahasa jawa. *Humaniora* 2008;**20(3)**:249–256.
13. Lewis, D.D.. Naive (bayes) at forty: The independence assumption in information retrieval. In: *European Conference on Machine Learning (EMCL)*. 1998, p. 4–15.
14. Srivastava, D.K., Bhambhu, L.. Data classification using support vector machine 2009;**12(1)**:1–7.
15. Ali, J., Khan, R., Ahmad, N., Maqsood, I.. Random forests and decision trees 2012;**9(3)**:272–278.
16. Tulkens, S., Hilte, L., Lodewyckx, E., Verhoeven, B., Daelemans, W.. Dictionary-based approach to racism detection in dutch social media. In: *First Workshop on Text Analytics for Cybersecurity and Online Safety (TA-COS)*. 2016, p. 11–17.
17. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 2002;**16**:321–357.
18. Hossin, M., Sulaiman, M.N.. A review on evaluation metrics for data classification evaluations 2015;**5**:1–11.