Ensemble Method for Indonesian Twitter Hate Speech Detection

M. Ali Fauzi[1], Anny Yuniarti[2]
[1]Faculty of Computer Science, Brawijaya University, Malang, Indonesia
[2]Informatics Department, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Article Info

Article history:

Received 27 Dec, 2017
Revised Jan 9, 2018
Accepted Jan 16, 2018

Keywords:

Classifier ensemble
Hate speech
Indonesian language
Text classification
Twitter

ABSTRACT

Due to the massive increase of user-generated web content, in particular on social media networks where anyone can give a statement freely without any limitations, the amount of hateful activities is also increasing. Social media and microblogging web services, such as Twitter, allowing to read and analyze user tweets in near real time. Twitter is a logical source of data for hate speech analysis since users of twitter are more likely to express their emotions of an event by posting some tweet. This analysis can help for early identification of hate speech so it can be prevented to be spread widely. The manual way of classifying out hateful contents in twitter is costly and not scalable. Therefore, the automatic way of hate speech detection is needed to be developed for tweets in Indonesian language. In this study, we used ensemble method for hate speech detection in Indonesian language. We employed five stand-alone classification algorithms, including Naïve Bayes, K-Nearest Neighbours, Maximum Entropy, Random Forest, and Support Vector Machines, and two ensemble methods, hard voting and soft voting, on Twitter hate speech dataset. The experiment results showed that using ensemble method can improve the classification performance. The best result is achieved when using soft voting with F1 measure 79.8% on unbalance dataset and 84.7% on balanced dataset. Although the improvement is not truly remarkable, using ensemble method can reduce the jeopardy of choosing a poor classifier to be used for detecting new tweets as hate speech or not.

Corresponding Author:

M. Ali Fauzi,
Faculty of Computer Science,
Brawijaya University, Malang, Indonesia
Email: moch.ali.fauzi@ub.ac.id

1. INTRODUCTION

Hate speech is any communicative acts that used to express hatred towards a person or a group on the basis of some characteristic such as race, ethnicity, gender, sexual orientation, nationality, religion, or other characteristic [1]. Due to the massive increase of user-generated web content, in particular on social media networks where anyone can give a statement freely without any limitations, the amount of hateful activities is also increasing. Social media technology make people able to express their opinion, including hate speech, quickly, then spread widely and become viral if the topics covered are 'interesting'. It can bring up disputes between groups in society. In Indonesia, based on the data of National Police Criminal Investigation Agency of Indonesia in 2015, there are 143 cybercrimes in the form of hate speech. This number increased to 199 in 2016. However, this data only cover hate speeches being criminalized and reported to the police. Obviously there are still many more hate speeches that exist in various social media.

One of the popular social media in Indonesia is Twitter [2]. Social media and microblogging web services, such as Twitter, allowing to read and analyze user tweets in near real time. Twitter is a logical source of data for hate speech analysis since users of twitter are more likely to express their emotions of an event by posting some tweet [3]. This analysis can help for early identification of hate speech so it can be prevented to be spread widely. It is also useful for content filtering and early detection of wrongful activities [4]. The manual way of detecting out hateful tweets is costly and not scalable. Therefore, the automatic way of hate speech detection is needed to be developed for tweets in Indonesian language.

Some previous works proposed in hate speech detection mostly for English [5-7]. Most of them used machine learning technique and the dataset is from Twitter. Meanwhile, the study of hate speech detection in Indonesian language is still very rare. As far as we know, [8] and [9] are the only works in hate speech detection in Indonesian language. These works provide datasets for hate speech detection in Indonesian language from Twitter. These works also used machine learning approach to tackle this problem. Basically we also consider the hate speech detection as a text classification problem. In this work, we focus on the problem of classifying a tweet as hate speech or not. Text classification technique mostly using bag of words features and machine learning methods such as Naïve Bayes (NB) [10], K-Nearest Neighbors (KNN) [11], Maximum Entropy (ME) [12], Random Forest (RF) [13], or Support Vector Machines (SVM) [1] for classification task.

In this works, we used ensemble method to tackle this problem. An ensemble of classifiers is a set of stand-alone classifiers which combined to classify new tweet in order to improve classification performance [14]. In general text classification, several works using ensemble method have been conducted and reported that ensembles method can enhance the classification performance (e.g. [15-17]). Several classifier that been used in this ensemble are NB, KNN, ME, RF, and SVM. We aim to improve the performance of some stand-alone classifiers by combining them.

2. RESEARCH METHOD

As seen in Figure 1, hate speech detection in this work consists of three main stages: 1) preprocessing; 2) training some stand-alone classifiers; and 3) combining the classifiers.

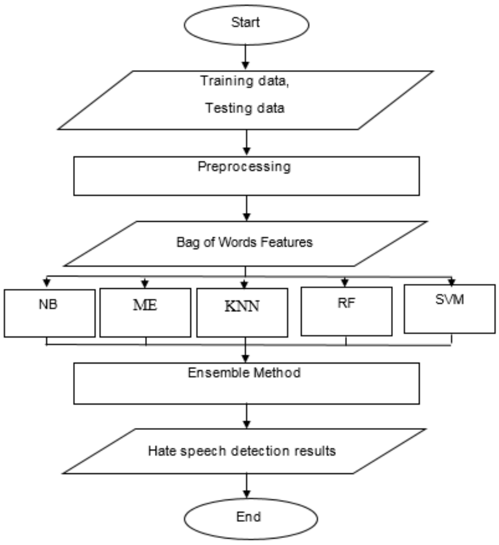

Figure 1. Hate Speech Detection Flowchart

2.1. Tweet Preprocessing

In tweet preprocessing, there are some steps to be conducted: 1) tokenization; 2) filtering; 3) stemming; and 4) term weighting. Tokenization is a task of splitting tweets into smaller units called tokens or terms. In this process, case folding and cleansing are also conducted. Case folding is a process of converting all of characters into lowercase. In the cleansing process, punctuation, numbers, html tag and characters outside of the alphabet were removed. The next step is filtering or Stopwords Removal. Stopwords or uninformative words were removed in this step based on an existing stoplist dictionary. In this work, we stoplist dictionary by Tala [18]. The fourth step is stemming or a process of reducing every words to its root or base form. The words ‘dilawan’, ‘melawan, and ‘perlawanan will be converted to the same word ‘lawan’ [19]-[21].

The last step in preprocessing is word or term weighting. In this work, we use bag of words (BOW) features with TF.IDF weighting. TF.IDF is the most popular term weighting method in text classification [22]. TF.IDF is a combination of term frequency (TF) and inverse document frequency (IDF). The TF.IDF weight of term or word t in tweet or document d is calculated as follows:

$$TF \cdot IDF(t,d) = (1+\log\left(f_{t,d}\right)) \cdot (1+\log\left(\frac{N_d}{df_t}\right))$$

where $f_{t,d}$ is the number of occurrences of term t in tweet d and N_d is the number of tweets in corpus and df_t is the number of tweets in corpus that contains term t. Finally, this stage produce a bag of words (BOW) features which will be used in the next stage.

2.2. Training Some Stand-alone Classifiers

In the second stage, several popular classifiers is trained. In this work, we used Naïve Bayes, K-Nearest Neighbours, Maximum Entropy, Random Forest, and Support Vector Machines. For Naïve Bayes, we used Multinomial distribution as it proves to show good performance in text classification. Meanwhile, for SVM, we used Linear kernel for the same reason. Finally, the classifiers is ensembled in the last stage.

2.3. Combining the Classifiers

In the last stage, several classifiers from the previous stage is combined. We conducted two types of ensemble methods: 1) hard voting; and 2) soft voting. In hard voting, each stand-alone classifier has one vote. As seen in Figure 2, the category of a tweet is selected by majority voting. The category selected is the one which have a majority, that is, more than half the votes. Meanwhile in soft voting, average category probabilities is used as voting score. As seen in Figure 3, the final category of a tweet is the one with the highest voting score or average probability from each classifiers.

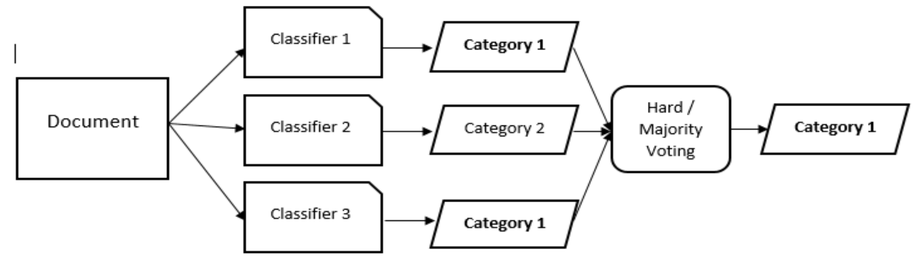

Figure 2. Hard Voting Ensemble Method

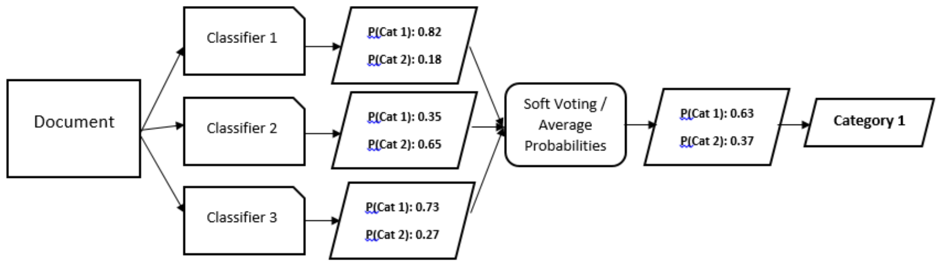

Figure 3. Soft Voting Ensemble Method

3. RESULTS AND ANALYSIS

We implemented the experiments using Scikit-Learn [23]. We used Twitter hate speech dataset in Indonesian language that have been collected and labelled by [9]. There are 260 tweets labelled as hate speech and 445 tweets labelled as non hate speech. We kept the dataset unbalanced in the first experiment. For the second experiment, we transform the unbalanced dataset into a balanced one using an undersampling method. We choose non hate speech tweets randomly so that the numbers of the non hate speech tweets become the same number of the hate speech tweets.

In the experiments, we compared the results of stand-alone classifiers with our ensemble method. We use 10 fold cross validation, which is mean the dataset is equally divided into 10 folds first. In each iteration of cross validation, tweets from 9 folds were used as training data and the remaining fold was used as testing data. We use average F1 Measure as the evaluation method in this experiments. Experiment results displayed in Figure 4 and Figure 5.

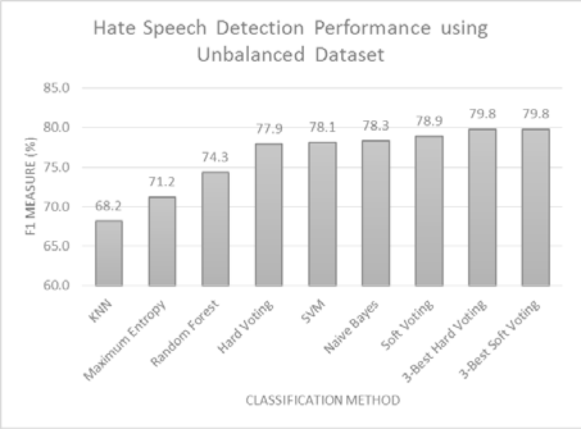

Figure 4. Hate Speech Detection Performance using Unbalanced Dataset

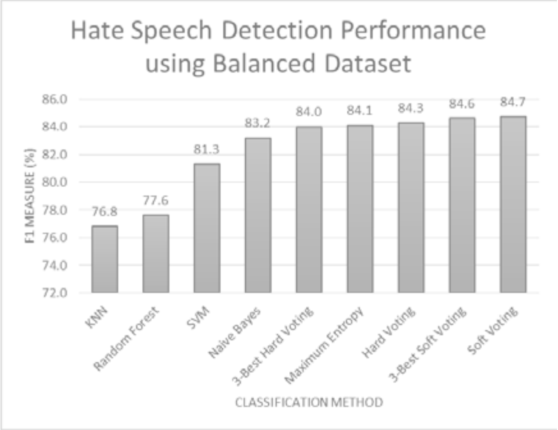

Figure 5. Hate Speech Detection Performance using Balanced Dataset

As seen in Figure 4, among all stand-alone classifiers, NB has the best performance on unbalanced dataset compared with other stand-alone classifier by 78.3% F1 measure. SVM performed almost the same to NB with F1 measure 78.1%. It is clear to see that KNN was the most inferior classifier with only 74.2% F1 measure. Meanwhile, RF and ME performed better than KNN by 71.2% and 74.3% F1 measure respectively.

Almost all of the ensemble methods have higher F1 measure over stand-alone classifiers on unbalanced dataset. However, on the hard voting strategy with 5 classifiers (NB, KNN, ME, RF, SVM), whose F1 measure is 77.9%, the ensemble methods can not exceed the NB performance. The decision in hard voting is equally determined by all of stand-alone classifiers. The F1 measure of hard voting method usually varies between the F1 measure of best classifier and the F1 measure of worst classifier. It is hard for the hard voting method to get higher F1 measure than the best classifier beacuse the difference in F1 measure is too far between the best classifier (78.3%) and the worst classifier (68.2%) that been combined. It is not happened when we use soft voting method. Soft voting method with 5 classifiers still surpass the performance of all stand-alone classifiers by 78.9% F1 measure. Although combining all of the classifiers, soft voting give votes for each category based on its average probability value from all of the classifiers. There is a possibility that winning categories based on hard voting will lose on soft voting because they have lower averages probability than other category. Soft voting simply provides a more robust voting scheme as it is often reduces overfit and creates a smoother model.

The ensemble methods by using only three best classifiers (NB, SVM, and RF) have the best performance when using hard voting or soft voting. Hard voting and soft voting based on this scheme have the same F1 measure, 79.8%. Since ensemble method is affected by the classifiers that compiled it, using only the best classifier can improve the possibility of ensemble method to get better performance.

Meanwhile, the result of the second experiment, which is using balanced dataset, can be seen in Figure 5. As predicted, all of the classification method got higher F1 measure on balanced dataset. KNN is still the worst classifier with F1 measure 76.8%. RF perfomed slightly better with 77.6% F1 measure. Surprisingly, ME has the best performance with 84.1% F1 measure value. NB and SVM are still below ME with only slight difference. All of the ensemble method have almost the same F1 measure value and also perform better than almost all of the stand-alone classifiers. These two exeperiments showed that we can improve the performance by using ensemble method even if not significant. Nevertheless, ensemble method surely reduce the jeopardy of choosing a week classifier to be used for detecting new tweets.

4. CONCLUSION

In this study, we we used ensemble method to for Hate Speech Detection in Indonesian language. We employed five stand-alone classification algorithms, including Naïve Bayes, K-Nearest Neighbours, Maximum Entropy, Random Forest, and Support Vector Machines, and two ensemble methods on Twitter hate speech dataset.

By using unbalance dataset, the experiment results show that Naïve Bayes offered the best performance among all five stand-alone classifiers with F1 measure value 78.3%. The experimental results also show that ensemble technique can improve the classification performance. The best result is achieved when using ensemble of three best classifier (Naïve Bayes, Support Vector Machine, and Random Forest) with F1 measure 79.8%.

Meanwhile, as predicted, all of the classification method got higher F1 measure when using balanced dataset. Surprisingly, Maximum Entropy has the best performance in this second experiment with 84.1% F1 measure value. Using balanced dataset, all of the ensemble method have almost the same F1 measure value and also perform better than almost all of the stand-alone classifiers. These two exeperiments showed that using ensemble method can improve the performance of the system. Although the improvement is not significant, using ensemble method can reduce the risk of selecting a poor classifier to be used for detecting new tweets as hate speech or not.

In the future work, instead of only using BOW features, applying ensembles of feature set maybe a promising direction to get better performance. Some feature set such as n-gram, lexicon, POS tagging, texual feature or twitter specific features can be applied for improvement. Another types of feature like Word2Vec or Paragraph2Vec also can be applied in the future.

REFERENCES

[1] Warner W, Hirschberg J, “Detecting hate speech on the World Wide Web”. In *Proceedings of the Second Workshop on Language in Social Media* 2012 Jun 7 (pp. 19-26). Association for Computational Linguistics.

[2] Sitorus AP, Murfi H, Nurrohmah S, Akbar A, “Sensing Trending Topics in Twitter for Greater Jakarta Area”, *International Journal of Electrical and Computer Engineering (IJECE).* 2017 Feb 1; 7(1):330-6.

[3] Burnap P, Williams ML, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making”. *Policy & Internet*. 2015 Jun 1; 7(2):223-42.

[4] Badjatiya P, Gupta S, Gupta M, Varma V, “Deep learning for hate speech detection in tweets”. In *Proceedings of the 26th International Conference on World Wide Web Companion,* 2017 Apr 3 (pp. 759-760). International World Wide Web Conferences Steering Committee.

[5] Waseem Z, Hovy D, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter”. InSRW@ HLT-NAACL 2016 Jun 12 (pp. 88-93).

[6] Kwok I, Wang Y, Locate the Hate: Detecting Tweets against Blacks. In *AAAI* 2013 Jul 14.

[7] Barbosa L, Feng J, “Robust sentiment detection on twitter from biased and noisy data”. In *Proceedings of the 23rd International Conference on Computational Linguistics*: Posters 2010 Aug 23 (pp. 36-44). Association for Computational Linguistics.

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

[9] Alfina I, Mulia R, Fanany MI, Ekanata Y. “Hate Speech Detection in the Indonesian Language: A Dataset and Preliminary Study”. In *Advanced Computer Science and Information Systems (ICACSIS), 2017 International Conference on 2017*. IEEE.

[10] Fauzi MA, Arifin AZ, Gosaria SC, “Indonesian News Classification Using Naïve Bayes and Two-Phase Feature Selection Model”. *Indonesian Journal of Electrical Engineering and Computer Science*. 2017 Dec 1;8(3).

[11] Suharno CF, Fauzi MA, Perdana RS, “Klasifikasi Teks Bahasa Indonesia Pada Dokumen Pengaduan Sambat Online Menggunakan Metode K-Nearest Neighbors Dan Chi-Square”. *Systemic: Information System and Informatics Journal*. 2017 Dec 7;3(1):25-32.

[12] El-Halees AM, “Arabic text classification using maximum entropy”. *IUG Journal of Natural Studies*. 2015 Dec 5;15(1).

[13] Wu Q, Ye Y, Zhang H, Ng MK, Ho SS, “ForesTexter: an efficient random forest algorithm for imbalanced text categorization”. *Knowledge-Based Systems*. 2014 Sep 30;67:105-16.

[14] Roli F. “Multiple classifier systems”, *Encyclopedia of Biometrics*. 2015:1142-7.

[15] Adeva JG, Atxa JP, Carrillo MU, Zengotitabengoa EA, “Automatic text classification to support systematic reviews in medicine”, Expert Systems with Applications. 2014 Mar 31;41(4):1498-508.

[16] Dong YS, Han KS, “A comparison of several ensemble methods for text categorization”. In *Services Computing, 2004.(SCC 2004). Proceedings. 2004 IEEE International Conference on*, 2004 Sep 15 (pp. 419-422). IEEE.

[17] Larkey LS, Croft WB, “Combining classifiers in text categorization”. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval,* 1996 Aug 18 (pp. 289-297). ACM.

[18] Tala F, “A study of stemming effects on information retrieval in Bahasa Indonesia”. *Institute for Logic, Language and Computation,* Universiteit van Amsterdam, The Netherlands. 2003 Jul.

[19] Pramukantoro ES, Fauzi MA, “Comparative analysis of string similarity and corpus-based similarity for automatic essay scoring system on e-learning gamification”, In *Advanced Computer Science and Information Systems (ICACSIS), 2016 International Conference on*, 2016 Oct 15 (pp. 149-155). IEEE.

[20] Fauzi MA, Arifin A, Yuniarti A, “Term Weighting Berbasis Indeks Buku dan Kelas untuk Perangkingan Dokumen Berbahasa Arab. Lontar Komputer”, *Jurnal Ilmiah Teknologi Informasi*. 2013;5(2).

[21] Fauzi, M.A., Utomo, D.C., Setiawan, B.D. and Pramukantoro, E.S, “Automatic Essay Scoring System Using N-Gram and Cosine Similarity for Gamification Based E-Learning”, In *Proceedings of the International Conference on Advances in Image Processing* (pp. 151-155). ACM.

[22] Fauzi MA, Arifin AZ, Yuniarti A, “Arabic Book Retrieval using Class and Book Index Based Term Weighting”, *International Journal of Electrical and Computer Engineering (IJECE),* 2017 Dec 1;7(6).

[23] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, “Scikit-learn: Machine learning in Python”, *Journal of Machine Learning Research*. 2011; 12(Oct):2825-30.