Cyberbullying Detection on Twitter using Multiple Textual Features

1st Jianwei Zhang
Faculty of Science and Engineering
Iwate University
Morioka, Japan
zhang@iwate-u.ac.jp

2nd Taiga Otomo
Graduate School of Science and Engineering
Iwate University
Morioka, Japan
g0319030@iwate-u.ac.jp

3rd Lin Li
School of Computer Science and Technology
Wuhan University of Technology
Wuhan, China
cathylilin@whut.edu.cn

4th Shinsuke Nakajima
Faculty of Information Science and Engineering
Kyoto Sangyo University
Kyoto, Japan
nakajima@cc.kyoto-su.ac.jp

***Abstract*—Due to the spread of PCs and smartphones and the rise of user-generated content in social networking service, cyberbullying is also increasing and has become a serious risk that social media users may encounter. In this paper, we focus on the Japanese text on Twitter and construct an optimal model for automatic detection of cyberbullying by extracting multiple textual features and investigating their effects with multiple machine learning models. The experimental evaluation shows that the best model with predictive textual features is able to obtain an accuracy of over 90%.**

***Index Terms*—cyberbullying, machine learning, text classification, textual feature**

I. Introduction

In recent years, due to the spread of PCs and smartphones and the rise of user-generated content in social networking service, cyberbullying is also increasing and has become a serious risk that social media users may encounter. Some reports indicate that this behavior can no longer be ignored. American national statistics show that in 2017, about 15% of high school students were bullied online or by text [1]. In Japan, according to a survey by the Ministry of Education, Culture, Sports, Science and Technology in October 2018, the number of cyberbullying awareness in 2017 was 12,632 and continues to increase year by year [2]. The main current method to detect cyberbullying in Japan is manual by questionaires or victims' reports. Research on cyberbullying is mainly conducted in the field of education and has been confined to the development of regulations and teaching materials. However, due to the massive amount of Web data, the effect of these methods is limited to the discovery of cyberbullying. On the other hand, in the field of IT, although there exist some researches on automatic detection of cyberbullying, most of them analyze English text and the features used for classifying cyberbullying are insufficient. There is a need for the technology that can extract predictive features and automatically detect cyberbullying with high accuracy for Japanese data.

This work was supported by JSPS KAKENHI Grant Number 19K12230.

In this paper, we aim at automatic cyberbullying detection and utilize machine learning methods to realize the purpose. Two most important aspects for classifying cyberbullying based on machine learning are what features to be extracted and what machine learning models to be selected. We focus on the Twitter text (hereinafter referred to as tweet) and intend to find the features that mostly contribute to cyberbullying detection. We use the technique of text mining and analyze a range of textual features including n-gram, Word2Vec, Doc2Vec, tweets' emotion values, and unique characteristics on Twitter. In addition, multiple machine learning models including linear models, tree-based models and deep learning models are investigated with multiple textual features to construct an optimal model. Based on the collected tweets, we evaluate the quality of automatic detection of cyberbullying, and find that the best model with predictive textual features can achieve the accuracy of over 90%.

The rest of this paper is structured as follows. Section 2 reviews related work. Section 3 provides an overview of the system. Section 4 describes data collection and preprocessing. The features and the machine learning models used for classification are introduced in Section 5 and Section 6 respectively. Section 7 reports the results of experimental evaluation. Finally, we conclude the paper and discuss future work in Section 8.

II. Related work

The majority of research on automatic cyberbullying detection has focused on English resources and intended to improve the accuracy of cyberbullying detection classifiers. Burnap et al. collected tweets on Twitter related to a murder incident using hashtags, and used n-grams and grammatical dependencies between words to automatically detect cyber hate speech with a focus on race, ethnicity, or religion [3]. Hosseinmardi

et al. collected Instagram data consisting of images and their associated comments, and investigated approaches to automatically detect incidents of cyberbullying over images by using the features such as the information of the image, the user who posted it, the interval between comments, n-grams [4]. Rafiq et al. collected a set of Vine video sessions, and designed approaches to automatically detect instances of cyberbullying over Vine media sessions by using the features such as the information of the video, the user who posted it, the emotion of the comment, n-grams [5]. Nobata et al. attempted to automatically detect abusive language from comments found on Yahoo! Finance and Yahoo! News by using n-grams, linguistic, syntactic and distributional semantics features [6]. Chatzakou et al. investigated bullying and aggressive behavior on Twitter and distinguished bullies and aggressors from regular users by using text, user, and network-based attributes [7]. In contrast with the work that focuses on the accuracy issue, Rafiq et al. attempted to address the issues of scalability and timeliness of cyberbullying detection systems by proposing a multi-stage cyberbullying detection solution comprised of a dynamic priority scheduler and an incremental classification mechanism [8].

A survey on cyberbullying detection [9] and the work of Notata et al. [6] have pointed out that it remains to be seen whether established approaches examined on English are equally effective on other languages. Van Hee et al. detected cyberbullying events on Dutch [10] and Ross et al. measured the reliability of hate speech annotations on German [11]. There is little work done with cyberbullying detection on Japanese. Ptaszynski et al. extracted patterns with a Brute-Force search algorithm and used them to classify Japanese harmful expressions [12]. They further improved the results by using deep learning models [13].

As far as we know, the investigated aspects for cyberbullying detection on Japanese resources are limited to only some basic features such as tokens, lemmas and POS. Our work intends to automatically detect cyberbullying from Japanese text on Twitter by studying not only surface features such as n-grams, but also multiple textual features including Word2Vec [14], Doc2Vec [15], emotional values [16], [17], Twitter-specific characteristics, and investigating the classification effect of multiple machine learning models with these features.

III. Overview of the proposed method

The flow of the proposed method is shown in Fig. 1.

1) Data collection (explained in Section IV)
   First, both the cyberbullying and non-cyberbullying tweets are collected from Twitter. The cyberbullying tweets are collected by retrieving Twitter with some bullying words and confirmed by crowds. The non-cyberbullying tweets are collected randomly.
2) Feature extraction (explained in Section V)
   After excluding unnecessary words from tweets, morphological analysis is performed. Then, textual features are extracted including n-gram, Word2Vec, Doc2Vec, emotion values of tweets, and Twitter-specific characteristics.
3) Model generation (explained in Section VI)
   The collected tweets are divided into training data and test data, and the models are constructed on the training data using each type of features and each type of machine learning algorithms. The machine learning algorithms include linear models (Linear support vector machine, Logistic regression), tree-based models (Decision tree, Random forest, Gradient boosting regression tree) and deep learning models (Multilayer perceptron). In addition, cross verification and grid search are used for constructing the best model.
4) Model evaluation (explained in Section VII)
   We evaluate how well the generated models can classify the cyberbullying and non-cyberbullying text on the test data. We use accuracy, precision, recall and F-measure as evaluation criteria.

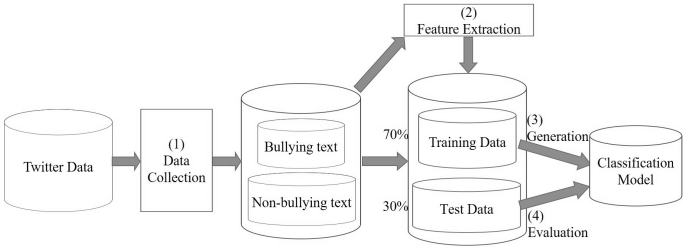

Fig. 1. Overview of the proposed method

IV. Data collection and preprocessing

A. Data collection

First, we collect tweets from Twitter. 36 Japanese bullying words introduced in previous researches [18], [19] are used for retrieving tweets either including any of the bullying words or tagged with any of the bullying words, and consequently 2,349,052 tweets are collected. Next, the tweets are ranked in descending order by the numbers of their bullying words and bullying tags and the top 3,450 tweets are shown to crowd workers. Each tweet is judged as cyberbullying or non-cyberbullying respectively by three workers. The tweets on which all the three workers make the same judgment are used as the following learning data. There are 1,395 tweets unanimously considered as cyberbullying and 282 tweets unanimously considered as non-cyberbullying. Since in this way the collected positive data (1,395 cyberbullying tweets) for learning are less than the negative data (282 non-cyberbullying tweets), we further randomly sample 1,113 non-cyberbullying tweets from Twitter checked by browsing their content. As a result, the number of non-cyberbullying tweets becomes 1,395, equalized to the number of cyberbullying tweets. Totally 2,790 tweets including 1,395 cyberbullying ones and 1,395 non-cyberbullying ones are used in our experiments.

B. Data preprocessing

Morphological analysis of 2,790 tweets is first performed before feature extraction. For morphological analysis, we exclude certain characters that are unnecessary in tweets. The excluded characters are URLs, Twitter-specific terms (e.g., RT), and some symbols (e.g., ♢, ★). Furthermore, newline characters, full-width spaces and multiple consecutive single-width spaces are converted into one single-width spaces.

MeCab [20], an open source text segmentation library for use with Japanese text, is used for morphological analysis. MeCab divides each tweet into word units, and attaches extra information to each word including the surface form (the form of the word on the tweet), the part of speech, the inflected form, the original form, the pronunciation, etc. "Word" that appears in Section V basically refers to the surface form. For the system dictionary of MeCab, we use mecab-ipadic-NEologd, that is customized by adding new words obtained from many Web language resources. The advantages of this dictionary are that it is updated twice every week and possible to cope with relatively new words and slang phrases emerging on the Web.

V. Feature extraction

Extraction of predictive features is especially important for distinguishing cyberbullying text from non-cyberbullying based on machine learning. In this paper, we investigate the features of n-grams, Word2Vec, Doc2Vec, emotion values of tweets, and Twitter-specific characteristics.

A. n-gram

N-gram is a language model that divides a string or document into n continuous characters or words and counts how many times they appear in the string or document. It is assumed that the probability of occurrence of a character or word depends on the previous character or word. We consider the characters or words that are often used for cyberbullying may be predictive features for classification. We use both character division (hereinafter character n-grams, n=2∼5) and word division (hereinafter word n-grams, n=1∼5). After extracting the character n-grams and word n-grams, a vector for a tweet is generated with each character n-gram or word n-gram as elements. The occurrence frequency of each element in the tweet is taken as the value of the element of the vector.

However, if all the n-grams are used as features, the amount of the features will become a huge number and there exist many elements that are obviously not important at all. Furthermore, most of the character n-grams are some combinations of characters that are meaningless in writing and tend to create noise. In order to solve this problem, we use principal component analysis on the extracted n-grams. Principal component analysis is a statistical method that uses an orthogonal transformation to convert possibly correlated features into linearly uncorrelated features. Generally, after the transformation, only some of the features that are important for representing the data are left for further analysis. The algorithm first finds the direction with the highest variance and labels it the "first principle component". The data have the most information in this direction. Next, among the directions orthogonal to the first principle component, the direction having the highest variance is selected as the "second principle component". The directions found in this way are called "principal components". Principal components have the same number as that of original features and are sorted in the descending order of importance for explaining the data. The number of principal components used for further analysis is usually decided by a parameter. It is possible to reduce the dimensions of feature space while only leaving the top components. In the experiments, we reduce each n-gram to 100 dimensions.

B. Word2Vec

Word2Vec is a method that vectorizes the meaning of a word taking a large corpus of text data as the input. It is possible to estimate semantic similarity between two words by calculating the inner product of the vectors of the words, or to understand the relationship between two words by performing addition/subtraction of the vectors of the words. We use Word2Vec to contribute to cyberbullying classification. In our experiment, Word2Vec is trained based on 2,349,052 tweets collected by using 36 bullying words described in Section IV-A. The word vector of each word is set to 100 dimensions. For each tweet, its Word2Vec feature is made by averaging the word vectors of the words contained in the tweet.

C. Doc2Vec

Doc2Vec is a method that vectorizes the meaning of a document. The difference from Word2Vec is that the order of words appearing in a sentence is taken into consideration, and that vectors are expressed in not a word unit but a document unit. These points make it possible to understand the relationship between documents. Doc2Vec is also trained based on 2,349,052 tweets, same with those for Word2Vec. The document (tweet) vector is also set to 100 dimensions. Since Doc2Vec expresses text in a document unit, each tweet can directly be represented by a Doc2Vec vector.

D. Emotion values of tweets

We extract emotion values of tweets as features for classification. Emotion values of tweets are calculated using the emotion dictionaries, in which each entry indicates the correspondence of a word and its emotion value. It is common that the cyberbullying text may probably contain negative words and thus the emotion of the text tends to be negative. In this paper, we use two emotion dictionaries: a three-dimension dictionary proposed by Zhang et al. [16] and a list of semantic orientations of words proposed by Takamura et al. [17].

1) Three-dimension dictionary: The three-dimension emotions are Happy⇔Sad, Glad⇔Angry, and Peaceful⇔Strained, which are decided based on a statistical analysis and a clustering analysis of 42 emotion words. A emotion value of a word on each dimension is a value between 0 and 1. The values close to 1 mean the emotions of the words are close to

Happy, Glad, or Peaceful, while the values close to 0 mean the emotions of the words are close to Sad, Angry, or Strained. The dictionary is constructed by examining the co-occurrence of each target word and two word sets having contrastive emotions in a newspaper corpus, under the assumption that one word (e.g., prize) with a certain emotion (e.g., Happy) is easy to co-occur with the emotional words (e.g., Happy, Enjoy, Joy) expressing this emotion, but hard to co-occur with the emotional words (e.g., Sad, Grieve, Sorrow) expressing the opposite emotion (e.g., Sad).

2) Semantic orientations of words: The semantic orientation of a word shows whether the word expresses positive emotion or negative emotion. The words registered in this emotion dictionary are extracted from Japanese Iwanami dictionary and the orientations of the words are computed based on a lexical network. The method assigns each word a score in the range of -1 to +1, where the words with values close to -1 are assumed to be negative and the words with values close to +1 are assumed to be positive.

3) Calculation of emotion values for tweets: For each tweet, we first perform a morphological analysis with MeCab to acquire the surface form and original form of each word. Next, we check whether the surface form or the original form of each word in the tweet is included in each emotion dictionary. If it exists in an emotion dictionary, the corresponding emotion value is recorded to a list. If it is not included in an emotion dictionary, the median value of each emotion dictionary (0.5 for the three-dimension dictionary, and 0 for semantic orientations of words) is recorded to the list. After all the words in the tweet are examined, the average value of the list is used as the emotion value of the tweet.

E. Twitter-specific characteristics

Twitter provides the functions such as "retweet" to repeat other people's tweet, "favorite" to register other people's tweet as favorite, and "hashtag" to categorize own tweet so as to facilitate search. For each tweet, the information of the numbers of retweet, favorite, hashtag, and URLs in it is also used as features for classification.

VI. Selection of machine learning models

Selection of machine learning models is also important for cyberbullying detection. We compare multiple machine learning algorithms including linear models (Linear support vector machine, Logistic regression), tree-based models (Decision tree, Random forest, Gradient boosting regression tree) and deep learning models (Multilayer perceptron).

VII. Experimental evaluation

A. Experimental setup

Using the collected tweets, experiments are conducted to confirm how correctly cyberbullying texts can be classified with each type of features and each type of machine learning models, and which features and which machine learning models are especially useful for classification. The tweets used for classification are 1,395 cyberbullying tweets and 1,395 non-cyberbullying tweets.

TABLE I
Feature types and contents

Feature type	Feature content
Word n-gram	Word unigrams after principal component analysis (100 dimensions)
	Word bigrams after principal component analysis (100 dimensions)
	Word trigrams after principal component analysis (100 dimensions)
	Word 4-grams after principal component analysis (100 dimensions)
	Word 5-grams after principal component analysis (100 dimensions)
Character n-gram	Character bigrams after principal component analysis (100 dimensions)
	Character trigrams after principal component analysis (100 dimensions)
	Character 4-grams after principal component analysis (100 dimensions)
	Character 5-grams after principal component analysis (100 dimensions)
Word2Vec	Average of word vectors (100 dimensions)
Doc2Vec	Document vector (100 dimensions)
Emotion values of tweets	Happy⇔Sad
	Glad⇔Angry
	Peaceful⇔Strained
	Positive⇔Negative
Twitter-specific characteristics	# of retweets
	# of favorites
	# of hashtags
	# of URLs

For testing the predictive effects of features, firstly six types of features (Tab. I) are used individually, including word n-gram (n=1∼5), character n-gram (n=2∼5), Word2Vec, Doc2Vec, emotion values (Happy⇔Sad, Glad⇔Angry, Peaceful⇔Strained, and Positive⇔Negative) and Twitter-specific characteristics (numbers of retweet, favorite, hashtag and URL). Next, we intend to confirm whether the emotion feature can further strengthen the classification effect when it is used with other type of features. Specially, the type of features individually achieving the best classification results (character n-gram) are combined with the emotion feature to test the classification effect. Finally, all types of features are used together for examining the classification effect. If all the word n-grams and character n-grams are used, feature space will have too many dimensions. Thus, for the type of word n-grams, we use decision tree to confirm which n-grams mostly contribute to classification and find that word unigram and word bigram are more predictive than others. In the same way, for the type of character n-grams, character 4-gram is found the most predictive. Word unigram, word bigram, and character 4-gram are finally used when testing the combination of all types of features.

For all types of machine learning models, 70% of tweets are used as training data and 30% as test data. Based on the training data, stratified 5-fold cross-verification is performed for fairly evaluating each model. The training data are split into five folds, one of which is left for evaluating the model and other four of which are used for training the model. This process is repeated five times by using each of five folds as evaluation data, and the accuracies of five times are averaged as the final accuracy of the model. With cross-verification, we can obtain more robust evaluation results. Furthermore, we use grid search to find the appropriate parameters for each machine learning model. Given each combination of parameters, the above-mentioned cross-verification is performed to test the accuracy of the model with this setting. We compare differ-

ent machine learning models using the highest classification accuracy that each model achieves by the optimal parameters.

We confirm how correctly the test data can be classified with each generated model. For the evaluation on the test data, four criteria are used: accuracy, precision, recall, and F-measure. Accuracy is the ratio of the correctly judged tweets (both cyberbullying tweets and non-cyberbullying tweets) to all the tweets. Precision is the ratio of the true cyberbullying tweets to the tweets that the model has classified as cyberbullying. Recall is the ratio of the true cyberbullying tweets that the model has identified to all the true cyberbullying tweets. F-measure is a harmonic mean of precision and recall. The equations for calculating each criterion are as follows.

$$\begin{aligned} accuracy &= \frac{TP+TN}{TP+FP+TN+FN} \\ precision &= \frac{TP}{TP+FP} \\ recall &= \frac{TP}{TP+FN} \\ F-measure &= \frac{2*recall*precision}{recall+precision} \end{aligned}$$

where TP represents True Positive (the number of the tweets which are judged as cyberbullying with the model and are true cyberbullying), TN represents True Negative (the number of the tweets which are judged as non-cyberbullying with the model and are true non-cyberbullying), FP represents False Positive (the number of the tweets which are judged as cyberbullying with the model but are false cyberbullying), and FN represents False Negative (the number of the tweets which are judged as non-cyberbullying with the model but are false non-cyberbullying).

B. Experimental results

The evaluation results are shown in Tab. II. The rows in the table represent different types of machine learning models and the columns represent different types of textual features. Specially, the first six columns of features are individual types of features, and the last two columns of features are the best individual type of feature (character n-gram) combined with the emotion feature and the combination of all types of features. The cells show the values of four criteria (accuracy, precision, recall and F-measure) for corresponding features and machine learning models. For one column (one type of feature), the best machine learning model is marked in bold type. For one row (one type of machine learning model, the best feature is marked in underlined type.

For individual types of features, character n-gram is the best, followed by Word2Vec and word n-gram. Doc2Vec acquires good results for some models, but has low values for Linear support vector machine, Decision tree and Multilayer perceptron. Emotion features and Twitter-specific features do not work well alone. In summary, we can see that the features of character n-gram, word n-gram and Word2Vec can greatly contribute to the correct classification.

As for the machine learning models, stable and high classification quality for this task is achieved by Logistic regression and Gradient boosting regression trees, followed by Random forest and Decision tree. Multilayer perceptron shows some good results with some features but has low values with other features. Linear support vector machine does not show particularly good results with each type of features. In summary, based on different types of features, the machine learning models with which the best classification results are achieved are Logistic regression and Gradient boosting regression tree. The experimental results show these models can greatly contribute to the correct classification for this task.

Comparing the results individually achieved by character n-gram with those of "character n-gram + emotion values", the difference of evaluation results is not obvious. The emotion feature combined with character n-gram can slightly improve the classification for some models, also slightly decrease the evaluation values for some models, in contrast with individual character n-gram. Overall, the highest values are achieved with all types of features and the models of Logistic regression or Gradient boosting regression tree, which exceed 90% for all the four evaluation criteria.

C. Discussion

We discuss the effect of features. N-grams show good classification quality mainly due to the reason that tweets labeled with cyberbullying in this study are collected using 36 specific bullying words, tweets labeled with non-cyberbullying rarely contain these bullying words, and thus n-grams in two data sets have clear difference. However, the classification quality is unclear if a cyberbullying text only contains bullying words which do not appear in the collected learning data or even has no bullying words such as sarcastic expressions.

The good quality achieved by Word2Vec and Doc2Vec is also considered due to the same reason as n-grams. However, the Word2Vec and Doc2Vec models are constructed based on the data set mainly collected by using the specific bullying words, which may fail to cover other words that are used in cyberbullying. Since the current n-gram, Word2Vec and Doc2Vec models may not cope well with the text containing unseen bullying words, a wider collection of cyberbullying text and the model generation on it will be our future work.

The reason that the emotion feature does not show high classification quality may be due to the coverage rate of emotion words in the emotion dictionaries. The vocabulary in the three-dimension dictionary is extracted from newspaper articles and the vocabulary in semantic orientations of words is taken from a Japanese language dictionary. In contrast with these formal words, many of the words in tweets are informal or newly emergent, which are not included in the two emotion dictionaries. Therefore, we consider constructing a new emotion dictionary corresponding to Twitter data and specialized to cyberbullying.

Twitter-specific characteristics do not cause sufficiently good results. This may be because the numbers of retweet, favorite, hashtag, or URL are commonly related to various factors and it is not reasonable enough to identify cyberbullying only by this type of features. Notice that in this study we

TABLE II
EVALUATION RESULTS (FROM TOP TO BOTTOM ARE ACCURACY, PRECISION, RECALL AND F-MEASURE)

	Word n-gram	Character n-gram	Word2Vec	Doc2Vec	Emotion values of tweets	Twitter-specific characteristics	Character n-gram+ Emotion values	All features
Linear SVM	0.488	0.512	0.827	0.725	0.727	0.697	0.524	0.488
	0.000	0.616	0.915	0.886	0.728	0.652	0.629	0.000
	0.000	0.123	0.731	0.530	0.745	0.873	0.170	0.000
	0.000	0.206	0.812	0.663	0.736	0.747	0.268	0.000
Logistic regression	0.892	**0.929**	**0.913**	0.872	0.737	0.702	**0.933**	**0.934**
	0.935	**0.936**	0.921	0.902	0.741	0.657	**0.932**	**0.934**
	0.848	0.925	**0.908**	0.841	0.745	0.871	**0.936**	0.936
	0.889	**0.930**	**0.915**	0.870	0.743	0.749	**0.934**	**0.935**
Decision tree	0.864	0.908	0.818	0.727	0.745	0.715	0.908	0.902
	0.879	0.903	0.872	0.731	0.708	0.664	0.903	0.882
	0.852	0.918	0.754	0.738	0.852	0.897	0.918	0.932
	0.865	0.910	0.809	0.734	0.774	0.763	0.910	0.906
Random forest	0.841	0.890	0.880	0.836	**0.769**	**0.719**	0.892	0.919
	0.845	0.878	0.912	0.837	**0.752**	**0.668**	0.884	0.918
	0.843	0.911	0.848	0.843	0.817	0.894	0.908	0.925
	0.844	0.894	0.878	0.840	**0.783**	**0.765**	0.896	0.922
Gradient boosting regression tree	**0.912**	0.925	0.904	**0.875**	0.755	0.715	0.918	0.933
	0.917	0.919	0.914	0.880	0.737	0.663	0.912	0.924
	0.911	**0.936**	0.897	**0.876**	0.808	**0.899**	0.929	**0.946**
	0.914	0.928	0.905	**0.878**	0.771	0.763	0.921	**0.935**
Multilayer perceptron	0.879	0.919	0.719	0.556	0.525	0.545	0.913	0.909
	0.899	0.914	**0.961**	**0.983**	**0.942**	0.566	0.927	0.909
	0.859	0.929	0.469	0.135	0.077	0.476	0.901	0.913
	0.879	0.922	0.631	0.238	0.142	0.517	0.914	0.911

conduct the automatic detection of cyberbullying in the tweet unit. If cyberbullying is analyzed in the user unit (not tweet unit), a user's action summary (the numbers of his retweets, etc.) may be effective features for classification.

VIII. CONCLUSIONS

In this paper, we have proposed the method to automatically detect cyberbullying text from Twitter. Multiple textual features and multiple machine learning algorithms are used to construct classification models. With the experiments based on the collected tweets, the quality of automatic cyberbullying detection is evaluated and the best model performs over 90% for the four criteria: accuracy, precision, recall, and F-measure.

Although the classification quality is high with these features and models, there is still much work to do in the future. Since the text used in this study is limited, we are planing to collect a wider range of texts with a bootstrapping method: extracting new bullying words from the current tweets and using them to obtain new tweets iteratively. An emotion dictionary specialized for cyberbullying and corresponding to the expressions on Twitter or newly emergent words will be constructed. Cyberbullying detection in user unit is also one of our research direction so as to find bullying users.

REFERENCES

[1] stopbullying.gov. Facts About Bullying, 2017. https://www.stopbullying.gov/media/facts/index.html.

[2] Investigation results in Japan, 2018. http://www.mext.go.jp/b_menu/houdou/30/10/1410392.htm.

[3] P. Burnap, M. L. Williams. Cyber Hate Speech on Twitter: An Application of Machine Classification and Statistical Modeling for Policy and Decision Making. Policy & Internet, vol.7, no.2, pp.223-242, 2015.

[4] H. Hosseinmardi, S. A. Mattson, R. I. Rafiq, R. Han, Q. Lv, S. Mishra. Analyzing Labeled Cyberbullying Incidents on the Instagram Social Network. SocInfo 2015, pp.49-66, 2015.

[5] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, S. A. Mattson. Careful what you share in six seconds: Detecting cyberbullying instances in Vine. ASONAM 2015, pp.617-622, 2015.

[6] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang. Abusive Language Detection in Online User Content. WWW 2016, pp.145-153, 2016.

[7] D. Chatzakou, N. Kourtellis, J. Blackburn, E. D. Cristofaro, G. Stringhini, A. Vakali. Mean Birds: Detecting Aggression and Bullying on Twitter. WebSci 2017, pp.13-22, 2017.

[8] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra. Scalable and Timely Detection of Cyberbullying in Online Social Networks. SAC 2018, pp.1738-1747, 2018.

[9] A. Schmidt, M. Wiegand. A Survey on Hate Speech Detection using Natural Language Processing. SocialNLP@EACL 2017, pp.1-10, 2017.

[10] C. Van Hee, E. Lefever, B. Verhoeven, J. Mennes, B. Desmet, G. De Pauw, W. Daelemans, V. Hoste. Detection and Fine-Grained Classification of Cyberbullying Events. RANLP 2015, pp.672-680, 2015.

[11] B. Ross, M. Rist, G. Carbonell, B. Cabrera, N. Kurowsky, M. Wojatzki. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. NLP4CMC 2017, pp.6-9, 2017.

[12] M. Ptaszynski, F. Masui, Y. Kimura, R. Rzepka, K. Araki. Automatic Extraction of Harmful Sentence Patterns with Application in Cyberbullying Detection. LTC 2015, pp.349-362, 2015.

[13] M. Ptaszynski, J. K. K. Eronen, F. Masui. Learning Deep on Cyberbullying is Always Better Than Brute Force. LaCATODA 2017, pp.3-10, 2017.

[14] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean. Distributed Representations of Words and Phrases and their Compositionality. NIPS 2013, pp.3111-3119, 2013.

[15] Q. Le, T. Mikolov. Distributed Representations of Sentences and Documents. ICML 2014, pp.1188-1196, 2014.

[16] J. Zhang, K. Minami, Y. Kawai, Y. Shiraishi, T. Kumamoto. Personalized Web Search Using Emotional Features. CD-ARES 2013, pp.69-83, 2013.

[17] H. Takamura, T. Inui, M. Okumura. Extracting Semantic Orientations of Words using Spin Model. ACL 2005, pp.133-140, 2005.

[18] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, K. Araki. Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. IJCNLP 2013, pp.579-586, 2013.

[19] S. Hatakeyama, F. Masui, M. Ptaszynski, K. Yamamoto. Statistical Analysis of Automatic Seed Word Acquisition to Improve Harmful Expression Extraction in Cyberbullying Detection. IJETI, vol.6, no.2, pp.165-172, 2016.

[20] T. Kudo, K. Yamamoto, Y. Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. EMNLP 2004, pp.230-237, 2004.