International Journal of Technology and Engineering Studies
volume 4 issue 1 pp. 7-15 doi: https://dx.doi.org/10.20469/ijtes.4.10002-1

Flame Prediction Based on Harmful Expression Judgement Using Distributed Representation

Kazuyuki Matsumoto*
Faculty of Science and Technology, Tokushima University, Tokushima, Japan

Seiji Tsuchiya
Faculty of Science and Engineering, Doshisha University, Kyoto, Japan

Takeshi Miyake
Shikoku Instrumentation Company Limited, Kagawa, Japan

Minoru Yoshida
Faculty of Science and Technology, Tokushima University, Tokushima, Japan

Kenji Kita
Faculty of Science and Technology, Tokushima University, Tokushima, Japan

Abstract: In recent years, flaming-that is, hostile or insulting interaction-on social media has been a problem. To avoid or minimize flaming, enabling the system to automatically check messages before posting to determine whether they include expressions that are likely to trigger flaming can be helpful. We target two types of harmful expressions: insulting expressions and expressions that are likely to cause a quarrel. We first constructed an original harmful expressions dictionary. To minimize the cost of collecting the expressions, we built our dictionary semi-automatically by using word distributed representations. The method used distributed representations of harmful expressions and general expressions as features, and constructed a classifier of harmful/general expressions based on these features. An evaluation experiment found that the proposed method was able to extract harmful expressions with an accuracy of approximately 70%. The proposed method was also able to extract unknown expressions; however, it tended to wrongly extract non-harmful expressions. The method is able to determine unknown harmful expressions not included in the basic dictionary and can identify semantic relationships among harmful expressions. Although the method cannot presently be applied directly to multi-word expressions, it should be possible to add such a capability by introducing time-series learning.

Keywords: *Flame prediction, harmful expression, distributed representation, support vector machine*

Received: 06 September 2017; ***Accepted:*** 21 November 2017; ***Published:*** 12 February 2018

I. INTRODUCTION

With the development of the Internet, social media, such as Social Networking Services (SNS), have been popular and widely used. By using such a service, an individual, company or association can easily transmit information, including opinions and advertisements, to a large audience. Some problems, however, have emerged. Flaming, whereby individuals or companies are criticized or confronted in a hostile manner by a third party, can cause serious damage. The leaking of personal information and accusations or harassment by e-mail or telephone can produce a severe deterioration of human relations and have a negative effect on individuals or organizations that can end in the loss of social credibility. To minimize these risks, finding ways to eliminate flaming should be an important priority.

There are various reasons for flaming. The targeted contents may have been too radical or offensive, having been posted without careful consideration by users with

*Correspondence concerning this article should be addressed to Kazuyuki Matsumoto, Faculty of Science and Technology, Tokushima University, Tokushima, Japan. E-mail: matumoto@is.tokushima-u.ac.jp

a shallow understanding of SNS. Many of the targeted users fail to understand the reasons why their posts were criticized and why the flaming occurred. By checking posted contents with an objective perspective and suggesting revisions by pointing out potential problems, we believe that flaming can be prevented or, at least, reduced. One way to achieve this goal is to collect examples of posts that caused flaming and assess the possibility of flaming for various cases by training the collected examples. However, collecting such data from comments or statements that cause flaming is difficult as the targeted posts are often quickly deleted.

Our study focused on expressions of malicious slander, which were thought to be more easily collected, as a key for flame prediction. We collected harmful expressions and trained their distributed expressions. We then created a classifier to judge whether an expression is harmful by using machine learning.

II. RELATED WORKS

There are numerous studies dealing with flaming in the engineering field, and various approaches to identify potential flame-inducing posts have been proposed. [1] proposed a moral judgment system by using distributed representations and association information. [2] proposed a way to identify flaming on CGM and its applications. To create a harmful expression dictionary, it would seem cost-prohibitive to manually collect the necessarily large number of harmful expressions. Therefore, we attempted to create a small dictionary manually, then automatically extend the small dictionary.

[3] extracted expressions of malicious slander from the Japanese Bulletin Board System (BBS) 2 channel [4]. The purpose of their study was to create a dictionary by automatically extracting expressions. They constructed a small malicious expressions dictionary by manually collecting expressions from the 2 channel and created a model of words neighboring malicious expressions. They used this model to further extract malicious expressions and expand the dictionary. Although their method was able to extract new expressions, it is difficult to significantly expand the dictionary without using a large-scale corpus that included malicious expressions from the BBS website.

[5] accomplished Multi-level Classification by extracting features at different conceptual levels, achieving greater accuracy than statistical or rule-based models in detecting offensive language. [6] proposed an LSF framework for a system that extracts lexical features and syntactic features from the users conversation history and judges offensiveness based on the framework rules. Their system is able to estimate the degree of offensiveness by extracting style features, structure features, and cyberbullying features from the sentences in the users conversation. The approach performed more effectively in detecting offensive language than conventional methods, achieving 98.24% precision and 94.34% recall.

[7] calculated semantic orientation scores with PMI-IR [8] based on seed words in three categories and succeeded in detecting cyberbullying entries with greater accuracy than the baseline method. [9] statistically analyzed automatic seed word acquisition to extract harmful expressions in cyberbullying cases. However, the cover rate of extraction tends to be low for such methods as they use comparatively small-sized seed word collections. In addition, because cyberbullies commonly use jargon or euphemisms to avoid discovery, it can be difficult to obtain direct co-occurrences with seed words.

We propose a more versatile method that can determine indirect similarities and relationships by using word distributed representations obtained from large-scale corpora without the need to establish a direct co-occurrence relationship among words. However, we recognize that harmfulness may be judged much differently for jargon depending on the character of the corpus used for training. It should also be noted that our method cannot directly judge the harmfulness of a sentence or phrase, but rather is limited to judging the harmfulness of individual words.

While existing studies tend to focus on malicious expressions, our study also targeted discriminatory expressions and other words that might cause flaming. We believed that discriminatory expressions should be monitored and restricted. By controlling the use of such expressions and words that were likely to cause flaming responses, it is possible that flaming might be prevented.

Our approach used word distributed representations produced by word 2 vec [10] as a feature. If only expanding a corpus, it would be easy to extract new harmful expressions resembling the words in the seed expression dictionary.

III. PROPOSED METHOD

A. *System Overview*

Fig. 1 shows the flow of constructing the harmful expressions dictionary. We first created a small harmful expressions dictionary by extracting harmful expressions from social media, such as Twitter and weblog, as well as books.

Next, a corpus that collected texts from Twitter at random was tokenized by the morphological analyzer MeCab [11]. Based on this tokenized corpus, the small

harmful expressions dictionary was trained with the word distributed representations by word2vec.

The harmful expressions dictionary was automatically expanded by extracting words with similar distributed representations to words included in the harmful expressions dictionary. Because automatic expansion sometimes causes the inclusion of semantically dissimilar words as noise words, we manually chose the similar words to refine the dictionary. Fig. 2 shows the construction flow of the harmful judgement model based on the word distributed representations. To create a classifier to judge flaming, we assembled another dictionary - a general expressions dictionary - by collecting general expressions from word dictionaries, etc. We used the word distributed expressions of the words included in the harmful expressions dictionary and the general expressions dictionary as features, and trained with a harmful judgment model (to classify words into the harmful or general class) by using Support Vector Machine [12].

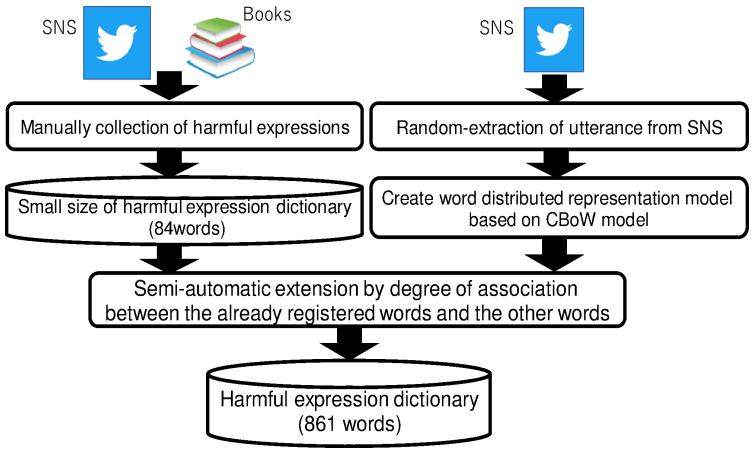

Fig. 1. Construction of the harmful expression dictionary

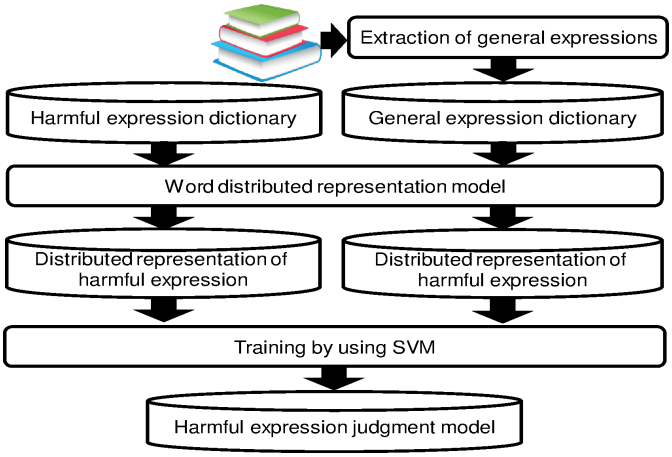

Fig. 2. Construction of harmful expression judgment model based on the distributed representation

B. *Extraction of Harmful Expressions*

1) Definition of harmful expressions: We targeted two types of harmful expressions:

1. Insulting expressions, including slanders and discriminative comments that are used to abuse or ridicule someone.

2. Expressions under the category of taboos or controversial/delicate issues regarding politics, history or religion.

TABLE 1
EXAMPLES OF HARMFUL EXPRESSIONS

Expression	Type Example
Insult expressions	Idiot, coward, spastic, dementia, hwabyung, etc.
Expressions under the category of taboos or controversial/delicate issues	Politics, persuasion, war, etc.

2) Process of extraction: First, we manually collected 84 harmful expressions from Twitter, news sites, weblog articles, BBS, and books. These 84 words were included in a small harmful expressions dictionary called the Basic Dictionary.

We used the dictionarys definition of harmful expression as the basis for our collection. To collect expressions similar to the expressions included in the Basic Dictionary, we trained the words distributed representation using word2vec.

We randomly collected one million tweets from Twitter and trained the words distributed representation model by inputting the tweets tokenized by a morphological analyzer as the training data. The vector dimension parameter was set at 200, and window size, that is, the maximum number of context words, was set at five; the sample value, which means the frequency with which to ignore words, was set at 0.001.

The default setting was used for the other parameters. Fig. 3 shows the mapping to 2-dimensional space of typical harmful expressions by using *t*-SNE algorithm [13].

We calculated the similarities between the vectors of words included in the Basic Dictionary and the vectors of words included in the word's distributed representation model, and collected similar expressions.

We manually selected the harmful expressions from the collected expressions. From a selection of approximately 3000 words, we identified 861 words as harmful expressions. We then defined a dictionary that included these 861 words as our harmful expressions dictionary.

C. Extraction of General Expressions

1) Definition of general expression: For our study, a "general expression" was defined as an expression used by the public that is not considered harmful. To define such expressions, we referred to "Nihongo no goitokusei Vol. 9." [14], where a familiarity value is annotated to each Japanese word. A words familiarity value represents the degree of public familiarity with the word, measured on a scale from 1 to 7 (1: not familiar, 7: very familiar).

For the pre-test, the target subjects were forty 18-30-year-old Japanese males and females (male: 20, female: 20). Each subject was given a word familiarity test; the average values among the subjects were then used as a words familiarity value. The data for 34 of the 40 subjects were used because of their high reliability. Table 2 shows an example of a words familiarity value.

In this paper, we defined words with a high familiarity value and not harmful as general expressions, and constructed a general expression dictionary as a negative example to train harmful expressions classification.

TABLE 2
EXAMPLES OF WORD'S FAMILIARITY VALUE

Word	Word's Familiarity Value
Starry sky	6.176
Arcade	5.688
Polarization	3.824
Aozari	1.750

2) Process of Extraction: We extracted 1,384 words that had a familiarity value over 6.3 and manually selected words that were included in the trained word's distributed representation model and judged to be not harmful expression. We then randomly chose 900 words from the expressions and constructed a dictionary that included these 900 words as our general expressions list. Table 3 shows the extracted expressions.

TABLE 3
LIST OF GENERAL EXPRESSIONS

Word	Word's familiarity value
Ice cream	6.562
Honesty	6.375
Noisy	6.531
Meet	6.594

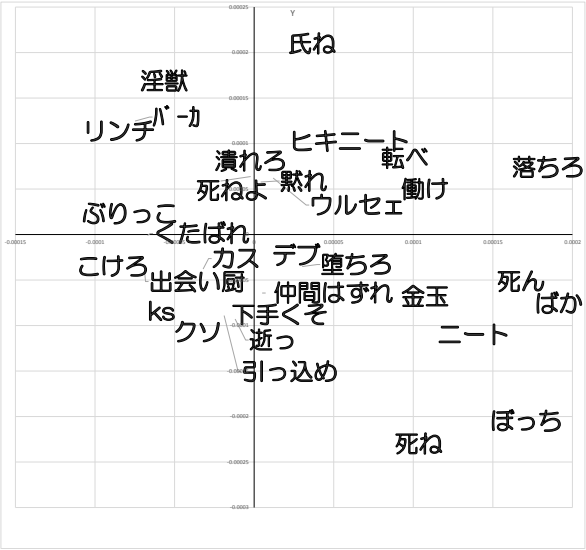

Fig. 3. Mapping to 2-dimensional space of typical harmful expressions by using *t*-SNE

D. Construction of Flame Judgment System

1) Creation of classifier: We trained the classifier based on SVM by using the collected harmful expressions dictionary and general expressions list as positive/negative examples. We converted the 864 harmful expressions and 900 general expressions into vectors by using the words distributed representation model. We created the harmful expressions classifier by using the vectors as training data and set the hyper parameters at the default setting.

2) Validation of classifier: We confirmed the accuracy of the classifier by applying a 10-fold cross-validation test.

TABLE 4
RESULT OF CROSS-VALIDATION

Averaged accuracy	± 2σ range
0.96	(+/- 0.03)

The results of the cross-validation are shown in Table 4. Accuracy and a ± 2σ range are shown in the table. The ± 2σ range was obtained by doubling the standard variation; it indicates the variation of the results. We found that the accuracy of the classifier was high and that the variation of the results was quite small.

3) Flame judgment system: We constructed a system to judge the possibility of flaming by using the created classifier. The system split the input sentence into units of words by morphological analysis. If a word was included in the words distributed representation model, we judged whether it was harmful by inputting the words vector into the harmful expressions classifier. As a harmful expression may cause flaming, the system displays the word and the word's harmfulness score (SVM judgment score). The flow of the system and an example are shown in Fig. 4, Fig. 5 shows the web application graphical user interface of the flame judgement system.

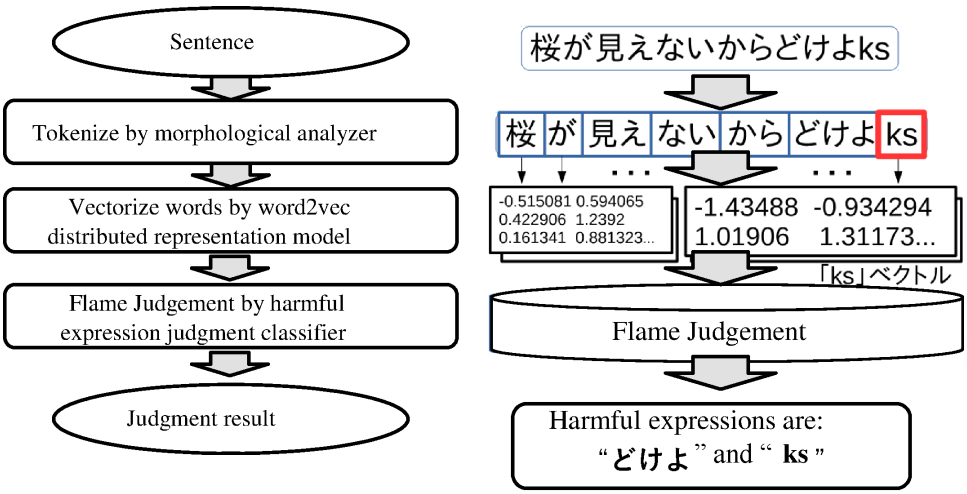

Fig. 4. Flow of flame judgement system

Fig. 5. Web GUI of the flame judgement system

IV. EVALUATION

A. Experimental Data

The official Twitter account [15], which is tweeted by Sanrio Company, Ltd., has received many malicious replies to its posts. We chose the malicious sentences contained in these replies to evaluate the performance of our flame prediction system.

TABLE 5
EXPERIMENTAL DATA

Data	Number of Data
Malicious sentences to Cinnamon	108 sentences
Number of correct expressions	35 words

Table 5 shows the experimental data. In all, 108 malicious sentences were collected from tweets posted in 2015. Table 6 shows examples.

TABLE 6
EXAMPLES OF MALICIOUS SENTENCES TO CINNAMON

Malicious Sentences	Correct Expression
Release the block, and fuck off and die in an accident.	Fuck off and die
Drop from the cloud.	Drop
Please stop acting cute.	acting cute

We manually checked these malicious sentences and judged 35 expressions to be harmful. These 35 expressions were to be regarded as correctly judged by our system if the system identified them as harmful. The list of these expressions is shown in Table 7.

TABLE 7
CORRECT ANSWERS

List of Correct Expressions			
Fuck off and die	slip	drop	die
Dating addiction	loneliness	agots	Work
Fall low	ks	Shut up	beast
Idiot	booby	coy	bugger
Delete	trip	hate	muggle
Fall	Die	annoying	crap
Recede	cow pattie	NEET	Chaff
Go to hell	Crumple up	lynch	porky
hiki-NEET	Pass away	dying	

B. Parameters of Classifier

The harmful expression classifier was constructed by using Support Vector Machines included in the Python package Scikit-learn [16]. We used word distributed representation vectors generated by word2vec as features.

The dimension of the vector was 200; the parameter setting was the default. The experiment was conducted with the C value set as 1.0, the gamma value = (1/number of feature), and the kernel type set as RBF.

We used the soft-margin SVM. The soft-margin SVM allows data (instances) to be included in a margin. Allowing data to be included within a margin and adding a penalty to the data within the margin enable us to create a separating hyperplane with high general versatility. The C value controls the trade-off between penalty and margin. When the C value is small, error is allowed; when the C value is large, no error is allowed. The gamma value indicates the complexity of the hyperplane. A smaller value means a simpler hyperplane; a bigger value means a more complex hyperplane. In this study, a simple separating hyperplane that allows small errors was used.

C. Evaluation Method

We evaluated the system according to the number of expressions that it could correctly judge harmful among the 35 previously-determined harmful expressions. We examined the recall, precision, and false detection rate (calculated as the frequency with which the system incorrectly judged a non-harmful expression to be harmful). The calculation formulas for recall, precision, and false detection are shown in Equation 1, Equation 2, and Equation 3.

$$Recall = \frac{N_c}{N_a} \tag{1}$$

$$Precision = N\frac{c}{N_o} \tag{2}$$

$$FalseDetectionRate = \frac{N_e}{N_o} \tag{3}$$

N_c indicates the number of expressions that were correctly judged as harmful by the system. N_a indicates the total number of the correct expressions. N_o indicates the number of words that were judged as harmful by the system. N_e shows the number of non-harmful expressions that were wrongly judged as harmful. We also evaluated how accurately the harmful expressions would be judged using a harmful expressions dictionary without SVM.

D. Experimental Results

Experimental results are shown in Table 8. As indicated, the system was able to judge unknown harmful expressions that were not included in the harmful expressions dictionary.

TABLE 8
EXPERIMENTAL RESULTS

	Proposed Method	Only Dictionary
Success Rate (%)	26.0	10.0
False Detection Rate (%)	20.0	0.0
Recall (%)	74.3	28.6
Precision (%)	56.6	100

In addition, all of the expressions included in the harmful expressions dictionary were successfully judged as harmful. Table 9 shows the expressions that were successfully judged. Table 10 shows the expressions that were undetected. As can be seen, rows of w were often judged as harmful. Words expressing disease or injury were included among them. Instruction words were also included. From our results, it appears that the proposed method does not always classify a word properly when the word has semantic polysemy. However, jargon is usually created from a common word by adding another meaning. The proposed method may be able to deal with the semantic polysemy of expressions if the distributed representations are defined from multiple viewpoints by splitting the learning corpus according to the users attributes or community.

V. DISCUSSION

Twenty-six harmful expressions were correctly judged by the proposed method. This exceeded the performance of a simple matching method based on the harmful expressions dictionary.

This result showed that classification by SVM using distributed representation expressions as features was efficient for correctly judging a harmful expression as harmful. On the other hand, 20 general expressions were incorrectly judged to be harmful expressions.

As shown in Table 10, we found there were many rows of w in the falsely detected expressions. The reason may be that w often appears in the harmful sentences used for the learning distributed representation expressions. In Internet slang, w is commonly used as a way of laughing at or making fun of other people, as !!!!!!! is often used with imperative sentences or invectives. Because many insulting expressions are derived from illness, insulting expressions and expressions related to injuries or diseases can be quite similar. This may explain why such expressions were incorrectly judged as harmful. In addition, instruction words can be quite similar to invectives or vilifications in their distributed representations; as a result, the system tended to wrongly judge them as harmful expressions.

TABLE 9
SUCCESS EXPRESSIONS AND FALSE EXPRESSIONS

List of Expressions			
Successful Expressions		Failure Expressions	Dictionary
Die	booby	Die	booby
Slip	hate	agots	hate
Drop	loneliness	Work	loneliness
Muggle	cow pattie	bugger	cow pattie
Idiot	NEET	Delete	NEET
Fall low	chaff	slip	chaff
Shut up	porky	crap	porky
Beast	Dying	Recede	Dying
Fall	Go to hell	Pass away	Go to hell
Crumple up	hiki-NEET		hiki-NEET
Annoying	Coy		
Ks	dating addiction		
Die	lyinch		
26 words		9 words	10 words

There are many expressions that might be perceived as impolite depending on ones viewpoint and may be judged harmful to some. If, for such expressions, the system were to display a caution to the user indicating that This expression may be offensive, the user might be motivated to modify the suspect sentence in advance. Insofar as much of the slang that appears on the Internet could be judged as possibly harmful, instructing users to avoid slang in order to prevent flaming might also be effective. [17, 18] proposed methods to extract and classify slang. Adapting such methods to extend our approach would seem to be a desirable next step.

TABLE 10
FALSE DETECTION EXPRESSIONS

Symbols	Injury/Illness	Instruction Word	Other
w	a vegetative state	Gangway	gnarly
www	comminuted fracture	Work	a lack in common sense
wwww	rest	Kick	
wwwww	sacrum	Flow	
w*25		Tincture	
! * 9			

VI. CONCLUSION

We first created a harmful expressions dictionary by collecting harmful expressions from social media and books.

We then collected a million tweets from Twitter and trained a model of a distributed representation vector by word2vec.

We expanded the harmful expressions dictionary by extracting words similar to the harmful expressions based on this model.

We also created a general expressions dictionary by collecting general and non-harmful expressions.

Our harmful expression classifier was trained by using the harmful expressions dictionary and the general expressions dictionary as training data for SVM learning.

A harmful expressions judgment system was created using this classifier. We then conducted an experiment using malicious sentences. It was found that the proposed system was able to correctly identify more harmful expressions than the baseline method in which only a dictionary was used.

In analyzing falsely detected expressions - general or non-harmful expressions that were identified as harmful - it was found that they were words that are likely to be mistakenly identified as harmful due to the influence of similar words or alternative common usage.

As a future task, we would like to expand the harmful expressions dictionary and propose a solution for declinable words or expressions that are split into multi-words by morphological analysis, as well as expressions that may be judged harmful in particular contexts.

ACKNOWLEDGMENT

This work was supported by JSPS KAKENHI Grant Number 15K00425, 15K00309, 15K16077.

REFERENCES

[1] M. Yamamoto and M. Hagiwara, “A moral judgment system using distributed representation and associative information,” *Transactions of Japan Society of Kansei Engineering*, vol. 15, no. 4, pp. 493–501, 2016.

[2] Y. Iwasaki, R. Orihara, Y. Sei, H. Nakagawa, Y. Tahara, and A. Ohsuga, “Identification of flaming and its applications in cgm,” in *Proceedings of the 6th International Conference on Agents and Artificial Intelligence-Volume 1*, Setúbal, Portugal, 2014.

[3] T. Ishisaka and K. Yamamoto, “Detecting nasty comments from bbs posts,” in *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, Sendai, Japan, 2010.

[4] 2 Channel, “Japanese text board,” n.d. [Online]. Available: https://goo.gl/T3U979

[5] A. H. Razavi, D. Inkpen, S. Uritsky, and S. Matwin, “Offensive language detection using multi-level classification,” in *Canadian Conference on Artificial Intelligence*, Berlin, Heidelberg, 2010.

[6] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detect-

ing offensive language in social media to protect adolescent online safety," in *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on Social Computing (SocialCom)*, Amsterdam, Netherlands, 2012.

[7] T. Nitta, F. Masui, M. Ptaszynski, Y. Kimura, R. Rzepka, and K. Araki, "Detecting cyberbullying entries on informal school websites based on category relevance maximization," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan, 2013.

[8] P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th annual meeting on association for computational linguistics*, Philadelphia, PA, 2002.

[9] S. Hatakeyama, F. Masui, M. Ptaszynski, and K. Yamamoto, "Statistical analysis of automatic seed word acquisition to improve harmful expression extraction in cyberbullying detection," *International Journal of Engineering and Technology Innovation*, vol. 6, no. 2, pp. 165–172, 2016.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, Lake Tahoe, NV, 2013.

[11] T. Kudo, "Yet another part-of-speech and morphological analyzer," 2005. [Online]. Available: https://goo.gl/AM5rmb

[12] V. Vapnik, "Pattern recognition using generalized portrait method," *Automation and Remote Control*, vol. 24, pp. 774–780, 1963.

[13] L. v. d. Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, no. 11, pp. 2579–2605, 2008.

[14] N. Amano, K. Kondo, S. Sakamoto, and Y. Suzuki, "NTT psycholinguistic databases Lexical Properties of Japanese, 1999," 2011. [Online]. Available: https://goo.gl/6Y94VV

[15] Cinnamon Official, "Cinnamon official Twitter account," n.d. [Online]. Available: https://goo.gl/7uDGq6

[16] Scikit Learn, "Machine learning in python," n.d. [Online]. Available: https://goo.gl/q1175

[17] K. Matsumoto, K. Akita, X. Keranmu, M. Yoshida, and K. Kita, "Extraction japanese slang from weblog data based on script type and stroke count," *Procedia Computer Science*, vol. 35, pp. 464–473, 2014.

[18] K. Matsumoto, S. Tsuchiya, M. Yoshida, and K. Kita, "Judgment of slang based on character feature and feature expression based on slangs context feature," in *International Conference on Soft Computing in Data Science*, Kuala Lumpur, Malaysia, 2016.