# Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization

**Taisei Nitta** † **Fumito Masui** † **Michal Ptaszynski** † **Yasutomo Kimura** §
**Rafal Rzepka** ‡ **Kenji Araki** ‡
† Department of Computer Science, Kitami Institute of Technology
nitta@ialab.cs.kitami-it.ac.jp, {f-masui,ptaszynski}@cs.kitami-it.ac.jp
§ Department of Information and Management Science, Otaru University of Commerce
kimura@res.otaru-uc.ac.jp
‡ Graduate School of Information Science and Technology, Hokkaido University
{kabura,araki}@media.eng.hokudai.ac.jp

## Abstract

We propose a novel method to detect cyberbullying entries on the Internet. "Cyberbullying" is defined as humiliating and slandering behavior towards other people through Internet services, such as BBS, Twitter or e-mails. In Japan members of Parent-Teacher Association (PTA) perform manual Web site monitoring called "net-patrol" to stop such activities. Unfortunately, reading through the whole Web manually is an uphill task. We propose a method of automatic detection of cyberbullying entries. In the proposed method we first use seed words from three categories to calculate semantic orientation score PMI-IR and then maximize the relevance of categories. In the experiment we checked the cases where the test data contains 50% (laboratory condition) and 12% (real world condition) of cyberbullying entries. In both cases the proposed method outperformed baseline settings.

## 1 Introduction

"Cyberbullying" is a new form of bullying. It is carried out on the Internet instead of classrooms. It takes form of hate messages sent through e-mails, electronic Bulletin Board System (BBS), etc., with the use of personal computers or mobile phones. Recently it has become a serious social problem in many countries, one of which is Japan (MEXT, 2008). Examples of cyberbullying that actually happened, include ridiculing student personality, body type, or appearance on informal school BBS, slandering students and insinuating they had performed deviate sexual intercourses. Some cases of cyberbullying lead the students who were bullied to assault or kill themselves or the student who wrote the bullying entry on the BBS.

To deal with the problem members of Parent-Teacher Association (PTA)[1] perform Web site monitoring activities called "net-patrol". When a harmful entry is detected the net-patrol member who found it sends a request to remove the entry to the Internet provider or Web site administrator. Some of the actual examples of harmful entries which were requested for deletion are represented in Table 1 (names, phone numbers and other personal information was changed).

Unfortunately, net-patrol has been carried out mostly manually. It takes much time and effort to find harmful entries (entries that contain harmful information and expressions) in a large amount of contents appearing on countless number of bulletin board pages. Moreover, the task comes with a great psychological burden on mental health to the net-patrol members. To solve the above problem and decrease the burden of net-patrol members, Matsuba et al. (2011) proposed a method to detect harmful entries automatically.

In their method they extended the method of relevance calculation PMI-IR, developed by Turney (2002) to calculate relevance of a document with harmful contents. With the use of a small number of seed words they were able to detect effectively large numbers of document candidates for harmful entries.

Their method was proved to determine harmful entries with an accuracy of 83% on test data for which about a half contained harmful entries. However, it was not yet verified how well the method would perform in real life conditions, where the ratio of cyberbullying entries and normal contents is not equal.

In this research, based on Matsuba et al.'s method of obtaining maximal relevance values for seed words, we propose a method for maximization of relevance score of seed words. In our

[1] An organization composed of parents, teachers and school personnel.



*International Joint Conference on Natural Language Processing*, pages 579–586,
Nagoya, Japan, 14-18 October 2013.

method we divide seed words into multiple categories and calculate maximal relevance value for each seed word with each category. By calculating the score, representing semantic orientation of "harmfulness", the method is expected to detect harmful entries more effectively than in the previous research. Moreover, we evaluate our method on data sets with different ratios of harmful contents to verify the usability of the method in the most realistic way.

The paper outline is as follows. Firstly, we describe research on extraction of harmful entries in Section 2. Next, we describe the proposed method in Section 3. Furthermore, in Section 4 we construct evaluation data sets with different ratios of cyberbullying entries, perform evaluation experiments based on these data sets, and describe the results of the experiments. We present a discussion and explain the results in detail in Section 5. Finally, in Section 6 we conclude the paper and propose some of the ideas for future improvement of the method.

## 2 Related research

There has been a number of research on extracting harmful information before. For example, Ishisaka and Yamamoto (2010) have focused on developing an abusive expression dictionary based on a large Japanese electronic bulletin board "2 channel". In their research Ishisaka and Yamamoto firstly defined words and paragraphs in which the speaker directly insults or slanders other people with the use of explicit words and phrases such as バカ (*baka*) "stupid", or マスゴミのクズ (*masugomi no kuzu*) "trash of mass-mudia". Next, they studied the use of abusive language, in particular which words appear the most often with abusive expressions, and based on this study they extracted abusive expressions from the surrounding context.

In other research Ikeda and Yanagihara (2010) have manually collected and divided separate sentences into harmful and non-harmful, and based on word occurrence within the corpus they created a list of keywords for classification of harmful contents. Next they utilized context of dependency structures of sentences containing harmful and non-harmful contents to improve the system performance. However, on the Web there are numerous variations of the same expressions differing with only one or two characters, such as 爆破 (*bakuha*) "blow up" and 爆ー破 (*baku–ha*) "blooow up". The weakness of this method is that all of the variations of the same expression need to be collected manually, which is very time-consuming.

Table 1: Examples of harmful entries which were requested for deletion. Japanese (above), transliteration (middle), English translation (below).

- 調子乗りすぎいっぺん殺らなあかんで
(*Chōshi nori sugi ippen yara na akan de*)
"Don't get excited that much or I'll kill ya!"
- 新田キモイつかキショイほんま死んで
(*Nitta kimoi tsuka kishoi honma shinde*)
"Nitta [proper noun], you're ugly,
or rather fugly, just die, man"
- ンな奴どつき回したれ
(*N'na yatsu dotsuki mawashi tare*)
"What an ass, slap him"
- 性格わるーい　ぶちゃいくー笑
(*Seikaku waru–i buchaiku– warai*)
"Baaad personality, and an ugly hag, lol"
- ＞＞17 あの女、昔、モバだったかグリに
登録してたヤリマンじゃん。
(*Ano onna mukashi Moba dattaka Guri ni tōroku shiteta yariman jan*)
"＞＞17 that woman is the same one
who was bitching around before
on *Mobage* or *Gree*[2]."
- すぐにヤレる。01234567890。
めっちゃカワイイで
(*Sugu ni yareru. 01234567890. Meccha kawaii de.*)
"You can take her out even now.
01234567890. She's a great lay."

Fujii et al. (2010) proposed a system for detecting documents containing excessive sexual descriptions using a distance between two words in a sentence. In their method they determine as harmful those words which are in closer distance to words appearing only in harmful context ("black words") rather than those in closer distance to words which appear in both harmful and non-harmful context ("grey words").

Hashimoto et al. (2010) proposed a method for detecting harmful meaning in jargon. In their method they assumed that the non-standard meaning is determined by the words surrounding the

[2] *Mobage* and *Gree* are online game service Web sites.

word in question. They detected the harmful meaning based on calculating co-occurrence of a word with its surrounding words.

In our research we did not consider the surrounding words. Instead, we determine the harmfulness of input by calculating the harmfulness score for all word sequences in input. Moreover, since we check the co-occurrence of word sequences on the Web, our method greatly reduces the cost of manual construction of training data. Furthermore, in calculating the harmfulness score we apply dependency relations between phrases. Therefore there is no need to check all words proceeding and succeeding the queried words, which greatly reduces processing time.

## 3 Proposed method

In this section we present an overview of the method for maximization of category relevance. In the proposed method we extend the method proposed by Turney (2002) to calculate the relevance of seed words with entries from the bulletin board pages. Moreover, we apply multiple categories of harmful words and calculate the degree of association separately for each category. Finally, as the harmfulness score (or polarity of "harmfulness") we choose the maximum value achieved by all categories. The method consists of three steps. (1) Phrase extraction, (2) Categorization and harmful word detection together with harmfulness polarity determination, (3) Relevance maximization. Each of the steps is explained in detail in the following paragraphs.

### 3.1 Phrase extraction

In cyberbullying entries the harmful character of an entry can be determined by looking at separate words. For other cases however, even if a word in itself is not harmful, it gains harmful meaning when used in a specific context, or in combination with other words. For example, for a pair of words 性格が悪い (*seikaku ga warui*) "bad personality", neither "bad", nor "personality" on their own express harmful meaning. However, when these words are used together in a dependency relation, they become harmful (negative depiction of a person's personality). Therefore, methods for detecting harmful contents using separate words only, will fail when they encounter an entry which gained harmful meaning by phrases containing words in dependency relation.

Table 2: Types of phrases applied in the proposed method with examples.

| Phrase | Example |
|---|---|
| **noun-noun** | サル顔<br>(*sarugao*)<br>"monkey face"<br>→Description ridiculing person's features |
| **noun-verb** | 新田を殺す<br>(*Nitta wo korosu*)<br>"Kill Nitta"<br>→Threatening expressions |
| **noun-adjective** | 性格が悪い<br>(*seikaku ga warui*)<br>"bad personality"<br>→Description criticizing person's features |

To solve this issue, we use the polarity calculation score for the morphemes[3] combined in the dependency relation. We define such a combination as a "phrase". One phrase consists of a morpheme pair in dependency relation. The dependency relation is calculated using a standard morphological analyzer for Japanese (MeCab[4]) and a Dependency parser for Japanese (Cabocha[5]). The phrases defined this way are extracted from all target entries.

### 3.2 Harmful word detection and categorization

In this process we detect words of potential harmful connotations, or "harmful words". Harmful words often include newly coined words or informal modifications of normal transcriptions, thus are not recognized by standard preprocessing tools, such as morphological analysers or dependency parsers. Therefore, it is possible such words, unless specifically annotated, would not be handled properly and cause error in morphological analysis. We investigated the entries of informal school Websites using the definition of harmful words proposed by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2008, later referred to as "Min-

[3] In this report, we use the words "word" and "morpheme" in the same meaning.

[4] http://mecab.sourceforge.net/

[5] http://code.google.com/p/cabocha/

istry of Education"), and registered 255 harmful words (nouns, adjectives and verbs) in the dictionary of morphological analyzer. In addition, we categorized harmful words into three categories: obscene, violent and abusive. The Ministry of Education defined words considered as cyberbullying to include obscene, violent, or abusive words used on BBS. In this study, we applied the above definition, and therefore we also classified harmful words into the three categories: obscene, violence, or abusive. Next, we selected from each category three most often occurring words as seed words and registered them in the system. The words we selected include セックス (*sekkusu*) "[to have] sex", ヤリマン (*yariman*) "slut", フェラ (*fera*) "fellatio" for the obscene category, 死ね (*shine*) "die [imperative]", 殺す (*korosu*) "[to] kill", 殴る (*naguru*) "[to] slap" for the violent category, and うざい (*uzai*) "annoying", きもい (*kimoi*) "gross", 不細工 (*busaiku*) "ugly" for the abusive category.

### 3.3 Maximization of relevance score

In this process we calculate harmfulness polarity score of phrases with each seed word for all three categories. We use pointwise mutual information (PMI) score as a measure of relevance between a phrase and harmfulness polarity words from each category. PMI here indicates a co-occurrence frequency of the queried phrase with the three words registered for each category. To calculate the co-occurrence frequency we use information retrieval (IR) score. Countless number of various pages exists on the Web, and thus various words are written there. Therefore, it is possible to obtain a high coverage by using the IR score.

We calculate the relevance of a phrase with words from each category according to the following equation (1). $p_i$ is a phrase extracted from the entry, $w_j$ are three words that are registered in one category of harmfulness polarity words, $hits(p_i)$ and $hits(w_j)$ are Web search hits for each category for $p_i$ and $w_j$ respectively, $hits(p_i \& w_j)$ is a number of hits when $p_i$ and $w_j$ appear on the same web page. Finally, $PMI - IR(p_i, w_j)$ is the relevance of $p_i$ and $w_j$.

$$PMI - IR(p_i, w_j) = log_2 \left\{ \frac{hits(p_i \& w_j)}{hits(p_i) hits(w_j)} \right\} \quad (1)$$

From all three scores calculated for the phrase with seed words from the three categories, we select the category which achieved the highest score as the one of the highest relevance with the phrase. We calculate the relevance score this way for all phrases extracted from the entry. Finally we select the category with the maximal overall score as the one with the highest relevance with the entry. The $score$ is calculated according to the following equation (2).

$$score = max(max(PMI - IR(p_i, w_j))) \quad (2)$$

In the baseline settings (Matsuba et al., 2011) the relevance was calculated as a sum of all scores for all phrases with each harmful word separately. In this method instead of taking all words separately we group them in categories and calculate the relevance with three most common harmful words from each category. By incorporating this improvement during the Web search the retrieved pages are those for which the phrase appeared not only with one of the harmful words, but with all three words from one category. This not only reduces the processing time, but also improves the calculation of the relevance score, since only the strongest (most harmful) phrases are selected. Moreover, since it is easier to find a Web page containing a phrase and the only one harmful word, than the phrase with three words together, calculating the relevance for all harmful words separately allowed phrases with low actual relevance to achieve high scores in the baseline system. Maximization of the relevance $score$ prevents low relevance phrases to erroneously achieve high scores.

We explain this process on the following example: 可愛いけど性格が悪い女 (*kawaii kedo seikaku ga warui onna*) "Cute girl, but bad personality". Firstly, from the entry the extracted phrases are: 可愛い-女 (*kawaii-onna*) "cute-girl", 性格-悪い (*seikaku-warui*) "bad-personality", 悪い-女 (*warui-onna*) "bad-girl". Next, we calculate the relevance between "cute-girl" and the three groups of words separately ("sex, slut, fellatio", "die, kill, slap", "annoying, gross, ugly"). The highest maximal score is selected as the relevance (harmfulness score) of this phrase. Similarly the score is calculated for "bad-personality" and "bad-girl". From all the scores for all phrases the highest overall score is considered as the maximized harmfulness $score$ of the phrase. All entries are then sorted beginning with the one with the highest harmfulness score. Finally, we set a

Table 3: The numbers of entries on informal school BBS including the number and percentage of cyberbullying entries.

| BBS | Overall number of entries | Cyberbullying entries | Percentage (%) |
|---|---|---|---|
| BBS(1) | 600 | 75 | 12.5 |
| BBS(2) | 736 | 90 | 12.2 |
| BBS(3) | 886 | 100 | 11.3 |

harmfulness threshold $n$ and consider $n$ entries with the highest score as harmful, and discard other as irrelevant to check how many of the entries within the specified threshold are in fact harmful.

## 4 Evaluation experiment

We performed and experiment to evaluate the performance of the proposed method, and compared the results with the baseline. Below we describe the preliminary study for carrying out the experiment (Section 4.1), explain the experiment settings (Section 4.2) and report the results of experiments (Section 4.3).

### 4.1 Preliminary study

It is necessary to create a test data with harmful and non-harmful entries mixed at an appropriate rate. In previous research the mixing was set at the same rate (half of the entries included cyberbulliyng). However, it cannot be assumed that harmful entries appear in real life with the same rate as normal ones. Therefore to evaluate our method in conditions closer to reality we performed a preliminary study to verify how much of the entries are harmful on actual Web pages. We counted the harmful entries mixing ratio on three informal school Websites, in particular we focused on informal school bulletin boards (BBS). The result of the study is represented in Table 3. We performed the study during four days between January 27 and 30, 2012. The number of obtained entries was 2,222.

As of the result of the study, the first BBS contained a total number of 600 entries from which 75 were harmful, which indicates that harmful entry appearance rate was 12.5%. Similarly, for the second BBS, 90 out of 736 total entries were harmful (12.2%). On the third BBS there were 886 total entries with 100 harmful ones (11.3%). From the above results, we concluded that about 12% of all entries appearing on informal school BBS can be accounted as cyberbullying. Therefore in the experiment we verified the performance of the method under the condition when harmful entries cover 12% of the whole data.

### 4.2 Experiment settings

In the evaluation experiment we compared the performance of the proposed method to the baseline. We did this firstly for the case where the test data contained 50% of harmful entries. Next, we prepared a different test data, which contained 12% of harmful entries and compared the performance under this condition for both the baseline and the proposed method.

The test data containing 50% of harmful entries contains 2,998 entries in total with 1,508 of harmful entries and 1,490 of non-harmful entries. The dataset contains actual collection gathered by the net-patrol members from bulletin boards, and additional data gathered manually by Matsuba et al. (2011) (the latter were collected from the BBS sites limited to schools from the Mie Prefecture, Japan). We performed a 10-fold cross validation on this dataset. We processed the dataset by both the baseline and the proposed method and calculated the harmful polarity score for entries where the phrases could be extracted. Then we ranked all entries decreasingly according to the harmful polarity score, and evaluated the performance looking at the top $n$ entries by increasing the threshold of $n$ by 50 each time.

To prepare the dataset for the real world condition (12% of harmful entries), we prepared five test sets by randomly extracting 60 harmful and 440 non-harmful, 500 entries in total from the original dataset. On these datasets we did not perform a 10-fold cross validation, since it would make the results not statistically relevant (each set for cross validation would contain only 60 harmful entries). Instead we calculated the results for each of the five sets separately. This allowed us to include all entries from the original test set in the evaluation. We processed these datasets with both the baseline and the proposed method and calculated the harmfulness polarity score for entries for which the phrases could be extracted. Then, similarly to the original dataset, we ranked all entries based on the harmfulness polarity score, and evaluated the performance by taking the top $n$ and increasing the threshold $n$ by 10 each time.

We considered automatic setting of the threshold using machine learning methods, however it was difficult due to the small size of test data for the real world condition. In the future for automatic threshold setting we plan to develop a machine learning method capable of handling small sized data. Therefore this time we increased the threshold manually each time and investigated Precision and Recall for each threshold.

As evaluation criteria we used Precision (P) and Recall (R), calculated according to the equations (3) and (4). The Precision is a ratio of the number of entries that could be properly determined as harmful to the number of all entries determined as harmful among the top $n$. The Recall is a ratio of the number of entries that could be properly determined to be harmful to the overall number of harmful entries. The final performance is calculated as an average of Recall and Precision for each test data in this experiment.

$$P = \frac{correct\ \ annotations}{all\ \ system\ \ annotations} \tag{3}$$

$$R = \frac{correct\ \ annotations}{all\ \ harmful\ \ annotations} \tag{4}$$

### 4.3 Results

The results showing Precision and Recall for both the baseline and the proposed method for both datasets (50% and 12% of cyberbullying entries) are represented in Figure 1. The horizontal axis and the vertical axis represent percentage of Recall and Precision for each threshold, respectively.

For the test data containing 50% of harmful entries, Precision was between 49% - 79% for the baseline, while for the proposed method Precision was between 49% - 88%.

For the test data containing 12% of harmful entries, Precision was between 11% - 31% for the baseline, while for the proposed method Precision was between 10% - 61%.

## 5 Discussion

The experiment results showed that the proposed method achieved higher overall performance comparing to the baseline. The shape of the correlation curve for Recall and Precision shows that that performance for the baseline is significantly reduced in general comparing the test data containing 12% of harmful entries to the test data containing 50% of harmful entries. However, for the proposed method, although the performance is reduced as well, there is no sudden drop in the shape of the correlation curve when comparing both datasets.

This could suggest that the performance is more stable in the proposed method than in the baseline. There were several cases of threshold $n$ where the Recall was slightly higher for the baseline than for the proposed method in the test data containing 12% of harmful entries. This happened because for some harmful entries the harmfulness score could not be calculated highly enough due to the fact that the score calculation is more strict (Precision-oriented) in the proposed method. Therefore, although Recall is slightly higher in the baseline for large thresholds, the Recall is higher for the proposed method for small and medium thresholds, with the Precision being constantly higher for the proposed method. Therefore, it can be said that the proposed method achieved higher general performance than the baseline.

Next we explain the results for the test data containing 12% of harmful entries. We investigated the threshold cases of entries where the Precision reaches 48%. Entries found there included, for example “アトピーのやつ死ねよ” (*atopii no yatsu shine yo*) “The bastard with atopy must die” and “ウザイキモイぶす” (*uzai kimoi busu*) “annoying, gross and ugly”. From those entries high relevance score was calculated for phrases like “アトピー-死ね” (*atopii-shine*) “atopy-die” and “ウザイ-ぶす” (*uzai-busu*) “annoying-ugly”. Since the phrases included seed words as well, this most probably increased the polarity value of the harmful entry.

On the other hand, there were many non-harmful entries classified as harmful with harmfulness polarity score equally high or higher than the actual harmful entries. An example of a non-harmful entry of this kind was “県外に住んでいる” (*kengai ni sun de iru*) “living outside of the prefecture”. The phrase extracted from this entry was “外-住ん” (*soto-sun*) “outside-live”. This is a neutral phrase, and appears similarly often in non-harmful entries as well as in harmful entries in cases of exposing personal information about where a person lives. Therefore, the relevance of a harmful entry containing such a phrase is increased, and as a result overall harmfulness po-

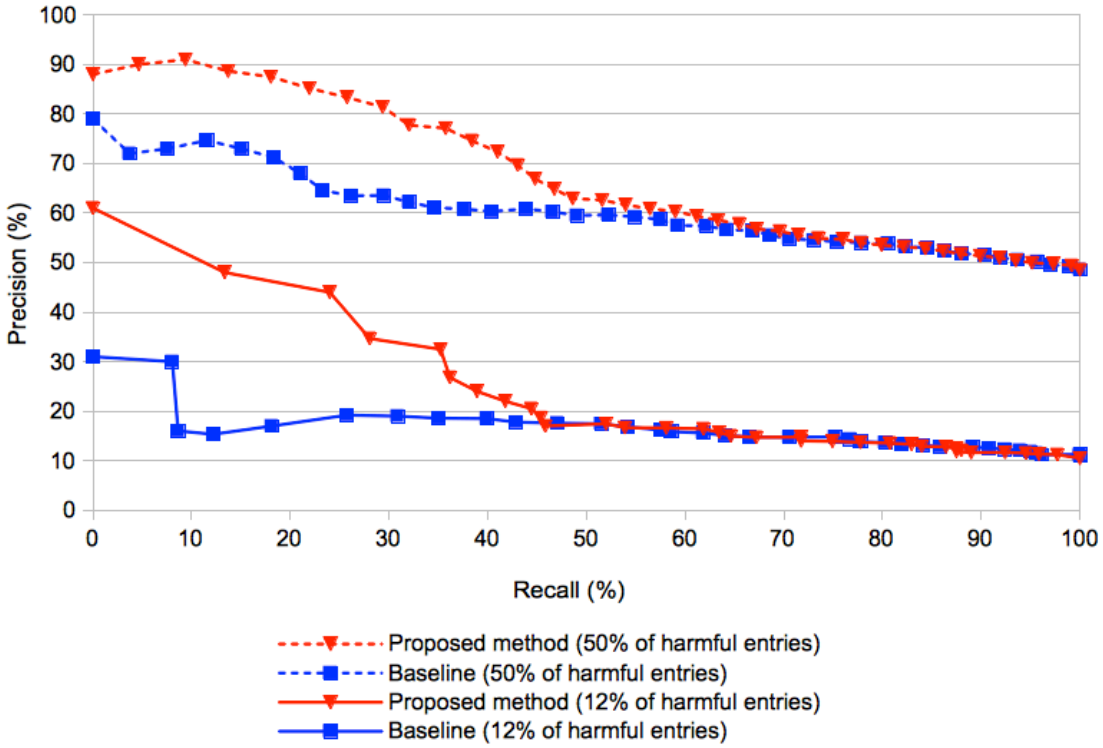


Figure 1: Precision and Recall for both the baseline and the proposed method for both datasets (50% and 12% of cyberbullying entries). The horizontal axis and the vertical axis represent percentage of Recall and Precision for each threshold, respectively. All results statistically significant. For the 50% data the results were extremely statistically significant on 0.0001 level. For the 12% data the results were statistically significant on 0.05 level.

larity score of such non-harmful entries becomes higher.

As a countermeasure it could be considered to register some words, which appear only in non-harmful entries, like "splendid", with non-harmful polarity and calculate the relevance of non-harmful entries as well. In particular, firstly, the non-harmful words could be registered in the dictionary. Then the relevance could be calculated for the phrases with both, the non-harmful polarity words and harmful polarity words. In such cases the phrase could be considered as non-harmful when the relevance score of the phrase with non-harmful words was higher than the relevance with the harmful words. This could reduce the influence of neutral phrases on the overall performance.

We also investigated the cases where Recall reaches 100%. These cases include entries which contain personal information only such as school names or person's names, such as "Nitta of Kitami Institute of Technology, 4th grade", etc. The relevance of such entries with harmful words registered at present in the dictionary is low, which influenced their overall harmfulness score as well. To solve this problem we plan to register in the dictionary words which have high relevance with personal information and use them in the relevance score calculation as well.

## 6 Conclusions and future work

In this study, we proposed a method of maximization of category relevance to automatically detect cyberbullying entries on the Internet. With this research we wish to contribute to reducing the burden of Internet patrol personnel who make efforts to manually detect harmful entries appearing on the Internet. In order to verify the actual usefulness of the proposed method we evaluated the performance for the test data containing similar percentage of harmful entries as in reality. Firstly, in a preliminary study we verified the usual ratio of harmful entries on the Internet. Next, we prepared test datasets containing the same amount of cyberbullying entries as in reality and evaluated the method on these test sets. In addition, we re-

produced the baseline method and compared the performance to the proposed method.

The experiment results showed that the proposed method obtained higher results than the baseline. Under the fair condition (test dataset with 50% of harmful entries) the proposed method achieved over 90% of Precision at 10% Recall and keeping up high Precision (80-70%) at Recall close to 50%. Under the real world condition (test dataset with 12% of harmful entries) the method achieved nearly 50% of Precision at about 10% of Recall. The relevance curve have decreased slowly with growing Recall for the proposed method, while for the baseline the relevance curve has dropped suddenly from 30% to around 15% at the same Recall rate. As for drawbacks in our method, harmful entries consisting of personal information were scored as less harmful due to the appearance of neutral phrases which appear often in both harmful and non-harmful entries.

In the near future, we plan to register non-harmful polarity words which have a high relevance with non-harmful entries to lower the overall harmfulness polarity score of non-harmful entries containing neutral phrases. We will also investigate a method for assessing the harmfulness score to entries including personal information. Furthermore, we plan to increase the data set, and determine the optimal threshold automatically by using machine learning.

## Acknowledgement

This work was supported by JSPS KAKENHI (Grants-in-Aid for Scientific Research) Grant Number 24600001.

## References

Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2008. *'Netto-jō no ijime' ni kansuru taiō manyuaru jirei shū (gakkō, kyōin muke)* ["Bullying on the Net" Manual for handling and collection of cases (for schools and teachers)] (in Japanese). Published by MEXT.

Tatsuaki Matsuba, Fumito Masui, Atsuo Kawai, Naoki Isu. 2001. *Gakkō hi-kōshiki saito ni okeru yūgai jōhō kenshutsu wo mokuteki to shita kyokusei hantei moderu ni kansuru kenkyū* [Study on the polarity classification model for the purpose of detecting harmful information on informal school sites] (in Japanese), In *Proceedings of The Seventeenth Annual Meeting of The Association for Natural Language Processing (NLP2011)*, pp. 388-391.

Peter D. Turney. 2002. Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews, In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 417-424.

Tatsuya Ishisaka, Kazuhide Yamamoto. 2010. *2chaeru wo taishō to shita waruguchi hyōgen no chūshutsu* [Extraction of abusive expressions from 2channel] (in Japanese), In *Proceedings of The Sixteenth Annual Meeting of The Association for Natural Language Processing (NLP2010)*, pp.178-181.

Kazushi Ikeda, Tadashi Yanagihara. 2010. *Kakuyōso no chūshōka ni motozuku ihō-, yūgai-bunsho kenshutsu shuhō no teian to hyōka* [Proposal and evaluation of method for illegal and harmful document detection based on the abstraction of case elements] (in Japanese), In *Proceedings of 72nd National Convention of Information Processing Society of Japan (IPSJ72)*, pp.71-72.

Yutaro Fujii, Satoshi Ando, Takayuki Ito. 2010. *Yūgai jōhō firutaringu no tame no 2-tango-kan no kyori oyobi kyōki jōhō ni yoru bunshō bunrui shuhō no teian* [Developing a method based on 2-word cooccurence information for filter harmful information] (in Japanese), In *Proceedings of The 24th Annual Conference of The Japanese Society for Artificial Intelligence (JSAI2010)*, paper ID: 3D2-4, pp. 1-4.

Hiromi Hashimoto, Takanori Kinoshita, Minoru Harada. 2010. *Firutaringu no tame no ingo no yūgai goi kenshutsu kinō no imi kaiseki shisutemu SAGE e no kumikomi* [The function that detect harmful word sense from slang built into the semantic analysis system SAGE for filtering] (in Japanese), IPSJ SIG Notes 2010-SLP-81(14), pp. 1-6.