Brute Force Works Best Against Bullying

Michal Ptaszynski **Fumito Masui**
Department of Computer Science,
Kitami Institute of Technology
{ptaszynski,f-masui}@cs.kitami-it.ac.jp

Yasutomo Kimura
Department of Information and Management
Science, Otaru University of Commerce
kimura@res.otaru-uc.ac.jp

Rafal Rzepka **Kenji Araki**
Graduate School of Information Science and Technology, Hokkaido University
{kabura,araki}@media.eng.hokudai.ac.jp

1 Introduction

The problem of harmful and offending messages on the Internet has existed for many years. One of the reasons such activities evolved was the anonymity of communication on the Internet, giving users the feeling that anything can go unpunished. Recently the problem has been officially defined and labeled as cyberbullying (CB)[1].

In Japan the problem has become serious enough to be noticed by the Ministry of Education [MEXT 2008]. In 2007 Japanese school personnel and members of Parent-Teacher Association (PTA) have started monitoring activities under the general name Internet Patrol (later: *net-patrol*) to spot Web sites containing such inappropriate contents. However, the net-patrol is performed manually as a volunteer work. Countless amounts of data on the Internet make this an uphill task. This situation motivated us to help and ease the burden of the net-patrol members and create a net-patrol crawler automatically spotting cyberbullying entries on the Web and reporting them to appropriate organs.

In the following sections we firstly present some of the previous research related to ours. Next, we describe our method and the dataset used in this research. In the method we apply a combinatorial approach to language modeling, resembling brute force search algorithms, to extract sophisticated patterns from sentences. Next, we use them in text classification task. Finally, we explain the evaluation settings, analyze and discuss the results. Evaluation on actual cyberbullying data showed our method outperformed previous ones while minimizing human effort.

2 Previous Research

[Ptaszynski et al. 2010] performed affect analysis of small dataset of cyberbullying entries to find out that their distinctive features were vulgar words. They applied a lexicon of such words to train an SVM classifier. With a number of optimizations the system was able to detect cyberbullying with 88.2% of F-score. However, increasing the data caused a decrease in results, which made them conclude SVMs are not ideal in dealing with frequent language ambiguities typical for cyberbullying. Next, [Matsuba et al.2011] proposed a method to automatically detect harmful entries by extending the SO-PMI-IR score to calculate relevance of a document with harmful contents. With the use of a small number of seed words they were able to detect large numbers of candidates for harmful documents with an accuracy of 83%. Finally, [Nitta et al. 2013] proposed an improvement to Matsuba et al.'s method. They calculated SO-PMI-IR score for three categories of seed words (abusive, violent, obscene), and selected the one with the highest relevance. Their method achieved 90% of Precision for 10% Recall.

Most of the previous research assumed that using vulgar words as seeds will help detect cyberbullying. However, all of them notice that vulgar words are only one kind of distinctive vocabulary and do not cover all cases. We assumed such a vocabulary can be extracted automatically. Moreover, we did not restrict the scope to words, but extended the search to sophisticated patterns with disjoint elements. To achieve this we applied a pattern extraction method based on the idea of brute force search algorithm.

3 Method Description

We assumed that applying sophisticated patterns with disjoint elements should provide deeper insight than the usual bag-of-words or n-gram approach. Such patterns, if defined as ordered combinations of sentence elements, could be extracted automatically. Algorithms using combinatorial approach usually generate a massive number of combinations - potential answers to a given problem. Thus they are often called brute-force search algorithms. We assumed that optimizing the combinatorial algorithm to the problem requirements should make it advantageous in language processing task.

In the proposed method, firstly, ordered non-repeated combinations are generated from all elements of a sentence. In every n-element sentence there is k-number of combination clusters, such as that $1 \leq k \leq n$, where k represents all k-element combinations being a subset of n. In this procedure all combinations for all values of k are generated. The number of all combinations is equal to the sum of all k-element combination clusters (see eq. 1).

$$\sum_{k=1}^{n} \binom{n}{k} = \frac{n!}{1!(n-1)!} + \frac{n!}{2!(n-2)!} + ... + \frac{n!}{n!(n-n)!} = 2^n - 1 \quad (1)$$

Next, all non-subsequent elements are separated with an asterisk ("*"). Pattern occurrences O for each side of the dataset is used to calculate their normalized weight w_j (eq. 2). The score of a sentence is calculated as a sum of weights of patterns found in the sentence (eq. 3).

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5\right) * 2 \quad (2) \qquad score = \sum w_j, (1 \geq w_j \geq -1) \quad (3)$$

[1] http://www.ncpc.org/cyberbullying

The weight can be further modified by:

- awarding pattern length k (LA),
- awarding length and occurrence O (LO).

The list of frequent patterns can be also further modified by:

- discarding ambiguous patterns which appear in the same number on both sides (harmful and non-harmful); later “zero patterns” (0P), as their weight is equal 0.
- discarding ambiguous patterns of any ratio on both sides

We also compared the performance of sophisticated patterns (PAT) to more common n-grams (NGR).

4 Evaluation Experiment

Experiment Setup

In the evaluation we used a dataset created by [Matsuba et al.2011]. The dataset was also used by [Ptaszynski et al. 2010] and recently by [Nitta et al. 2013]. It contains 1,490 harmful and 1,508 non-harmful entries collected from unofficial school Web sites and manually labeled by Internet Patrol members according to instructions included in the manual for dealing with cyberbullying [MEXT 2008].

The dataset was further preprocessed in three ways:

- **Tokenization:** All words, punctuation marks, etc. are separated by spaces (TOK).
- **Parts of speech (POS):** Words are replaced with their representative parts of speech (POS).
- **Tokens with POS:** Both words and POS information is included in one element (POS+TOK).

We compared the performance for each kind of dataset preprocessing using a 10-fold cross validation and calculated the results using standard Precision, Recall and balanced F-score. There were several evaluation criteria. We checked which version of the algorithm achieves top scores within the threshold span. We also looked at break-even points (BEP) of Precision and Recall and checked the statistical significance of the results. We also compared the performance to the baselines [Matsuba et al.2011; Nitta et al. 2013].

Results and Discussion

Although highest occasional precision (P=.93) was achieved by POS feature set based on ngrams (NGR), its Recall and F-score were the lowest (R=.02, F=.78). Also high P with much higher R (P=.89, R=.34) was achieved by tokens with parts of speech based on either patterns or ngrams (TOK+POS/PAT|NGR). This feature set also achieved the highest general F-score (F=.8). Tokenization with POS tagging also achieved the highest break-even point (BEP) (P=.79, R=.79). In most cases deleting ambiguous patterns yielded worse results, which suggests that such patterns, despite being ambiguous (appearing in both cyberbullying and non-cyberbullying entries), are in fact useful in practice.

Comparison with Previous Methods

In the comparison with previous methods we used the ones by [Matsuba et al.2011], and [Nitta et al. 2013]. Moreover, since the latter extracts cyberbullying relevance values from the Web, we also repeated their experiment to find out how the performance of the Web-based method changed during the two years since being proposed. Also, to make the comparison fair, we used our best and worst settings. As the evaluation metrics we used area under the curve (AUC) of Precision and Recall (Fig. 1). The highest overall results were obtained by the best settings of the proposed method (TOK+POS/PAT). Although the highest score was still by [Nitta et al. 2013], their performance quickly decreases due to quick drop in Precision for higher thresholds. Moreover when we repeated their experiment in January 2015, the results greatly dropped. This could happed due to: (1) fluctuation in page rankings which pushed the information lower making it not extractable anymore; (2) frequent deletion requests of harmful contents by Internet Patrol members; (3) tightening of usage and privacy policies by most Web service providers. This advocates more focus on corpus-based methods such as the one proposed in this paper.

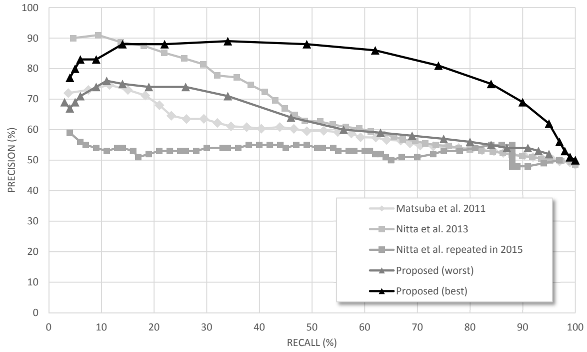

Figure 1: Comparison between the proposed method (best and worst performance) and previous methods.

5 Conclusions

In this paper we proposed a method for automatic detection of cyberbullying – a recently noticed social problem influencing mental health of Internet users.

We applied a combinatorial algorithm in automatic extraction of sentence patterns, and used those patterns in text classification of CB entries. The evaluation experiment performed on actual CB data showed our method outperformed previous methods.

References

[Matsuba et al.2011] T. Matsuba, F. Masui, A. Kawai, N. Isu. 2011. A study on the polarity classification model for the purpose of detecting harmful information on informal school sites (in Japanese), In *Proceedings of NLP2011*, pp. 388-391.

[MEXT 2008] Ministry of Education, Culture, Sports, Science and Technology (MEXT). 2008. “Bullying on the Net” Manual for handling and collection of cases (for schools and teachers) (in Japanese). Published by MEXT.

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

[Ptaszynski et al. 2010] M. Ptaszynski, P. Dybala, T. Matsuba, F. Masui, R. Rzepka, K. Araki, Y. Momouchi. 2010. In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis. *IJCLR*, Vol. 1, Issue 3, pp. 135-154.