Proceedings of the Linguistic And Cognitive Approaches To Dialog Agents Symposium, Rafal Rzepka (Ed.), at the AISB 2010 convention, 29 March – 1 April 2010, De Montfort University, Leicester, UK

Machine Learning and Affect Analysis Against Cyber-Bullying

Michal Ptaszynski[1] **Pawel Dybala**[1] **Tatsuaki Matsuba**[2] **Fumito Masui** [3] **Rafal Rzepka**[1] **Kenji Araki**[1]

Abstract. Online security has been an important issue for several years. One of the burning online security problems lately in Japan has been online slandering and bullying, which appear on unofficial Web sites. The problem has been becoming especially urgent on unofficial Web sites of Japanese schools. School personnel and members of Parent-Teacher Association (PTA) have started Online Patrol to spot Web sites and blogs containing such inappropriate contents. However, countless number of such data makes the job an uphill task. This paper presents a research aiming to develop a systematic approach to Online Patrol by automatically spotting suspicious entries and reporting them to PTA members and therefore help them do their job. We present some of the first results of analysis of the inappropriate data collected from unofficial school Web sites. The analysis is performed firstly with an SVM based machine learning method to detect the inappropriate entries. After analysis of the results we perform another analysis of the data, using an affect analysis system to find out how the machine learning model could be improved.

1 INTRODUCTION

Online security has been an urgent problem ever since the creation of the Internet. Some of the most well known issues of online security include hacking, cracking, data theft and online espionage. However, for several years, a problem that has become much more visible and therefore influential and socially harmful, is the problem of exploitation of online open communication means, such as BBS forum boards, or social networks to convey harmful and disturbing information. In USA, a great focus on this issue began in 2001 after the 9.11 terrorist attack. However, similar cases have been noticed before in other countries not on such a scale. In Japan, on which this research is focused, a great social disturbance was caused by cases of sending alarming messages by criminals or hijackers on the Internet just before a committing a crime. One famous case of this kind happened in May 2000, a year before the 9.11 in USA, when a frustrated young man sent an message on, then fairly popular, Japanese BBS forum *2channel*, informing readers he was going to hijack a bus, just before he proceeded with his plan. Growing number of similar cases around the world opened a public debate on whether such suspicious messages could not be spotted early enough to prevent the crimes from happening [1] and on the freedom of speech on the Internet in general [2]. Some of the famous research in this matter was performed by the team of Hsinchun Chen, who started the project aiming to analyze Dark Web Forums in search of alarming entries about planned terrorist attacks [3, 4]. Another research of this kind was performed by Gerstenfeld [5], who focused on extremist groups.

However, there have been little research performed on a problem less lethal, although equally serious, namely online slandering and bullying of natural persons, known generally as 'cyber-bullying'. In Japan the problem has become serious enough to be noticed by the Ministry of Education, Culture, Sports, Science and Technology (later: MEXT) [6]. At present school personnel and members of Parent-Teacher Association (PTA) have started Online Patrol to spot Web sites and blogs containing such inappropriate contents. However, countless number of such data makes the job an uphill task. Moreover, the Online Patrol is performed manually and as a volunteer work.

Therefore we started this research to help the Online Patrol members. The final goal of this research is to create a machine Online Patrol crawler automatically spotting the cyber-bullying cases on the Web and reporting them to the Police.

In this paper we present some of the first results of this research. We first focused on developing a systematic approach to spotting online cyber-bullying entries automatically to ease the burden of the Online Patrol volunteers. In our approach we use machine learning to train a system for spotting the undesirable contents and then perform affect analysis of these contents to find out how the system could be further improved.

The paper outline is as follows. In Section 2 we describe the problem of cyber-bullying in more details. In Section 3 we present the prototype SVM-based method for detecting the cyber-bullying entries and evaluate it. In Section 5 we present the results of affect analysis of these data and propose some ideas about how the system described in Section 3 could be improved. Finally, in Section 5 we conclude the paper and provide some hints on further work in this area.

2 WHAT IS CYBER-BULLYING?

Although the problem of sending harmful messages in the Internet has appeared for several yaers, it has been officially defined only lately and named as cyber-bullying[4]. The National Crime Prevention Council in America state that cyber-bullying happens "when the Internet, cell phones or other devices are used to send or post text or images intended to hurt or embarrass another person." [5] Other definitions, such as the one by Bill Belsey, a teacher and an anti-cyber-bullying activist, say that cyber-bullying "involves the use of

[1] Graduate School of Information Science and Technology, Hokkaido University, {ptaszynski,paweldybala,kabura,araki}@media.eng.hokudai.ac.jp
[2] Graduate School of Engineering, Mie University, matsuba@ai.info.mie-u.ac.jp
[3] Department of Computer Science, Kitami Institute of Technology, f-masui@mail.kitami-it.ac.jp

[4] other terms used are cyber-harassment, and cyber-stalking
[5] http://www.ncpc.org/cyberbullying

information and communication technologies to support deliberate, repeated, and hostile behavior by an individual or group, that is intended to harm others." [7]

Some of the first robust research on cyber-bullying was done by Hinduja and Patchin, who performed numerous surveys about the subject in the USA [8, 9]. They found out that the harmful information may include threats, sexual remarks, pejorative labels, false statements aimed at humiliation. When posted on a BBS forum or a social network, such as Facebook, it may disclose personal data of the victim which, published with a humiliating material about them or in their name, aims to defame or ridicule them personally.

2.1 Cyber-bullying and Online Patrol in Japan

In Japan, after a several cases of suicides of cyber-bullying victims who could not bare the humiliation, MEXT has considered the problem serious enough to start a movement against the problem. In a manual for spotting and handling the cases of cyber-bullying [6], the Ministry puts a great importance on early spotting of the suspicious entries and messages, and distinguishes several types of cyber-bullying noticed in Japan. These are:

1. Cyber-bullying appearing on BBS forums, blogs and on private profile web-sites;
   (a) Entries containing libelous, slanderous or abusive contents;
   (b) Disclosing personal data of natural persons without their authorization;
   (c) Entries and humiliating online activities performed in the name of another person;
2. Cyber-bullying appearing in electronic mail;
   (a) E-mails directed to a certain person/child, containing libelous, slanderous or abusive contents;
   (b) E-mails in the form of chain letters containing libelous, slanderous or abusive contents;
   (c) E-mails send in the name of another person, containing humiliating contents;

In this research we focused mostly on the cases of cyber-bullying appearing on informal web sites of Japanese secondary schools. Informal school web sites are web sites where school pupils gather to exchange information about school subjects or contents of tests, etc. However, as was noticed by Watanabe and Sunayama [10], on such pages there have been a rapid increase of entries containing insulting or slandering information about other pupils or even the teachers. Cases like that make other users uncomfortable using the Web sites and cause undesirable misunderstandings.

To deal with such malicious entries, a movement of Online Patrol have started. In this movement usually participate teachers and PTA members, who, based on the MEXT definition of cyber-bullying, read through all available entries, decide whether an entry is dangerous or not and, if necessary, send a deletion request to the web page administrator or the Internet provider and send a report to the police. The typical Online Patrol route is presented on Figure 1

Unfortunately, at present state of affairs, the school personnel and PTA members taking part in Online Patrol perform all tasks manually as voluntary work, beginning from reading the countless numbers of entries and deciding about the appropriateness of their contents,

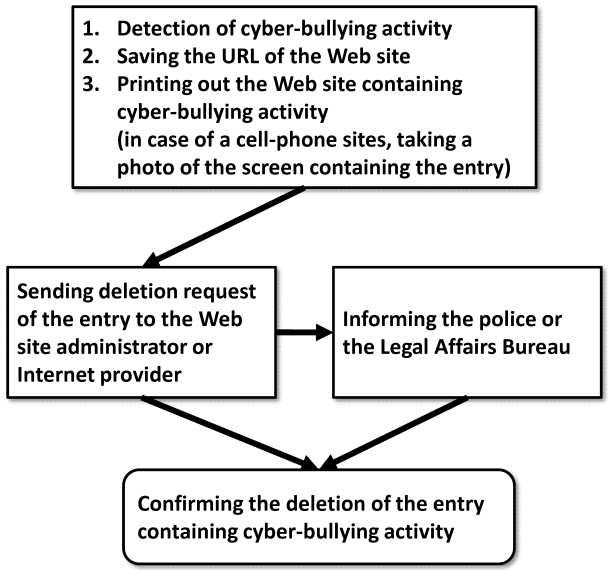

Figure 1. Route of Online Patrol.

through printing out or taking photos of the pages containing cyber-bullying entries, and finally sending the reports and deletion requests to the appropriate organs. Moreover, since the number of entries rises by the day, surveillance of all of the Web becomes an uphill task for the small number of patrol members.

With this research we aim to create a Web crawler capable to perform this difficult task instead of humans, or at least to ease the burden of the Online Patrol volunteers.

3 MACHINE LEARNING METHOD FOR CYBER-BULLYING DETECTION

In this section we describe a machine learning method developed to handle cyber-bullying activities. The method consists of several stages, including creation of a lexicon of vulgar, slanderous and abusive words, slanderous information detection module, ranking of the information according to the level of their harmfulness, and visualization of the harmful information. The system flow chart is represented on Figure 2. The creation of the system has separated into two general phases: training phase and processing (test) phase. Below we present the details of each phase.

1. Training phase;
   (a) Crawling the school Web sites;
   (b) Detecting manually the cyber-bullying entries;
   (c) Extraction of vulgar words and adding them to the lexicon;
   (d) Estimating word similarity with Levenshtein distance;
   (e) Part of speech analysis;
   (f) Training with SVM;
2. Processing (test) phase;
   (a) Crawling the school Web sites;
   (b) Detecting the cyber-bullying entries with SVM model;

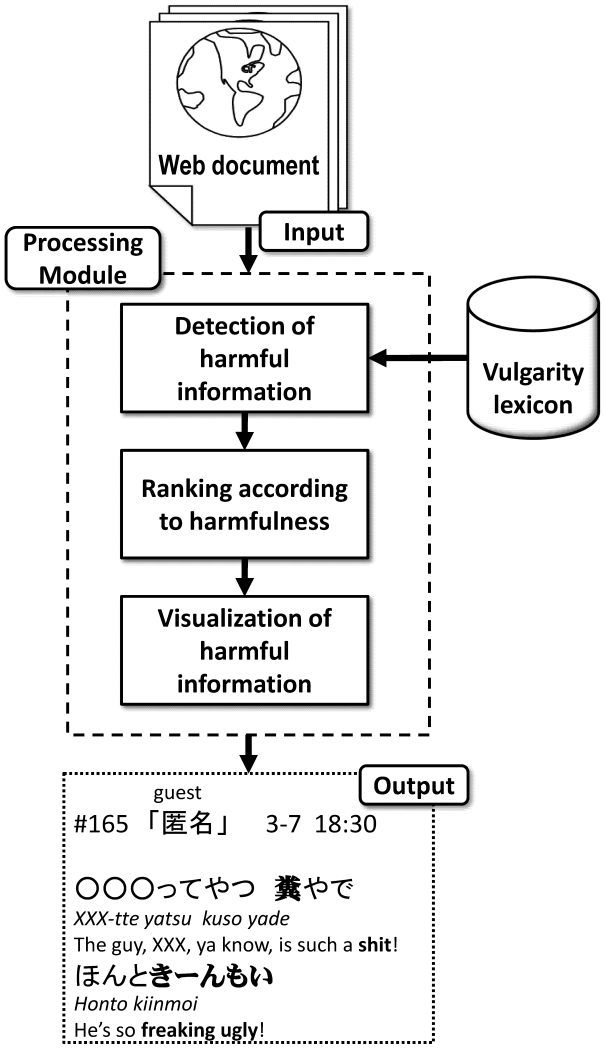

Figure 2. Flow chart of the cyber-bullying activity detecting system.

(c) Part of speech analysis of the detected harmful entry;

(d) Estimating word similarity with Levenshtein distance;

(e) Marking and visualization of the key sentence;

3.1 Construction of Vulgarity Lexicon

We defined the keywords distinguishable for cyber-bullying as vulgarities. Vulgarities are defined as "obscene or vulgar words which connote offenses against particular persons or society." Examples of such words in English are *shit*, *fuck*, *bitch*. Examples in Japanese include, e.g., *uzai* (freaking annoying), or *kimoi* (freaking ugly).

Since vulgar words are often not detected by part of speech (POS) taggers, or marked as "unknown word", we decided to create list of vulgarities and add it to a basic lexicon of POS tagger. This was performed according to the procedure described below. At first we performed a study on the vulgar keywords constituting the harmful information. We obtained a set of informal school Web pages, read them and performed a manual categorization into "harmful" and "non-harmful" entries based on the MEXT classification of cyber-bullying. From these Web pages we obtained 750 harmful entries including 216 unique vulgar keywords distinguishable for cyber-bullying activity. The extracted keywords were finally added to the list of vulgarities. Finally, we added the grammatical information to the extracted keywords and added them to the POS tagger. An example of addition of grammatical information is presented in Table 1.

Table 1. Example of a newly registered word.

kimoi (freaking ugly)
POS: Adjective;
Headword: *kimoi* (hit-rate: 294);
Reading: kimoi;
Pronunciation: kimoi;
Conjugated form: uninflected;

3.2 Estimation of Word Similarity with Levenshtein Distance

Users would often change spelling of words and write them in a unnormalized way. It is a part of jargonization of the language used online. Examples of this phenomenon in English, would be writing phrases like, e.g., "C.U." in the meaning of "See you [later]!" (usually on the end of e-mails or online messages), or "brah" in the meaning of "bro[ther], friend" [6], etc. Some examples of such colloquial transformations in the Japanese online language are shown in Table 2.

Table 2. Three examples of colloquial transformations of Japanese words in online jargon.

original word	colloquial transformation
kimoi (freaking ugly, gross)	*kimosu*, *kishoi*, *kisho*, ...
uzai (freaking annoying)	*uzee*, *UZAI*, *uzakkoi*, ...
busaiku (ugly bitch)	*buchaiku*, *bussaiku*, ...

With this variation, words having the same meaning would be classified as separate samples, which would cause hit-rate dispersion. Therefore to unify the same words written with slightly different spelling we calculated the similarity of the extracted words. The similarity was calculated using Levenshtein Distance [11] in a way similar to [12, 13]. The Levenshtein Distance between two strings is calculated as the minimum number of operations required to transform one string into another, where the available operations are only deletion, insertion or substitution of a single character.

However, as Japanese is transcribed using three character types, Chinese characters (*kanji*) encapsulating from one to several syllables and two additional syllabary (*katakana* and *hiragana*), calculating the distance would become very imprecise. To solve this problem, every word was automatically transformed in their alphabetical transcription. An example of distance calculation is presented on the back-transformation of the word *kimosu* to its original spelling *kimoi* in Table 3. The distance between both words is calculated as 2.

3.3 SVM Based Classification of Cyber-bullying

To classify the entries into either harmful (cyber-bullying) or non-harmful, we used Support Vector Machines. Support Vector Ma-

[6] http://www.internetslang.com/

Table 3. Three examples of colloquial transformations of Japanese words in online jargon.

transformed word	performed operation
kimosu	
→ *kimoiu*	substitution of 's' to 'i'; distance = 1;
→ *kimoi*	deletion of final 'u'; distance = 2;

chines (SVMs) are a method of supervised machine learning developed by Vapnik [14] and used for classification of data. They are defined as follows. With a set of training samples, divided into two categories A and B, SVM training algorithm generates a model for prediction of whether test samples belong to either category A or B. In the traditional description of SVM model, samples are represented as points in space (vectors). SVM constructs a hyperplane, in a space of a higher dimension than the base one, with the largest distance to the nearest training data points (support vectors). The larger the margin the lower the generalization error of the classifier. Since SVM has been successfully used for text classification [15] we decide to use them in this research as well. In our research the category A contains cyber-bullying cases and the category B contains all other cases, which do not consist of socially harmful information. As the software for building SVM models we used SVM_light (ver6.02)[7].

3.4 Extraction of Key Sentences

In the process of automation of Online Patrol, apart from the classification of cyber-bullying entries, there is a need to appropriately determine how harmful is a certain entry. A ranking according to the harmfulness of entries is important to detect the most dangerous cases. In our approach an entry is considered as the more harmful, the more vulgar keywords appear in the entry.

The harmfulness of an entry is calculated using T-score. T-score is a measure answering the question of how confident can one be that the association measured between two words is an actual collocation and not a matter of chance. The higher occurrence frequency a word has in a corpus, the higher is the value of T-score. A T-score of a word associating with words A and B is calculated according to the equation below:

$$T_{score} = \frac{a}{b} \tag{1}$$

where,

$$a = [word\ co-occurrence\ frequency] - \_\frac{([occurence\ of\ word\ A] * [occurence\ of\ word\ B])}{[all\ words\ in\ the\ corpus]}$$

and,

$$b = \sqrt{[word\ co-occurrence\ frequency]}$$

We calculate the harmfulness of the whole entry as a sum of T-scores calculated for all vulgar words. This way the more frequently occurring words there are in the entry, the higher rank the entry achieves in the ranking of harmfulness.

3.5 Evaluation of the Method

To verify the performance of the method we evaluated three procedures:

1. Classification of Cyber-bullying entries with SVM;
2. Word similarity calculation with Levenshtein distance;
3. Extraction of key sentences;

3.5.1 Evaluation of SVM Model

To apply SVM to the detection of harmful information from unofficial school BBS sites, we needed to prepare the data for training the SVM model. At first we performed morphological analysis of the BBS entries to be used as the training data. For every part of speech from the analyzed and parsed data we used the POS label names and strings of characters to identify the strings. As the features in the part of speech identification we used parts of speech like nouns (person's name), nouns (other than name), verbs and adjectives. In the identification of the whole entries we used feature sets consisting of the features of each part of speech and the strings of characters containing the whole entry. Based on the amount of identified features, SVM model calculates the probability of affiliation of a character string to a certain class.

As the training data we used 966 entries gathered during an actual Online Patrol, from which human annotators (Online Patrol members) classified 750 entries as harmful and 216 as non-harmful. This time however, we did not apply the string similarity calculation to the SVM model, to evaluate both techniques separately.

The above conditions are applied to test the model trained with SVM_light. In the evaluation we calculated the system's result as balanced F-score, using 10-fold cross validation for Precision and Recall. In 10-fold cross validation data is first broken into 10 sets of size n/10. Then, 9 datasets are used to train on and 1 as a test. This procedure is repeated 10 times and the overall score is the mean accuracy from all 10 tests. The balanced F-score, Precision (P) and Recall (R) are calculated as follows,

$$F_{score} = 2 * \frac{P * R}{P + R} \tag{2}$$

where,

$$Precision = \frac{s}{n} \qquad Recall = \frac{n}{c}$$

and,
$s = cases\ correctly\ classified\ by\ the\ system\ as\ harmful$
$n = all\ cases\ classified\ by\ the\ system\ as\ harmful$
$c = all\ harmful\ cases$

The result was Precision = 79.9%, Recall = 98.3% and the F-score = 88.2%.

3.5.2 Evaluation of Similarity Calculation

When preparing the conditions for evaluation of word similarity calculation method with Levenshtein distance, we noticed that the larger was the threshold, the higher was the probability of matching a word with completely different meaning. Therefore we performed an optimization of the similarity calculation algorithm. In the optimization we applied two heuristic rules shown in Table 4.

On BBS sites, where informal language is widely used, such as unofficial school Web sites, prolonging of syllables is used mostly to express changes in user attitude, or mood, usually highlighting the unofficial character of the entry. However, such operation has no influence on the semantics of a word. Therefore before the phase of similarity calculation we added a rule deleting all syllable prolongations.

[7] http://svmlight.joachims.org

Table 4. Two heuristic rules applied in optimization of similarity calculation algorithm.

Rule	Example
1. deletion syllable prolongations	*kimoooi* → *kimoi*
2. unification of word first letter	In case of *uzai* we will consider only the words beginning with *u*

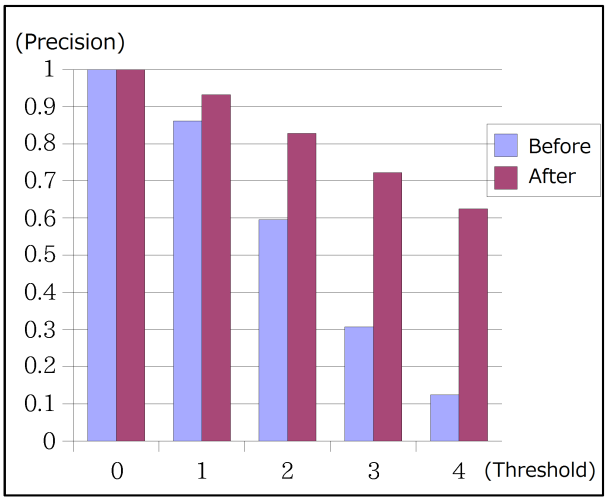

Figure 3. Precision of similarity calculation before and after applying the heuristic rules.

The second rule was unification of the first letters of matched words. This was done because in the Japanese language only the final letters change during conjugation. Therefore we could assume that it is irrelevant to calculate similarity for words differing with first letters.

The results for Precision of similarity calculation before and after applying the heuristic rules are presented on Figure 3. The Precision was greatly improved after applying the rules. With the threshold set on 2, the Precision before applying the rules was 58.9% and was improved to 85.0%.

3.5.3 Evaluation of Key Sentence Extraction

To set the ranking of cyber-bullying entries according to how harmful they are, we calculated the T-score of vulgar vocabulary they contain. The results are represented in Table 5. As one can see in Table 5, the first three scores are much higher then the other ones. This is caused by numerous repetitions of the same vulgar words in one entry, which artificially increases the occurrence rate. Moreover, in the cases, appearing in the table on places 4-6, one of the words was used repeatedly, which also caused artificial increase of occurrence and eventually the result of T-score, although smaller than for the first three cases. In such cases, although the occurrence of a certain pair is not high, T-score is artificially increased and therefore biased.

To solve this problem we considered the same vulgar words appearing in one entry as a one word. However, some vulgar words may appear as collocations. In such cases the identical numerous collocations are considered as one. The change in the results is represented in Table 6. The results indicate the bias have disappeared. However, the T-score became too small causing many word pairs appear on the same place in the ranking making the ranks difficult to set. This is caused by the fact that the words with the same meaning appear on the Web sites in different transcription and therefore there is a large number of word sets with a small occurrence.

We solved this problem by calculating word similarity before calculating T-score. The results are represented in Table 7. Thanks to word similarity calculation a rise in T-score was observed and the number of diversified word pairs increased, which made rank setting much easier. There was 202 vulgar word pairs, from which 40 pairs occurred more than twice.

Table 5. The results of T-score calculation for sets of vulgar words. As mentioned in section 3.2 words in Japanese can be usually transcribed in three different systems: *hiragana*, *katakana* and *kanji*. The differences in transcription are represented in the table as markers after the words, with [h] for *hiragana*, [k] *katakana* and [K] for *kanji*.

word A	word B	co-occurrence	T-score
baka [h] (stupid)	*baka* [h]	861	29.34
shine [K] (fuck you)	*shine* [K]	552	23.49
shine [h]	*shine* [h]	58	7.62
tarashi [k] (pimp)	*shine* [K]	17	4.12
busu [k] (ugly bitch)	*shine* [K]	16	4.00
kimosu [h]	*shine* [h]	16	4.00
shine [h]	*busu* [h]	6	2.45

Table 6. Change in the results of T-score calculation for sets of vulgar words, when two or more words were considered as one.

word A	word B	co-occurrence	T-score
shine [K]	*shine* [K]	7	2.65
shine [h]	*shine* [h]	4	2.00
pashiri [k] (looser)	*shine* [K]	3	1.73
debu [k]	*kiero* [K]	3	1.73
kiero [K] (get lost)	*kiero* [K]	2	1.41
shine [K]	*kiero* [K]	2	1.41
uzai [h]	*kimoi* [k]	2	1.41

Table 7. Change in the results of T-score calculation for sets of vulgar words, with calculated word similarity.

word A	word B	co-occurrence	T-score
shine [K]	*shine* [K]	11	3.32
kimoi [h]	*shine* [K]	11	3.32
kimoi [h]	*busaiku* [h]	8	2.83
uzai [h]	*kimoi* [h]	7	2.65
panko [k] (slut)	*panko* [k]	6	2.45
kimoi [h]	*kimoi* [h]	6	2.45
busaiku [h]	*busaiku* [h]	6	2.45

3.6 Discussion

This time in morphological analysis of vulgar words we used only a small set of manually found words, which we added to the lexicon. It is difficult to include all existing vulgar words and more such vocabulary will appear in the future as well. There fore as a further work we need to develop a method for automatic extraction of vulgar vocabulary from the Internet.

The results of SVM model used to distinguish between harmful and non-harmful information were 79.9% of Precision and 98.3% of Recall. However, on the unofficial school Web pages used as the data in this research there were numerous entries consisting of only one sentence, or even one word. Therefore the feature set for training was not sufficient and the overall result was not ideal (F-score = 88.2%).

As for word similarity calculation, many vulgar words are short and a change of even one letter might cause a change of meaning. Therefore, two different words would be matched as similar by Levenshtein distance, when the threshold is too wide. this problem might be solved by automatically setting the threshold according to the word length.

As for the extraction of key sentences, although we were able to calculate non biased T-score for vulgar expressions and set the ranking, over 80% of vulgar words appeared only once. This caused over half of the cases to be attached with similar ranks. This problem could be solved by increasing the number of training data or applying a different method of rank setting.

As was shown by Chen and colleagues [3, 4], analysis of affect intensity of Dark Web Forums often helps specifying the character of the forum. Aiming to find any dependencies between expressing emotions and cyber-bullying activities we performed an additional study and analyzed affective level of the cyber-bullying data.

4 AFFECT ANALYSIS OF CYBER-BULLYING DATA

For the comparative affect analysis we obtained additional Online Patrol data containing both cyber-bullying activities and normal entries. The data used in affect analysis contained 1,495 harmful and 1,504 non-harmful entries. The affect analysis was performed with ML-Ask system developed by Ptaszynski and colleagues [16].

4.1 ML-Ask - Affect Analysis System

The affect analysis system employed in the analysis of the cyber-bullying data described in this paper is ML-Ask developed by Ptaszynski et al. [16]. ML-Ask (eMotive eLements / Emotive Expressions Analysis System) was developed for analyzing the emotive contents of utterances. The system uses a two-step procedure: 1) Analyzing the general emotiveness of an utterance by detecting emotive elements, or emotemes, expressed by the speaker and classifying the utterance as emotive or non-emotive; 2) Recognizing the particular emotion types by extracting expressions of particular emotions from the utterance. This analysis is based on Ptaszynski's [21] idea of two-part classification of realizations of emotions in language into:

1) *Emotive elements* or *Emotemes*. Elements conveyed in an utterance indicating that the speaker was emotionally involved in the utterance, but not detailing the specific emotions. The same emotive element can express different emotions depending on context. This group is linguistically realized by subgroups such as interjections, exclamations, mimetic expressions, or vulgar language. Examples are: *sugee* (great!), *wakuwaku* (heart pounding), *-yagaru* (a vulgarization of a verb);

2) *Emotive expressions*. Parts of speech used to describe emotional states. However, they function as expressions of the speaker's emotions only in utterances where the speaker is emotionally involved. In non-emotive sentences they fulfill the function of simple descriptive expressions. The group is realized by various parts of speech, like nouns, verbs, adjectives, etc. Examples are: *aijou* (love), *kanashimu* (feel sad), *ureshii* (happy), respectively.

The hand-selected emotive element database consists of interjections, mimetic expressions, endearments, vulgarities, and representations of non-verbal emotive elements, such as exclamation marks or ellipses. The emoteme database collected and divided in this way contains 907 elements in total.

A system for analysis of emoticons was also added, as they are symbols commonly used in everyday text-based communication to convey emotions.

The database of emotive expressions is based on Nakamura's collection [22] and contains 2100 emotive expressions, each classified into the emotion type they express.

4.1.1 Affect Analysis Procedure

On textual input provided by the user, two features are computed in order: the emotiveness of an utterance and the specific type of emotion. To determine the first feature, the system searches for emotive elements in the input to determine whether the input is emotive or non-emotive. In order to do this, the system uses MeCab [17] for morphological analysis and separates every part of speech. MeCab recognizes some parts of speech which belong to the group of emotemes, such as interjections, exclamations or emphatic sentence-final particles, like *-zo*, *-yo*, or *-ne*. If these appear, they are extracted from the utterance as emotemes. Next, the system searches and extracts every emoteme based on the system's emoteme databases (907 items). The input is then processed by CAO [18], a system for analysis of emoticons. Emoticons belong to both emotemes, and emotive expressions, as their appearance in an input always indicates emotional attitude of the user and for most emoticons it is possible to specify the emotion type they convey. All of the extracted elements mentioned above (exclamations from MeCab, emotemes and emoticons) indicate the emotional level of the utterance. The emotional value of an input is calculated quantitatively, the more emotemes there were, the higher was the emotional value, however with the upper limit of 5.

Secondly, in utterances classified as emotive, the system uses a database of emotive expressions to search for all expressions describing emotional states. Here, emotions specified by CAO are added to the result of ML-Ask basic procedure. This determines the specific emotion type (or types) conveyed in the utterance. Examples of analysis performed by ML-Ask (system output) are represented below. In these examples, from top line there are: an example in Japanese, emotive information annotation (emotemes-underlined, emotive expressions-bold type font) and English translation.

4.1.2 Contextual Valence Shifters in ML-Ask

One of the problems in the procedure described above was confusion of the valence polarity of emotive expressions. The cause of this problem was extracting from the utterance only the emotive expression keywords without its grammatical context. One case of such an input is presented below in example (3). In this sentence the emotive expression is the verb *akirameru* (to give up [verb]) but the phrase *-cha ikenai* (Don't- [particle+verb]) suggest that the speaker is in fact negating and forbidding the emotion expressed literally. Such phrases are called Contextual Valence Shifters (CVS).

The idea of Contextual Valence Shifters (CVS) application in Sentiment Analysis was first proposed by Polanyi and Zaenen [23]. They distinguished two kinds of CVS: negations and intensifiers. The group of negations contains words and phrases like "not", "never", and "not quite", which change the valence polarity of the semantic orientation of an evaluative word they are attached to. The group of intensifiers contains words like "very", "very much", and "deeply",

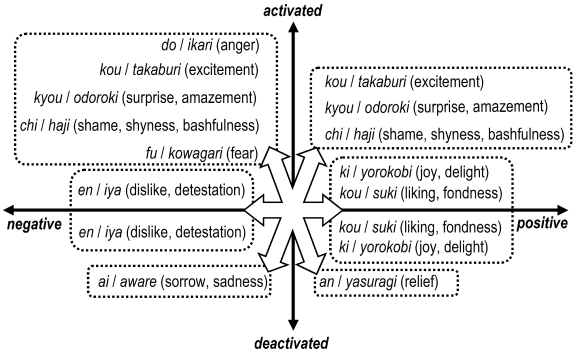

Figure 4. Grouping Nakamura's classification of emotions on Russell's space.

which intensify the semantic orientation of an evaluative word. Examples of CVS negations in Japanese are grammatical structures such as: *-nai* (not-), *amari -nai* (not quite-), *-to wa ienai* (cannot say it is-), or *-te wa ikenai* (cannot+[verb]-). Intensifiers are: *totemo-* (very-), *sugoku-* (-a lot), or *kiwamete-* (extremely). In this research we focused mostly on negations, since they have an immediate and significant influence on the meaning of emotive expressions. We applied Contextual Valence Shifters to change the valence polarity of emotive expressions in utterances containing CVS structures. Our hand-crafted database of CVS contains 71 negation structures.

However, using only the CVS analysis we would be able to find out about the appropriate valence of emotions conveyed in the utterance, but we would not know the exact emotion type. Therefore, to specify the emotion types in such utterances we applied the idea of the two-dimensional model of affect. To specify the emotion types after changing polarity with CVS, we applied the idea of the 2-dimensional model of affect [24] which assumes that all emotions can be described in 2-dimensions: the emotion's valence polarity (positive/negative) and activation (activated/deactivated). An example of positive-activated emotion could be "excitement"; a positive-deactivated emotion is, e.g., "relief" (see Figure 4).

Emotion types distinguished by Nakamura [22] were mapped on this model and their affiliation to one of the spaces determined. The emotion types with ambiguous affiliation were mapped on two possible fields. When a CVS structure is discovered, ML-Ask changes the valence polarity of the detected emotion. The appropriate emotion after valence changing is determined as the one with valence polarity and activation parameters different to the contrasted emotion (note arrows in Figure 4).

4.2 CAO - Emoticon Analysis System

Since one of the most popular strategies of expressing emotions in online communication is using emoticons, we supported ML-Ask with emoticon analysis system CAO. It is a system for estimation of emotions conveyed through emoticons developed by Ptaszynski and colleagues [18]. Emoticons are - sets of symbols widely used in text-based online communication to convey emotions. CAO, or *emotiCon Analysis and decOding of affective information*, extracts an emoticon form an input (a sentence) and determines specific emotion types expressed by it using a three-step procedure. Firstly, matching the input with a predetermined raw emoticon database containing over ten thousand emoticons. The emoticons, which could not be estimated with only the database are automatically divided into semantic areas, such as representations of "mouth" or "eyes", based on the idea of *kinemes*, or minimal meaningful body movements, applied from the theory of kinesics [19, 20]. The areas are automatically annotated according to their co-occurrence in database. The annotation is firstly based on eye-mouth-eye triplet. If no triplet was found, all semantic areas are estimated separately. This provides hints about potential groups of expressed emotions giving the system a coverage of over 3 million possibilities. CAO is used as a supporting procedure in ML-Ask to improve performance of the affect analysis system in utterances, which do not include emotive expressions, like in the example (2) below.

(1) *Kyo wa* *nante* ***kimochi ii*** *hi* *nanda* *!*
Today:TOP ML:nante **MX:joy** day:SUB ML:nanda ML:!
Translation: Today is such a nice day!

(2) *Iya∼,* *sore* *wa* *sugoi* *desu* *ne–* *!* ^o^
ML:iya∼ this :TOP ML:sugoi COP ML:ne– ML:! ML:^o^
Translation: Whoa, that's great! ^o^ **MX:joy**

(3) ***Akirame*** *cha* *ikenai* *yo* *!*
MX:dislike ML:cha| CVS:cha-ikenai{→joy} ML:yo ML:!
Translation: Don't cha give up!

(4) ***Hitoribocchi*** *nante* ***iya*** *da* *∼∼*
MX:sadness ML:nante–da **MX:dislike** COP ML:∼∼
Translation: Being alone sucks...

4.3 Affect Analysis Results

At first we calculated the number of all emotive entries among both sets of data. There was 956 emotive samples among 1,495 harmful (63.95%) and 1,029 among 1,504 non-harmful (68.42%) entries. The difference was not high and therefore the number of emotive entries cannot be considered as a highly distinctive feature. However, we made the first assumption, that harmful data are less emotively emphasized than non-harmful. This would be a reasonable assumption, since cyber-bullying is often based on irony or sarcasm, which is not highly emotive, however deliberately uses some amount of emotive information to slander the object of sarcasm. To confirm this thesis we performed other comparisons.

We calculated emotive values of all emotive entries. Although the number of emotive entries and emotive value, both relate to the idea of emotiveness of a corpus in general, they are not directly related. One can easily imagine the difference by imagining a corpus of many slightly emotive entries (high number of emotive entries, but low average emotive value) and another corpus with a small number of highly emotive entries (low number of emotive entries, but high average emotive value). The approximated emotive value for harmful and non-harmful data was 1.47 and 1.5, respectively. Here also the difference is not high, although, since both values are not directly related, this can moderately support the thesis.

Next, we took a closer look on the extracted emotemes. There are four groups of emotemes distinguished in ML-Ask: i) interjections, ii) exclamations, iii) vulgarities and iv) mimetic expressions (*gitaigo* in Japanese) - arranged in order of their emotional weight. Distribution of the extracted emotemes within both entry sets is represented in Table 8. There were relatively more emotemes with high emotive weight in non-harmful data and more low weight emotemes in harmful data, which is another confirmation of the thesis set on the

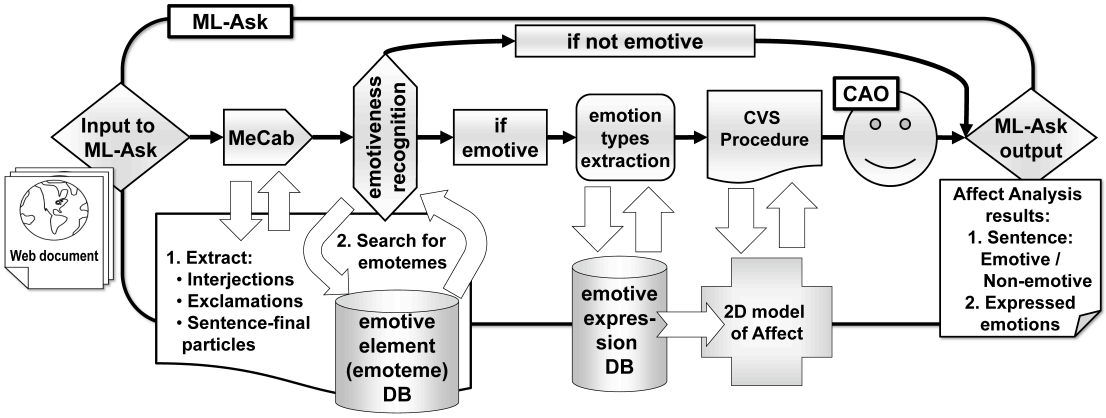

Figure 5. FLow chart of ML-Ask system.

beginning of this section. The biggest difference between the number of extracted emotemes was for found for vulgarities, which can be regarded as a distinctive feature of cyber-bullying entries. What is also important is the fact that this provides an experimental proof for the assumption made in the design of the machine learning based system for cyber-bullying detection, which is based on a vulgarity lexicon. A similar difference, although lower, also appeared in mimetic expressions, which could be included in further study on enlarging lexicon for the system. As for interjections and exclamations, although they appeared in a large number in both datasets, more of them appeared in non-harmful entries. This could be caused by the fact that these two emoteme types are used to express emotive attitude in a straightforward way. Therefore, although there certainly are cyber-bullying cases where the victims are slandered straightforwardly, not emotional and cold sarcasm is also an often phenomenon.

Table 8. Distribution of the extracted emotemes within both entry sets.

type of emoteme	type of data	
	non-harmful	harmful
interjections	**859**	784
exclamations	**284**	174
vulgarities	8	**149**
mimetic expressions	7	**23**

Another comparison was made with the closer study of only emotive utterances. ML-Ask is capable of detecting emotive utterances with a very high reliability, although it specifies particular emotion types with low Recall (although with high Precision) [16], because its lexicon (Nakamura's dictionary, cf. [22]) is said to be out of date. Therefore there a certain number of samples is always described as emotive but with no specified emotion types. Such a phenomenon is reasonable from the linguistic point of view, since there are many sentences that are emotive, although the emotion they convey depends on their context.

Therefore we first compared how many were there specified vs unspecified emotive utterances. The result was 13.18% vs 86.82% for cyber-bullying and 11.95% vs 88.05% for the normal entries. The higher ratio of specified emotion types in cyber-bullying data might suggest that people more often use traditional emotive expressions to slander people than they do to express emotions usually. However, low difference comparing to the non-harmful entries makes this statement weakly reliable.

Next, we compared the number of particular emotive expressions extracted by ML-Ask. The results have been represented in Table 9. The analysis of the number of emotion types represented by the extracted emotive expressions revealed interesting tendencies, although here as well the strength of the proof is low and therefore unequivocal. However, anger scored much higher in the cyber-bullying data than in the normal data, which is a reasonable result, since Web site entries meant to slander others are expected to express anger more often. Fear on the other hand scored very low in cyber-bullying data. This is also reasonable, since it is difficult to bully others by expressing one's fears. Dislike scored slightly higher in the harmful data. As for the positive emotion types, e.g., joy scored higher in normal dataset, which was expectable. On the other hand, fondness scored unexpectedly higher in cyber-bullying dataset. Detailed analysis revealed that people would often express strong sarcasm with the use of positive expressions. Some examples of such entries have been represented in Table 10.

Table 9. The number of particular emotive expressions extracted by ML-Ask from both datasets.

type of data			
non-harmful		harmful	
emotion type	no. extr.	emotion type	no. extr.
dislike	49	dislike	56
joy	32	fondness	49
fondness	21	joy	12
relief	18	relief	9
fear	11	anger	8
gloom/sadness	7	gloom/sadness	7
surprize	6	excitement	3
excitement	5	fear	3
anger	3	shame	3
shame	0	surprize	3

Finally, we compared tendencies in the annotated emotion types with regard to the two-dimensional affect space. The results are represented on Table 11.

In the valence dimension, negative emotions were annotated most often on harmful data, and positive emotions, on non-harmful data, which is a reasonable and predictable result. The differences were not

Table 10. Four examples of cyber-bullying entries gathered during Online Patrol from the unofficial school BBS Web site. The upper three represent expressing strong sarcasm despite of the use of positive expressions in the sentence. Emotemes - underlined; Emotive expressions - bolt type font.

>>104 Senzuri koite shinu nante, sonna hageshii senzuri sugee naa. "Senzuri masutaa" toshite isshou ***agamete*** *yaru yo.* Analysis results: Emotemes: *nante, sugee, naa, -yo, senzuri*; Emotive expression: ***agameru*** (worship, respect) → **fondness** ; Translation: >>104 Dying by 'flicking the bean'? Cannot imagine how one could 'flick the bean' so fiercely. I'll worship you forever, as a 'master-bator'.
2-nen no tsutsuji no onna meccha busu ***suki*** *na hito barashimashouka? 1-nen no ano ko desu yo ne? kimogatteru n de yamete agete kudasai* Analysis results: Emotemes: *meccha, -yo, -ne, -nde, kimogaru*; Emotive expressions: ***suki*** (like) → **fondness** ; Translation: Ya wanna know who likes 2nd-grade ugly azalea girls? Its that 1st-grader isn't it? He's looks disgusted, so leave him mercifully in peace.
*Aitsu wa busakute se ga takai dake no onna, busakute se takai dake ya no ni yatara otoko-****zuki*** *meccha tarashi de panko anna onna owatteru* Analysis results: Emotemes: *yatara, meccha, busai, tarashi, panko*; Emotive expressions: ***-zuki*** (-lover; an amateur of-) → **fondness** ; Translation: She's just tall and apart of that she's so freakin' ugly, and despite of that she's such a cock-loving slut, she's finished already.
Shinde kureeee, daibu ***kiraware-mono*** *de yuumei, subete ga* ***itaitashii****...* | Analysis results: Emotemes: *–eee* (syllable prolongation), *...* (elipsis); Emotive expressions: ***kiraware-mono*** (disliked by others) → **dislike**, ***itaitashii*** (pathetic, pitiful) → **gloom, sadness** ; Translation: Please, dieee, you're so famous for being disliked by everyone, everything in you is so pathetic

that obvious, however, after exclusion of fondness, which, as mentioned above, was often used mostly in sarcasm, the differences became clearer. The smallest difference was observed in emotion types which can be classified as both positive or negative, as their valence usually depends on the particular context.

As for the dimension of activation, non-harmful data was annotated as more vivid in the groups of deeply activated and deeply deactivated emotions. On the other hand, harmful data was annotated more often on emotions with moderately activated emotions, which provides another proof for the thesis set on the beginning of this section.

5 CONCLUSIONS AND FUTURE WORK

In this paper we presented a research on cyber-bullying, a new social problem that emerged recently, together with the development of social networking portals, etc. Cyber-bullying consist in sending messages containing slanderous expressions, harmful for other people, or verbally bullying other people in front of the rest of online community. In Japanese society, on which we focused in particular, this problem is particularly vivid on unofficial school Web sites. To handle the problem, teachers and PTA members perform voluntarily Online Patrol to spot and delete the online entries harmful for other people. Unfortunately, there is already an enormous number of cyber-bullying cases and keeps growing. Therefore the Online Patrol becomes an uphill task when performed manually.

Therefore we started this research to create an artificial Online Patrol agent. As the first step, we created a machine learning-based system for cyber-bullying detection and evaluated it. At first, we manually gathered a lexicon of vulgar words distinctive for cyber-bullying entries. To recognize the vulgar words, but written in an informal or jargonized way, we calculated word similarity with Levenshtein distance. As the result, with the threshold equal 2 the system was able to correctly determine the similar words with a 85% of Precision. The Support Vector Machine model trained for discrimination between harmful and non-harmful data achieved the results of 79.9% of Precision and 98.3% of Recall (with balanced F-score = 88.2%). Finally the rank setting of the online entries according to their harmfulness was set using T-score. We were able to eliminate undesirable bias in the rank setting, unfortunately, the procedure is not yet ideal.

Since new vulgar words appear frequently, we need to find a way to automatically extract new vulgarities from the Internet to keep the lexicon up to date. There is also a need for more experiments with the system including its different variations and improvements (ex. Levenshtein distance threshold optimization).

Looking for clues for further improvement of the system, we performed comparative affect analysis of the cyber-bullying data and normal entries. As a result, we noticed that, that the harmful data were less emotively emphasized than non-harmful. The thesis is reasonable, since the harmful entries are written with premeditation and aim not in expressing ones own emotions, but in evoking in other online community members negative emotions against victim of the cyber-bullying. The results of comparing different dimensions of emotional emphasis suggested the thesis was true, although the reliability of the proof was not satisfactory and further analysis on more robust data is necessary. Another discovery, although an expected one, was that positive emotions appeared more often in non-harmful data and negative emotions appeared more often in harmful data. However, detailed analysis revealed that, especially for fondness, the expressions of positive emotions are often used in strong sarcastic meaning. Therefore there is a need to analyze the data taking into consideration also other dimensions than the valence and the activation of an emotion. As one of such means we plan to apply Ptaszynski et al's [25, 26] system for contextual affect analysis verifying whether an emotion expressed in an utterance is appropriate for its context. We assume this will help in developing a sufficient model of formalization of cyber-bullying activities.

The problems concerning online security have been escalating ever since the birth of the Internet. Some of them are widely known to the society, such as spam e-mails or hacking, however more and more such problems, including cyber-bullying, appear by the day and social consciousness about them is not yet sufficient. There have been developed solutions for some of these problems (e.g., automatic spam e-mail detection, firewall protection, etc.), while others, like cyber-bullying, only keep escalating. Recognizing such problems and developing the remedy is one of the most urgent matters in the field of Artificial Intelligence today.

ACKNOWLEDGEMENTS

This research is partially supported by: a Research Grant from the Nissan Science Foundation; the GCOE Program founded by Japan's Ministry of Education, Culture, Sports, Science and Technology; the Research Project on Development of a Internet Word Book System for Dynamic Following of Information Transition (Project Number: 20500833).

The authors thank Mr. Motoki Matsumura from Human Rights Research Institute Against All Forms for Discrimination and Racism-MIE for providing data from unofficial school Web sites.

Table 11. Comparison of tendencies in annotation of emotion types with regard to the two-dimensional affect space.

valence (polarity) of emotions								
negative emotion			**negative/positive (both possible)**			**positive emotion**		
emotion type	non-harmful	harmful	emotion type	non-harmful	harmful	emotion type	non-harmful	harmful
gloom/sadness	7	7	excitement	**5**	3	joy	**32**	12
fear	**11**	3	shame	0	**3**	fondness	21	**49**
anger	3	**8**	surprize	**6**	3	relief	**18**	9
dislike	49	**56**				SUM	**71**	70
SUM	70	**74**	SUM	**11**	9	SUM (fondness excluded)	**50**	21

activation of emotions								
deactivated emotion			**moderately activated (deactivated/activated)**			**activated emotion**		
emotion type	non-harmful	harmful	emotion type	non-harmful	harmful	emotion type	non-harmful	harmful
gloom/sadness	7	7	joy	**32**	12	shame	0	**3**
relief	**18**	9	fondness	21	**49**	excitement	**5**	3
			dislike	49	**56**	fear	**11**	3
						anger	3	**8**
						surprize	**6**	3
SUM	**25**	16	SUM	102	**117**	SUM	**25**	20

REFERENCES

[1] L. Robinson. Debating the events of September 11th: Discursive and interactional dynamics in three online fora. *Journal of Computer-Mediated Communication*, 10(4), article 4 (2005).

[2] L. Leets. Responses to Internet hate sites: Is speech too free in cyberspace?", *Comm. Law and Policy*, vol. 6(2), pp. 287-317 (2001).

[3] H. Chen, J. Qin, E. Reid,W. Chung,Y. Zhou,W. Xi, G. Lai, T. Elhourani, A. Bonillas, F.-Y. Wang, and M. Sageman. The dark web portal: Collecting and analyzing the presence of domestic and international terrorist groups on the web,In: *Proc. 7th IEEE Int. Conf. Intelligent Transportation Systems*, Washington, DC, pp. 106-111 (2004).

[4] A. Abbasi, and H. Chen. Affect Intensity Analysis of Dark Web Forums,In: *IEEE Intelligence and Security Informatics*, pp. 282-288 (2007).

[5] G. A. Gerstenfeld, D. R.Grant, C. P. Chiang. Hate online: A content analysis of extremist Internet sites", *Anal. of Soc. Issues and Pub. Policy*, vol. 3(1), pp. 29-44 (2003).

[6] Ministry of Education, Culture, Sports, Science and Technology. *'Netto jou no ijime' ni kansuru taiou manyuaru jirei shuu (gakkou, kyouin muke)* ["Bullying on the net" Manual for handling and the collection of cases (directed to school teachers)] (in Japanese). Ministry of Education, Culture, Sports, Science and Technology (2008).

[7] B. Belsey. Cyberbullying: An Emerging Threat for the "Always On" Generation, http://www.cyberbullying.ca/pdf/Cyberbullying_Presentation_Description.pdf

[8] J. W. Patchin & S. Hinduja. Bullies move beyond the schoolyard: A preliminary look at cyberbullying. *Youth Violence and Juvenile Justice*, 4(2), 148-169 (2006).

[9] S. Hinduja, & J. W. Patchin. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Thousand Oaks, CA: Corwin Press (2009).

[10] H. Watanabe, W. Sunayama. *Denshi keijiban ni okeru yuuza no seishitsu no hyouka* [User's nature evalution on Bulletin Board System] (in Japanese). *IEICE Technical Report*, 105(652), 2006-KBSE, pp. 25-30 (2006).

[11] V. I. Levenshtein. Binary Code Capable of Correcting Deletions, Insertions and Reversals. *Doklady Akademii Nauk SSSR*, Vol. 163, No. 4, pp. 845-848 (1965).

[12] H. Minoru, E. Hirohide. *Nihongo OCR bun ni okeru eiji, katakana no superu ayamari teisei-hou* [Spelling Correction Method for English and Katakana in Japanese OCR Text] (in Japanese). *Transactions of Information Processing Society of Japan*, 38(7), pp. 1317-1327 (1997).

[13] K. Ryozo, H. Koudai, S. Tatsuya. *ProductionRule wo mochiita shisutemu hyougen to koshou shindan e no ouyou* [Modeling and Fault Diagnosis of Controlled Plant based on Production Rule] (in Japanese). *The Robotics and Mechatronics Conference 2005*, p. 16 (2005).

[14] V. Vapnik. *Statistical Learning Theory*, Springer (1998).

[15] T. Hirotoshi, M. Takafumi, H. Masahiko. *Support Vector Machine ni yoru tekisuto bunrui* [Text Categorization Using Support Vector Machines] (in Japanese), *IPSJ SIG Notes*, 98(99), pp.173-180 (1998).

[16] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Affecting Corpora: Experiments with Automatic Affect Annotation System - A Case Study of the 2channel Forum -', In *Proceedings of The Conference of the Pacific Association for Computational Linguistics 2009 (PACLING-09)*, pp. 223-228 (2009).

[17] T. Kudo, 'MeCab: Yet Another Part-of-Speech and Morphological Analyzer', 2001. http://mecab.sourceforge.net/

[18] M. Ptaszynski, P. Dybala, R. Rzepka and K. Araki. Towards Fully Automatic Emoticon Analysis System (^o^), In *Proceedings of The Fifteenth Annual Meeting of The Association for Natural Language Processing (NLP-2010)*, Tokyo (2010).

[19] R. L. Birdwhistell, *Introduction to kinesics: an annotation system for analysis of body motion and gesture*, Univ. of Kentucky Press (1952).

[20] R. L. Birdwhistell, *Kinesics and Context*, University of Pennsylvania Press, Philadelphia (1970).

[21] M. Ptaszynski. Boisterous language. Analysis of structures and semiotic functions of emotive expressions in conversation on Japanese Internet bulletin board forum - 2channel -. (in Japanese). M.A. Dissertation, UAM, Poznan (2006).

[22] A. Nakamura. *Kanjo hyogen jiten* [Dictionary of Emotive Expressions] (in Japanese). Tokyodo Publishing, Tokyo (1993).

[23] A. Zaenen and L. Polanyi. Contextual Valence Shifters. In *Computing Attitude and Affect in Text*, J. G. Shanahan, Y. Qu, J. Wiebe (eds.), Springer Verlag, Dordrecht, The Netherlands, pp. 1-10 (2006).

[24] J. A. Russell. A circumplex model of affect. *J. of Personality and Social Psychology*, 39(6):1161-1178 (1980).

[25] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki. Towards Context Aware Emotional Intelligence in Machines: Computing Contextual Appropriateness of Affective States. In *Proceedings of Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California, USA, pp. 1469-1474 (2009).

[26] M. Ptaszynski, P. Dybala, W. Shi, R. Rzepka and K. Araki. Contextual Affect Analysis: A System for Verification of Emotion Appropriateness Supported with Contextual Valence Shifters. *International Journal of Biometrics*, 2(2), pp. 134-154 (2010).