2016 Joint 8th International Conference on Soft Computing and Intelligent Systems and 2016 17th International Symposium on Advanced Intelligent Systems

Detecting Cyberbullying with Morphosemantic Patterns

Michal Ptaszynski Fumito Masui
Department of Computer Science
Kitami Institute of Technology
Kitami, 090-8507, Japan
Email: {ptaszynski,f-masui}@
cs.kitami-it.ac.jp

Yoko Nakajima
Department of
Information Engineering
Kushiro National
College of Technology
Kushiro, 084-0916, Japan
Email: yoko@
kushiro-ct.ac.jp

Yasutomo Kimura
Department of Information and
Management Science
Otaru University
of Commerce
Otaru, 047-0034, Japan
Email: kimura@
res.otaru-uc.ac.jp

Rafal Rzepka Kenji Araki
Graduate School of
Information Science
and Technology
Hokkaido University
Sapporo, 060-0808, Japan
Email: {rzepka,araki}@
ist.hokudai.ac.jp

***Abstract*—In this paper we study the validity of a novel method for extraction of morphosemantic patterns from sentences in the detection of cyberbullying, or humiliating and slandering people in the Internet. The patterns, consisting of both semantic and morphological information, are extracted from actual cyberbullying entries, provided by Human Rights Center, with a combinatorial algorithm and applied to a language classification task. The results were promising indicating that morphosemantic sentence representation is useful in the context of deceptive and provocative language used in cyberbullying.**

I. 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). The National Crime Prevention Council states that CB 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."[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.

To contribute to mitigating the problem of cyberbullying, in the present research we aim at developing a solution which would 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 this paper we specifically focus on developing a systematic approach to automatically detecting and classifying cyberbullying entries.

The outline of this paper is as follows. Firstly, we present some of the previous research related to ours on cyberbullying detection. Next, we describe the applied method and the dataset used in this research. Finally, we explain the evaluation settings, thoroughly analyze the results and discuss possible improvements.

II. Previous Research

Some of the first robust research on CB was done by Hinduja and Patchin, who performed numerous surveys about the subject in the USA [Patchin & Hinduja 2006], [Hinduja & Patchin 2009]. They found out that the harmful information may include threats, sexual remarks, pejorative labels, or false statements aimed to humiliate others. When posted on a social network, such as Facebook or Twitter, it may disclose humiliating personal data of the victim defaming and ridiculing them personally.

Cyberbullying has also been thoroughly studied and analyzed by Dooley, Pyżalski, and Cross (2009) [Dooley et al. 2009], who performed an in-depth comparative analysis of traditional face-to-face bullying and cyberbullying, while Lazuras et al. (2012) [Lazuras et al. 2012] discussed implications of cyberbullying for teachers in school environments.

There has been a small number of research on extracting harmful information from the Internet. For example, [Ishisaka and Yamamoto 2010] developed a dictionary of abusive expressions based on a large Japanese electronic bulletin board (BBS) *2channel*. In their research they labeled words and paragraphs in which the speaker explicitly insults other people with words and phrases like *baka* ("stupid"), or *masugomi no kuzu* ("trash of mass-mudia"). Based on which words appeared most often with abusive vocabulary, they extracted abusive expressions from the surrounding context.

Ptaszynski et al. performed affect analysis of small dataset of cyberbullying entries [Ptaszynski et al. 2010] to find out that distinctive features for cyberbullying 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

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

978-1-5090-2678-4/16 $31.00 © 2016 IEEE
DOI 10.1109/SCIS&ISIS.2016.89

them conclude SVMs are not ideal in dealing with frequent language ambiguities typical for cybrbullying.

Next, [Matsuba et al.2011] proposed a method to automatically detect harmful entries, in which they extended the SO-PMI-IR score [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 large numbers of candidates for harmful documents with an accuracy of 83% on test data.

Later, [Nitta et al. 2013] proposed an improvement to Matsuba et al.'s method. They used seed words from three categories (abusive, violent, obscene) to calculate SO-PMI-IR score and maximized the relevance of categories. Their method achieved 90% of Precision for 10% Recall. We used both of the above methods as a baselines for comparison due to similarities in used datasets and experiment settings.

Unfortunately, method by [Nitta et al. 2013], based on *Yahoo!* search engine API, faced a problem of a sudden drop in Precision (about 30 percentage-points) across two years, since being originally proposed. This was caused by change in information available on the Internet. In section IV-E we discuss the possible reasons for this change. Recently [Hatakeyama et al. 2015] tried to improve the method by automatically acquiring and filtering harmful seed words, with a considerable success.

Most of the previous research assumed that using vulgar words as seeds will help detecting cyberbullying. However, all of them notice that vulgar words are only one kind of distinctive vocabulary and do not cover all cases. We assumed that the harmfulness of the entry does not depend only on such words, but rather is expressed through patterns within the sentence structure. Therefore in this research we first of all did not focus on detecting vulgar words, as it was done in previous methods. We also did not restrict the scope of analyzed patterns to words, or phrases, but extended the search to sophisticated patterns with disjoint elements. Moreover, the success of detecting such entry would rely on how accurately the sentence structure is represented. Thus in our research to represent the sentences we used a novel representation method incorporating both morphological as well as semantic information.

III. Morphosemantic Patetrn Extraction Method

In this section we describe our method for extraction of morphosemantic patterns from sentences. The method consists of two stages. Firstly, the sentences are represented using a combination of semantic role labeling with morphological information. Secondly, frequent combinations of such patterns are extracted from training data using an automatic pattern extraction architecture.

A. Morphosemantic Patterns

In the first stage of the method, all sentences included in the dataset (see section IV-A for details), are represented in **morphosemantic structure** (MS). From sentences represented this way **morphosemantic patterns** (MoPs) are extracted during the second stage.

TABLE I: Examples of future referring words and phrases with their semantic and morphological representation.

Surface	Semantic (Semantic role, Category, etc.) and grammatical representation
mezasu ("aim to")	No change (activity)-action aiming to solve [a problem]-pursuit; Verb;
hōshin ("plan to")	Other; Noun;
mitooshi ("be certain to")	Action; Noun;
kentō ("consider to")	No change (activity)-action aiming to solve [a problem]-act of thinking; Noun;
-*suru* ("to do")	Change-creation or destruction-creation (physical); Verb;
-*iru* ("is/to be")	Verb;

The idea of morphosemantic structure has been described widely in linguistics and structural linguistics. For example, Levin and Rappaport Hovav (1998) [Levin and Malka 1998] distinguish them as one of the two basic types of morphological operations on words (mostly on verbs), which modify the Lexical Conceptual Structure (LCS), or the semantic representation of a word. As for practical application of the idea, Kroeger (2007) [Kroeger, 2007] applied morphosemantic structure to analyze an Indonesian suffix *–kan*. Later, Fellbaum et al. (2009) [Fellbaum et al. 2009] applied morphosemantic patterns to improve links between the synsets in WordNet. More recently, Raffaelli (2013)[Raffaelli 2013] used morphosemantic patterns to analyze a lexicon in Croatian, a language rich both morphologically and semantically. More recently [Nakajima et al. 2016] applied morphosemantic patterns to extract patterns of future referring sentences and applied them to the task of reasoning about the future unfolding of events in Japanese.

In our research we also used datasets in Japanese language, and applied morphosemantic structure for the same reason. Using only one representation (lexical, morphological, or semantic) narrows the spectrum of information encoded in the language.

We generated the morphosemantic model using semantic role labeling with additional morphological information. Below we describe in detail the process of morphosemantic representation of sentences.

At first, sentences from the datasets are analyzed using semantic role labeling (SRL). SRL provides labels for words and phrases according to their role in sentence context. For example, in a sentence "John killed Mary" the labels for words are as follows: John=`Actor`, kill[past]=`Action`, Mary=`Patient`. Thus the semantic representation of the sentence is "`[Actor]-[Action]-[Patient]`".

For semantic role labeling in Japanese we used **ASA**[2], a system, developed by Takeuchi et al. (2010) [Takeuchi et al. 2010], which provides semantic roles for

[2] http://cl.it.okayama-u.ac.jp/study/project/asa

TABLE II: One example of sentence analysis by ASA.

Example I: Romanized Japanese (RJ): *Ashita kare wa kanojo ni tegami o okuru darō.* / **Glosses:** Tomorrow he TOP her DIR letter OBJ send will (TOP: topic particle, DIR: directional particle, OBJ: object particle.) / **English translation (E):** He will [most probably] send her a letter tomorrow.

No.	Surface	Label
1	*ashita*	[Time-Point]
2	*kare ha*	[Agent]
3	*kanojo ni*	[Patient]
4	*tegami o*	[Object]
5	*okuru darou*	[State_change]-[Place_change]-[Change_of_place(physical)(between_persons)]-[Movement_of_property_to_others]-[Provide]

words and generalizes their semantic representation using an originally developed thesaurus. Examples of labels ASA provides for certain words are represented in Table I. Two examples of SRL provided by ASA are represented in Table II.

However, not all words are semantically labeled by ASA. The omitted words include those not present in the thesaurus, as well as grammatical particles, or function words not having a direct influence on the semantic structure of the sentence, but in practice largely contributing to the overall meaning. For such cases we used a morphological analyzer MeCab[3] in combination with ASA to provide morphological information, such as "Proper Noun", or "Verb". However, in its basic form MeCab provides morphological information for all words separately. Therefore, there often occurs a situation where a compound word is divided. For example "Japan health policy" is one morphosemantic concept, but in grammatical representation it takes form of "Noun Noun Noun". Therefore as a post-processing procedure we added a set of linguistic rules for specifying compound words in cases where only morphological information is provided.

Moreover, as it is shown in Table II, some labels provided by ASA are too specific. Therefore in order to normalize and simplify the patterns, we specified the priority of label groups in the following way.

1) Semantic role (Agent, Patient, Object, etc.)
2) Semantic meaning (State_change, etc.)
3) Category (Dog → Living animal → Animated object)
4) In case of no label by ASA perform compound word clustering for parts of speech (e.g., "Japan Health Policy" → `[Noun][Noun][Noun]` → `[Proper_Noun]`)

Furthermore, post-processing in the case of no semantic information is organized as follows.

- If a compound word can be specified, output the part-of-speech cluster (point 4 above).
- If it is not a compound word, output part-of-speech for each word.

Below is an example of a sentence generalized with the morphosemantic structure labeling method applied in this research.

- **Sentence** (in Romanized Japanese): *Nihon unagi ga zetsumetsu kigushu ni shitei sare, kanzen yōshoku ni yoru unagi no ryōsan ni kitai ga takamatte iru.*
- **English**: As Japanese eel has been specified as an endangered species, the expectations grow towards mass production of eel in full aquaculture.
- **MS**: `[Object][Agent][State_change][Action][Noun][State_change][Object][State_change]`

B. Automatic Extraction of Frequent Patterns

Having all sentences represented in morphosemantic structure as described in section III-A, we used SPEC, a system for extraction of sentence patterns developed by Ptaszynski et al. (2011) [Ptaszynski et al. 2011]. **SPEC**, or **S**entence **P**attern **E**xtraction ar**C**hitecturte is a system automatically extracting frequent sentence patterns distinguishable for a corpus (a collection of sentences). Firstly, the system generates ordered non-repeated combinations from all sentence elements. In every n-element sentence there is k-number of combination groups, such as that $1 \leq k \leq n$, where k represents all k-element combinations being a subset of n. The number of combinations generated for one k-element group of combinations is equal to binomial coefficient, represented in equation 1. In this procedure the system creates all combinations for all values of k from the range of $\{1, ..., n\}$. Therefore the number of all combinations is equal to the sum of combinations from all k-element combination groups, like in the equation 2.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1}$$

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

Next, the system specifies whether the elements appear next to each other or are separated by a distance by placing a wildcard ("*", asterisk) between all non-subsequent elements. SPEC uses all patterns generated this way to extract frequent patterns appearing in a given corpus and calculates their weight.

The weight can be calculated in several ways. Two features are important in weight calculation. A pattern is the more representative for a corpus when, firstly, the longer the pattern is (length k), and the more often it appears in the corpus (occurrence O). Thus the weight can be calculated by

- awarding length (`LA`),
- awarding length and occurrence (`LOA`),
- awarding none (normalized weight, `NW`).

The normalized weight w_j is calculated according to equation 3. Normalization is performed to make weights fit in range from +1 to -1, and is achieved by subtracting 0.5 from the initial score and multiplying the intermediate product by 2.

$$w_j = \left(\frac{O_{pos}}{O_{pos} + O_{neg}} - 0.5\right) * 2 \tag{3}$$

The generated list of frequent patterns can be also further modified. When two collections of sentences of opposite

[3] http://taku910.github.io/mecab/

TABLE III: Four examples of cyberbullying entries gathered during Internet Patrol. The upper three represent strong sarcasm despite of the use of positive expressions in the sentence. English translation below Japanese content.

>>104 Senzuri koi te shinu nante? sonna hageshii senzuri sugee naa. "Senzuri masutaa" toshite isshou agamete yaru yo.
>>104 Dying by 'flicking the bean'? Can't imagine how one could do it so fiercely. I'm gonna worship her as a 'master-bator', that's for sure.
2-nen no tsutsuji no onna meccha busu suki na hito barashimashoka? 1-nen no anoko desuyo ne? kimogatterunde yamete agete kudasai
Wanna know who likes that awfuly ugly 2nd-grade Azalea girl? Its that 1st-grader isn't it? He's disgusting, so let's leave him mercifully in peace.
Aitsu wa busakute sega takai dake no onna, busakute se takai dake ya noni yatara otoko-zuki meccha tarashide panko anna onna owatteru
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...
Please, dieeee, you're so famous for being disliked by everyone, everything in you is so pathetic

features (such as "harmful vs. non-harmful") is compared, the list will contain patterns that appear uniquely in only one of the sides (e.g., uniquely positive patterns and uniquely negative patterns) or in both (ambiguous patterns). Thus pattern list can be modified by

- using all patterns (`ALL`),
- erasing all ambiguous patterns (`AMB`),
- erasing only those ambiguous patterns which appear in the same number on both sides (due to their normalized weight being equal to 0, later called 'zero patterns", `0P`).

Moreover, a list of patterns will contain both the sophisticated patterns (with disjoint elements) as well as more common n-grams. Therefore the system can be trained on a model using

- all patterns (`PAT`), or
- only n-grams (`NGR`).

All combinations of the above modifications are further tested in the evaluation experiment.

IV. Evaluation Experiment

A. Dataset

At first we needed to prepare a dataset. We used the dataset created originally by [Matsuba et al. 2010] and developed further 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. The original data was provided by the Human Rights Research Institute Against All Forms for Discrimination and Racism in Mie Prefecture, Japan[4] and contains data from unofficial school Web sites and fora. The harmful and non-harmful sentences were manually labeled by Internet Patrol members according to instructions included in the MEXT manual for dealing with cyberbullying [MEXT 2008]. Some of those instructions are explained shortly below.

The MEXT definition assumes that cyberbullying happens when a person is personally offended on the Web. This includes disclosing the person's name, personal information and other areas of privacy. Therefore, as the first feature distinguishable for cyberbullying MEXT defines private names. This includes such information as:

- Private names and surnames,
- Initials and nicknames,
- Names of institutions and affiliations,

As the second feature distinguishable for cyberbullying MEXT defines any other type of personal information. This includes:

- Address, phone numbers,
- Questions about private persons (e.g. "Who is that tall guy straying on Computer Science Dept. corridors?"),
- Entries revealing other personal information (e.g. "I hate that guy responsible for the new project against cyberbullying.").

Also, according to MEXT, vulgar language is distinguishable for cyberbullying, due to its ability to convey offenses against particular persons. This is also confirmed in other literature [Patchin & Hinduja 2006], [Ptaszynski et al. 2010]. Examples of such words are, in English: *sh*t*, *f*ck*, or *b*tch*, in Japanese: *uzai* (freaking annoying), or *kimoi* (freaking ugly).

In the prepared dataset all entries containing any of the above information was classified as harmful. Some examples from the dataset are represented in Table III.

B. Dataset Preprocessing

As mentioned in section III-A, we propose representing the sentences in morphosemantic structure as a novel approach to detecting cyberbullying. However, we needed to verify empirically whether it is useful to use morphosemantics for this kind of data, or is it sufficient to only chose only one kind of representation. Therefore in the experiment we applied the following sentence preprocessing.

- **Parts of speech (`POS`):** Words are replaced with their representative morphemes and parts of speech.
- **Semantic roles (`SR`):** Words and phrases are replaced with their semantic representations within sentence context (semantic roles).
- **Morphosemantic patterns (`MoPs`):** The sentences are preprocessed using combined morphological and semantic information.

C. Experiment Setup

The preprocessed original dataset provided three separate training and test sets for the experiment (`POS`, `SR`, `MoPs`). The experiment was performed three times, one time for each kind of preprocessing to choose the best option. Using these preprocessed datasets we performed the classification as follows. Each test sentence was given a score calculated as a sum of weights of patterns extracted from training data and found in the input sentence (equation 4).

[4] http://www.pref.mie.lg.jp/jinkenc/hp/

TABLE IV: Comparison of best F-scores within threshold span and BEP for each version of the classifier. Best classifier version within each preprocessing kind - highlighted in bold type font; best overall - underlined.

	Highest F-score within threshold												BEP (P=R=F)		
	POS				Semantic Roles				MoPs				POS	Sem Rol	MoPs
	Pr	Re	F1	Acc	Pr	Re	F1	Acc	Pr	Re	F1	Acc			
PAT-ALL	0.53	0.95	0.68	0.55	0.59	0.80	0.68	0.64	**0.61**	**0.76**	**0.68**	**0.64**	0.61	0.67	0.64
PAT-0P	**0.53**	**0.95**	**0.68**	**0.55**	0.59	0.80	0.68	0.64	0.57	0.83	0.68	0.61	0.61	0.66	0.64
PAT-AMB	**0.53**	**0.95**	**0.68**	**0.55**	0.58	0.81	0.68	0.62	0.58	0.82	0.68	0.62	0.61	**0.67**	0.64
PAT-LA	**0.53**	**0.95**	**0.68**	**0.55**	0.60	0.78	0.68	0.64	0.61	0.74	0.67	0.64	0.61	0.67	0.64
PAT-LA-0P	0.52	0.95	0.68	0.54	0.59	0.79	0.68	0.63	0.59	0.80	0.68	0.62	0.59	0.66	0.62
PAT-LA-AMB	**0.53**	**0.95**	**0.68**	**0.55**	0.63	0.73	0.68	0.66	0.58	0.82	0.68	0.62	0.60	0.67	0.63
NGR-ALL	0.52	0.96	0.67	0.53	0.58	0.82	0.68	0.63	0.59	0.82	0.68	0.63	**0.61**	0.67	0.64
NGR-0P	0.52	0.95	0.67	0.54	0.58	0.82	0.68	0.63	0.59	0.81	0.68	0.63	0.57	0.60	0.54
NGR-AMB	0.50	1.00	0.67	0.50	0.54	0.89	0.67	0.57	0.49	1.00	0.66	0.49	0.61	0.67	**0.64**
NGR-LA	0.53	0.94	0.68	0.55	0.63	0.75	0.68	0.66	0.58	0.82	0.68	0.61	0.60	0.67	0.63
NGR-LA-0P	0.52	0.95	0.67	0.54	**0.63**	**0.74**	**0.68**	**0.67**	0.58	0.81	0.68	0.62	0.60	0.62	0.58
NGR-LA-AMB	0.57	0.76	0.65	0.59	0.56	0.82	0.67	0.60	0.56	0.74	0.64	0.58	0.61	0.67	0.63

$$score = \sum w_j, (1 \geq w_j \geq -1) \tag{4}$$

The results were calculated using standard Precision (P), Recall (R) and balanced F-score (F1), and additionally with standard Accuracy, for the whole threshold span. However, if the initial collection of sentences was biased toward one of the sides (e.g., sentences of one kind were in larger number or longer), there will be more patterns of a certain type. Thus, using a rule of thumb in evaluation (e.g., fixed threshold above which a new sentence is classified as either harmful or non-harmful) would not provide sufficiently objective view on results. Therefore we additionally performed threshold optimization to find the threshold for which the classifier achieved the highest scores.

For each version of the dataset preprocessing a 10-fold cross validation was performed. In one experiment 14 different versions of the classifier were compared. Since the experiment is performed for three different versions of preprocessing, we obtained overall number of 420 experiment runs. There were several evaluation criteria. Firstly, we looked at which version of the algorithm achieved the top scores within the threshold span. We also looked at break-even points (BEP) of Precision and Recall. Finally, we checked the statistical significance of the results. We used paired t-test because the classification results could represent only one of two classes (harmful or non-harmful). To chose the best version of the algorithm we compared separately the results achieved by each group of modifications, eg., "different pattern weight calculations", "pattern list modifications" and "patterns vs n-grams". We also compared the performance to the baseline [Nitta et al. 2013].

D. Results and Discussion

To summarize the results, we looked at which version of the algorithm achieved the top scores within the threshold span.

Firstly, we looked at standard balanced F-score to see if the clear winner can be selected by the simplest measure. Best F-score for all three kinds of preprocessing (Parts-of-speech [POS], Semantic Roles [SR] and Morphosemantic Patterns [MoPs]) reached the same maximum of 0.68. Therefore there was no clear winner, however, within this evaluation context, Semantic Roles, achieved the highest balance of F-score and Accuracy for the version of classifier trained on ngrams with length awarded and zero-patterns deleted (`NGR-LA-0P`). Morphosemantic Patterns were the second highest with classifier trained on all patterns with unmodified pattern list (`PAT-ALL`). The results were represented in Table IV.

To provide additional support for the results, we also looked into BEP (Break-Even Point of Precision and Recall). Here similarly Semantic roles achieved the highest score of 0.67, for classifier trained on patterns list with ambiguous patterns discarded (`PAT-AMB`). MoPs were second-best (0.64) when trained on ngrams with ambiguous patterns discarded (`NGR-AMB`). This could suggest, that, regardless of which preprocessing achieved was the highest scores, that training the classifier on a pattern list with ambiguous patterns deleted could result in achieving high BEP also in the future.

In the process of detecting cyberbullying messages, sometimes net-patrol members may want to focus not on finding many suspicious messages, but on the most harmful ones, or those which are harmful without a doubt. Therefore, we also looked at the highest Precision within the threshold. The results were represented in Table V.

The highest P was achieved by SR (`PAT-LA-AMB`) and POS (`NGR-LA`) (0.93). Both of those classifier versions incorporated length of the pattern in patetrn weight calculation (`LA`), thus it suggests that to achieve the highest P it could be useful to apply this pattern list modification also in the future.

However, for such high P both SR and POS achieved very low R (0.11 for SR and 0.06 for POS). Therefore, when it comes to P optimized for F, the highest score was achieved by MoPs (P=0.85 for F=0.18, for `NGR-ALL` *ex aequo* with `NGR-0P`).

We also looked at standard Accuracy as a supportive mean for evaluation. Similarly to previous results, SR achieved the highest maximum score (0.69), with MoPs being second (0.65).

TABLE V: Comparison of best Precision and Accuracy within the threshold span for each version of the classifier. Best classifier version within each preprocessing kind - highlighted in bold type font; best overall - underlined.

	Highest Precision within threshold												**Highest Accuracy within threshold**											
	POS				**Semantic Roles**				**MoPs**				**POS**				**Semantic Roles**				**MoPs**			
	Pr	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**	**Acc**	**Pr**	**Re**	**F1**	**Acc**
PAT-ALL	0.78	0.05	0.09	0.52	0.87	0.20	0.33	0.60	0.81	0.15	0.25	0.56	0.58	0.78	0.66	0.60	0.67	0.67	0.67	0.68	0.61	0.76	0.68	0.64
PAT-0P	0.78	0.05	0.09	0.52	0.87	0.28	0.42	0.63	0.81	0.15	0.25	0.56	0.58	0.78	0.66	0.60	0.67	0.67	0.67	0.68	0.61	0.75	0.67	0.64
PAT-AMB	0.80	0.02	0.04	0.51	0.89	0.04	0.07	0.53	0.79	0.17	0.28	0.56	**0.58**	**0.78**	**0.66**	**0.61**	0.64	0.70	0.67	0.66	0.61	0.74	0.67	0.64
PAT-LA	0.78	0.05	0.09	0.52	0.89	0.03	0.06	0.53	0.79	0.17	0.27	0.56	0.58	0.78	0.66	0.60	0.67	0.66	0.67	0.68	0.61	0.74	0.67	0.64
PAT-LA-0P	0.76	0.11	0.20	0.54	0.92	0.03	0.05	0.52	0.72	0.14	0.24	0.56	0.59	0.63	0.61	0.60	0.71	0.56	0.63	0.68	0.61	0.73	0.67	0.64
PAT-LA-AMB	0.76	0.12	0.21	0.54	**0.93**	**0.06**	**0.11**	**0.54**	0.70	0.17	0.27	0.56	0.60	0.56	0.58	0.60	0.63	0.73	0.68	0.66	0.61	0.72	0.66	0.64
NGR-ALL	0.92	0.02	0.04	0.51	0.87	0.20	0.33	0.60	**0.85**	**0.10**	**0.18**	**0.55**	0.63	0.53	0.58	0.61	**0.80**	**0.49**	**0.61**	**0.69**	0.64	0.64	0.64	0.65
NGR-0P	0.92	0.02	0.04	0.51	0.87	0.20	0.33	0.60	**0.85**	**0.10**	**0.18**	**0.55**	0.63	0.53	0.58	0.61	**0.80**	**0.49**	**0.61**	**0.69**	**0.62**	**0.72**	**0.67**	**0.65**
NGR-AMB	0.65	0.21	0.32	0.55	0.69	0.14	0.23	0.55	0.54	0.71	0.61	0.56	0.54	0.83	0.65	0.56	0.54	0.89	0.67	0.57	0.54	0.71	0.61	0.56
NGR-LA	**0.93**	**0.03**	**0.06**	**0.51**	0.88	0.02	0.05	0.52	0.83	0.14	0.25	0.56	0.62	0.54	0.58	0.61	0.78	0.50	0.61	0.69	0.63	0.69	0.66	0.64
NGR-LA-0P	0.91	0.03	0.06	0.51	0.89	0.03	0.05	0.52	0.83	0.15	0.25	0.56	0.59	0.72	0.65	0.61	0.78	0.50	0.61	0.69	0.63	0.68	0.65	0.64
NGR-LA-AMB	0.66	0.31	0.42	0.57	0.75	0.24	0.36	0.59	0.60	0.49	0.54	0.59	0.58	0.71	0.64	0.60	0.56	0.53	0.67	0.60	0.56	0.73	0.63	0.59

To confirm whether the above results are not a matter of chance, we also calculated statistical significance of the results using the paired two-tailed Student's T-test for F-score and Accuracy results for those classifier versions which achieved highest BEP. We selected this significance test due to the fact that the classification could result in only one of two labels, namely, either "harmful" or "non-harmful." The differences between POS and SR or MoPs were always statistically significant. This means that when SR or MoPs achieve higher scores than POS the improvement can be considered as reliable and not bound by chance. On the other hand, the differences between SR and MoPs were always not statistically significant. This suggests, that although SR achieved in some cases higher scores than MoPs, this advantage could be a matter of chance. This means that both SR and MoPs remain viable and further experiments on larger datasets are required to finally specify which of the dataset preprocessing is more effective.

E. Comparison with Previous Methods

After analyzing various multiple settings for the proposed method, we compared it to previous methods. In the comparison we used the method by [Matsuba et al.2011], [Nitta et al. 2013], and its most recent improvement by [Hatakeyama et al. 2015]. However, since the latter extracts cyberbullying relevance values from the Web, apart from comparing to the results reported in the papers we also repeated their experiment to find out how the performance of the Web-based method changed during the three years since being originally proposed. Finally, to make the comparison more fair, we compared our best and worst results. As the evaluation metrics we used area under the curve (AUC) on the graph showing Precision and Recall, the same metrics used in the above mentioned research. The results were represented in Figure 1.

TABLE VI: Results of the paired two-tailed Student's T-test for F-score and Accuracy for the classifier versions which achieved highest BEP.

	F-score			Accuracy		
	POS	SR	MoPs	POS	SR	MoPs
POS		0.0248* (p<0.05)	0.0079** (p<0.01)		0.0004*** (p<0.001)	0.0001*** (p<0.001)
SR			0.3077 (p>0.05)			0.2079 (p>0.05)

The highest overall results when it comes to AUC were obtained by the best settings of the proposed method (trained on pattern list with semantic roles, length awarded in weight calculation and ambiguous patterns discarded - `SemRol/PAT-LA-AMB`), which starts from a high 93% and retains the Precision between 90% to 70% for major part of the threshold. The highest Precision score (93%) outperformed the one by [Nitta et al. 2013] (91%). Moreover, the Precision-performance of their method decreases more quickly. However, when we repeated their experiment in 2015, the results of their method greatly dropped. After thorough analysis of the experiment data we noticed that most of the information extracted in 2013 was not available in 2015. [Hatakeyama et al. 2015] in their discussion provides three most probable reasons for this drop, namely, (1) fluctuation in page rankings (hindering information extraction), (2) the net-patrol movement itself (frequent deletion requests of harmful contents sent to service providers by PTA members), and (3) recent tightening of usage policies by most Web service providers, such as Google[5], Twitter[6] and *Yahoo!* used by [Nitta et al. 2013]. The two latter ones are fact positive reasons, and it is difficult to consider "improving" the situation. Therefore the area for improvement is in modifying the information extraction procedure. Initial study in this matter was performed by [Hatakeyama et al. 2015].

V. Conclusions and Future Work

In this paper we proposed a novel method for the detection of cyberbullying (CB) by automatically extracting morphosemantic patterns from sentences and applying them in classification of messages on the Internet. Cyberbullying is a recently noticed social problem which influences mental

[5] https://www.google.com/events/policy/anti-harassmentpolicy.html

[6] https://blog.twitter.com/2014/building-a-safer-twitter

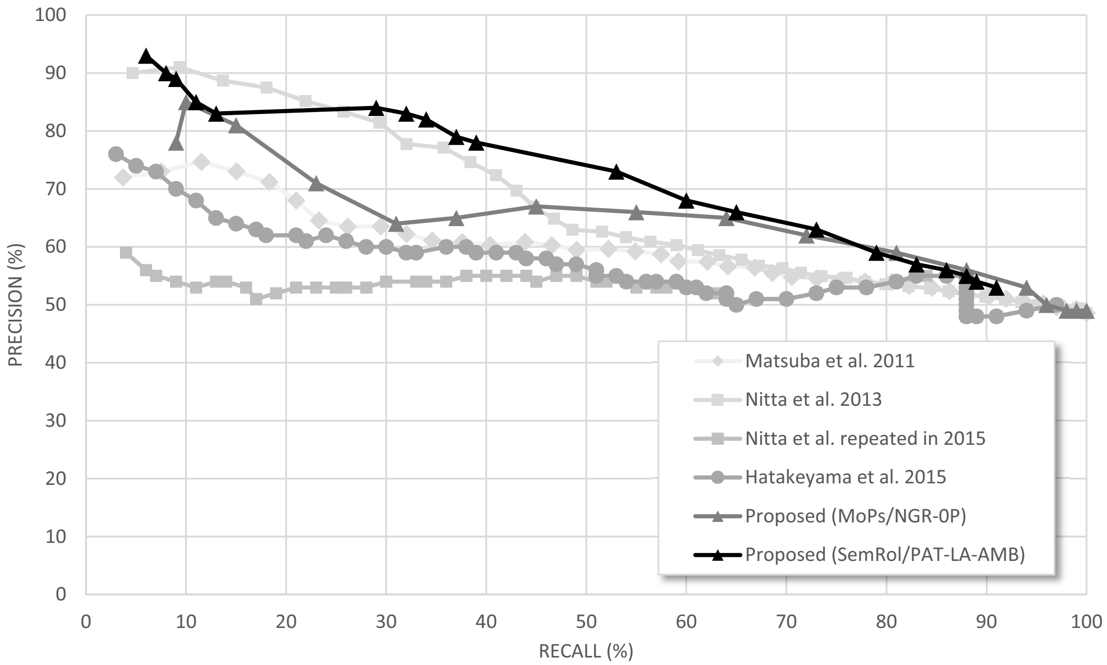

Fig. 1: Comparison between the proposed method (best and worst performance) and previous methods.

health of Internet users, and might lead to self-mutilation and even suicide of CB victims.

The morphosemantic patterns, containing both semantic and morphological information, were extracted from actual cyberbullying entries, provided by Human Rights Center, with a combinatorial algorithm and applied to a language classification task. The results show our method outperformed previous methods. It is also more efficient as it requires minimal human effort.

In the near future we plan to apply the proposed method in practice and propose an improvement to the original method by proposed originally by [Nitta et al. 2013] and presently developed by [Hatakeyama et al. 2015]. In this regard we plan to apply the proposed method to extract specific morphosemantic patterns most related to cyberbullying contents and apply them in information extraction of the original method.

Although the dataset applied in this study was sufficient to fully evaluate the proposed method, we also plan to obtain new data to evaluate the method even more thoroughly, and apply different classifiers. Finally, we plan to verify the actual amount of CB information on the Internet and reevaluate the method in more realistic conditions.

Acknowledgemets

This research was supported by (JSPS) KAKENHI Grant-in-Aid for Encouragement of Young Scientists (B) (Project Number: 15K16044).

References

[Amrita 2014] Amrita Paul. 2014. Effect of imbalanced data on document classification algorithms. Diss. Auckland University of Technology.

[Belsey 2007] Bill Belsey. 2007. Cyberbullying: An Emerging Threat for the "Always On" Generation, http://www.cyberbullying.ca/pdf/Cyberbullying_Presentation_ Description.pdf

[Fujii et al. 2010] 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 co-occurence information for filtering 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.

[Hatakeyama et al. 2015] Suzuha Hatakeyama, Fumito Masui, Michal Ptaszynski, Kazuhide Yamamoto. 2015. Improving Performance of Cyberbullying Detection Method with Double Filtered Point-wise Mutual Information, In *Demo Session of The 2015 ACM Symposium on Cloud Computing 2015 (ACM-SoCC 2015)*, Kohala Coast, Hawaii, August 27 - 29, 2015.

[He and Garcia 2009] Haibo He, and Edwardo A. Garcia. 2009. Learning from imbalanced data, *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 9 (2009), pp. 1263-1284.

[Hashimoto et al. 2010] 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* [Implementing a function for filtering harmful slang words into the semantic analysis system SAGE] (in Japanese), *IPSJ SIG Notes* 2010-SLP-81(14), pp. 1-6.

[Hinduja & Patchin 2009] Sameer Hinduja, J. W. Patchin. 2009. *Bullying beyond the schoolyard: Preventing and responding to cyberbullying*. Corwin Press.

[Ikeda and Yanagihara 2010] Kazushi Ikeda, Tadashi Yanagihara. 2010. *Kakuyōso no chūshōka ni motozuku ihō-, yūgai-bunsho kenshutsu shuhō no teian to hyōka* [Proposal and evaluation of a 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.

[Ishisaka and Yamamoto 2010] 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.

[Kilgarriff 2007] Adam Kilgarriff. 2007. Googleology is bad science. *Computational linguistics*, Vol. 33, No. 1, pp. 147-151.

[Krippendorff 1986] Klaus Krippendorff. 1986. Combinatorial Explosion, In: Web Dictionary of Cybernetics and Systems. Principia Cybernetica Web.

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

[Matsuba et al. 2010] Tatsuaki Matsuba, Fumito Masui, Atsuo Kawai, Naoki Isu. 2010. *Gakkou hikoushiki saito ni okeru yuugai jouhou kenshutsu* [Detection of harmful information on informal school websites] (In Japanese). In *Proc. of The 16th Annual Meeting of The Association for Natural Language Processing (NLP2010).*

[Matsuba et al.2011] 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* [A 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.

[MEXT 2008] 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.

[Nakajima et al. 2014] Yoko Nakajima, Michal Ptaszynski, Hirotoshi Honma, Fumito Masui. 2014. Investigation of Future Reference Expressions in Trend Information, In *Proceedings of the 2014 AAAI Spring Symposium Series*, "Big data becomes personal: knowledge into meaning For better health, wellness and well-being ", pp. 31-38, Stanford, USA, March 24-26, 2014.

[Nitta et al. 2013] Taisei Nitta, Fumito Masui, Michal Ptaszynski, Yasutomo Kimura, Rafal Rzepka, Kenji Araki. 2013. Detecting Cyberbullying Entries on Informal School Websites Based on Category Relevance Maximization. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013)*, pp. 579-586.

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

[Ptaszynski et al. 2009] Michal Ptaszynski, Pawel Dybala, Rafal Rzepka and Kenji Araki. 2009. 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 (PACLING-09)*, pp. 223-228.

[Ptaszynski et al. 2010] Michal Ptaszynski, Pawel Dybala, Tatsuaki Matsuba, Fumito Masui, Rafal Rzepka, Kenji Araki, and Yoshio Momouchi. 2010. In the Service of Online Order: Tackling Cyber-Bullying with Machine Learning and Affect Analysis. *International Journal of Computational Linguistics Research*, Vol. 1, Issue 3, pp. 135-154.

[Ptaszynski et al. 2011] Michal Ptaszynski, Rafal Rzepka, Kenji Araki, Yoshio Momouchi. 2011. Language combinatorics: A sentence pattern extraction architecture based on combinatorial explosion. *International Journal of Computational Linguistics (IJCL)*, Vol. 2, No. 1, pp. 24-36.

[Ptaszynski et al. 2014] Michal Ptaszynski, Fumito Masui, Rafal Rzepka, Kenji Araki. 2014. Emotive or Non-emotive: That is The Question, In *Proceedings of the 5th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2014)*, pp. 59-65, held in conjunction with *The 52th Annual Meeting of The Association for Computational Linguistics (ACL 2014)*, Baltimore, USA, June 22-27, 2014.

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

[Watanabe & Sunayama 2006] H. Watanabe, W. Sunayama. 2006. *Denshi keijiban ni okeru yūza no seishitsu no hyōka* [User nature evalution on BBS] (in Japanese). *IEICE Technical Report*, 105(652), 2006-KBSE, pp. 25-30.

[Dooley et al. 2009] Dooley J. J., Pyżalski J., and Cross D. 2009. Cyberbullying Versus Face-to-Face Bullying: A Theoretical and Conceptual Review, *Zeitschrift für Psychologie / Journal of Psychology*, Vol. 217(4), pp. 182-188.

[Lazuras et al. 2012] Lazuras L., Pyżalski J., Barkoukis V., Tsorbazoudis H. 2012. Empathy and Moral Disengagement in Adolescent Cyberbullying: Implications for Educational Intervention and Pedagogical Practice, *Studia Edukacyjne*, nr 23, pp. 57-69.

[Levin and Malka 1998] Beth Levin and Malka Rappaport Hovav, *Morphology and Lexical Semantics*, In Spencer and Zwicky, eds., pp. 248271, 1998.

[Kroeger, 2007] Paul Kroeger, "Morphosyntactic vs. morphosemantic functions of Indonesian *-kan*." In *Architectures, Rules, and Preferences: Variations on Themes of Joan Bresnan*, edited by A. Zaenen, J. Simpson, T. H. King, J. Grimshaw, J. Maling and C. Manning, pp. 229251, Stanford, CA: CSLI Publications, 2007.

[Fellbaum et al. 2009] Christiane Fellbaum, Anne Osherson, and Peter E. Clark, "Putting semantics into WordNet's "morphosemantic" links." *Human Language Technology. Challenges of the Information Society*, Springer Berlin Heidelberg, pp. 350-358, 2009.

[Raffaelli 2013] Ida Raffaelli, "The model of morphosemantic patterns in the description of lexical architecture." *Lingue e linguaggio*, Vol. 12, No. 1 (2013), pp. 47-72, 2013.

[Takeuchi et al. 2010] Koichi Takeuchi, Suguru Tsuchiyama, Masato Moriya, Yuuki Moriyasu, "Construction of Argument Structure Analyzer Toward Searching Same Situations and Actions." *IEICE Technical Report*, Vol. 109, No. 390, pp. 1-6, 2010.

[Nakajima et al. 2016] Yoko Nakajima, Michal Ptaszynski, Fumito Masui, Hirotoshi Honma, "A Method for Extraction of Future Reference Sentences Based on Semantic Role Labeling", *IEICE Transactions on Information and Systems*, Vol.E99-D, No.2, pp.514-524, Feb. 2016.