Cyberbullying Detection and Classification Using Information Retrieval Algorithm

B. Sri Nandhini
Student Pondicherry
Engineering college
Pondicherry-605014
+918122101602
nandhinibs.18@pec.edu

J.I. Sheeba
Assistant Professor
Pondicherry Engineering
college Pondicherry-605014
+919443084976
sheeba@pec.edu

ABSTRACT

Social networking site is being rapidly increased in recent years, which provides platform to connect people all over the world and share their interests. However, Social Networking Sites is providing opportunities for cyberbullying activities. Cyberbullying is harassing or insulting a person by sending messages of hurting or threatening nature using electronic communication. Cyberbullying poses significant threat to physical and mental health of the victims. Detection of cyberbullying and the provision of subsequent preventive measures are the main courses of action to combat cyberbullying. A cyberbully detection system to identify and classify cyberbullying activities such as Flaming, Harassment, Racism and Terrorism in Social network is proposed. Cyberbully detection is done using Levenshtein algorithm and classification of Cyberbully activity is carried out using Naïve Bayes classifier.

Categories and Subject Descriptors

Social networking sites- *Security, Information Retrieval algorithm.*

General Terms

Bullying, Detection, Classification.

Keywords

Social Network, Cyberbullying, Lavenshtein algorithm, Naïve Bayes classifier, Network security.

1. INTRODUCTION

With the proliferation of the Internet, security is becoming an important concern. While Web 2.0 provides easy, interactive, anytime and anywhere access to the online communities, it also provides platform for cybercrimes like cyberbullying. Life annoying cyberbullying experiences among young people have been reported internationally, thus drawing attention to its negative impact. In the USA, traces of cyberbullying is highly increasing and it has officially been identified as a social threat.

ICARCSET '15, March 06 - 07, 2015, Unnao, India Copyright 2015 ACM 978-1-4503-3441-9/15/03…$15.00
http://dx.doi.org/10.1145/2743065.2743085

There is an urgent need to study cyberbullying in terms of its detection, prevention and mitigation.

Bullying as a form of social turmoil has occurred in various forms over the years with the WWW and communication technologies being used to support deliberate, repeated and hostile behavior by an individual or group, in order to harm others[1]. Cyberbullying is defined as an intentional aggressive act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who cannot easily defend him or herself[2].

Recent research has shown that most teenagers experience cyberbullying during their online activities including mobile phone usage[3], and also while involved in online gaming or social networking sites. As highlighted by the National Crime Prevention Council, approximately 50% of the youth in America are victimized by cyberbullying[4]. The implications of cyberbullying[5] become serious (suicidal attempts) when the victims fail to cope with emotional strain from abusive, threatening, humiliating and aggressive messages. The impact of cyberbullying is exasperated by the fact that children are reluctant to share their predicament with adults, driven by the fear of losing their mobile phone and/or Internet access privileges[6]. The challenges in fighting cyberbullying include: detecting online bullying when it occurs; reporting it to law enforcement agencies, Internet service providers and others and identifying predators and their victims.

The Levenshtein distance is a metric for measuring the amount of difference between two sequences (i.e. an edit distance). The Levenshtein distance between two strings is defined as the minimum number of edits needed to transform one string into the other, with the allowable operations like adding, removing, or substituting a single character.

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set.

The growth of cyberbullying activities is increasing as equally as the growth of social networks. Cyberbullying activity poses a significant threat to mental and physical health of the victims. Study about effects of bullying is present but implementation for monitoring social network to detect cyberbullying activities is less.

Hence, the proposed system focuses on detecting the presence of cyberbullying activity in social networks using Levenshtein algorithm which helps government to take action before many users becoming a victim of cyberbullying.

2. RELATED WORK

In a recent study on cyberbullying detection, gender specific features were used and users are categorized into male and female groups. It is limited only to gender feature.

In other study[7], NUM and NORM features were devised by assigning a severity level to the bad words list (nosewaring.com). NUM is a count and NORM is a normalization of the bad words respectively. The dataset consisted of 3,915 posted messages crawled from the Web Site, Formspring.me. It showed only 58.5% accuracy, which is very less accuracy.

In [8] system allowing OSN users to have a direct control on the messages posted on their walls. This is done by using flexible rule-based system, this system allows users to customize the filtering criteria to be applied to their walls, and a Machine Learning based classifier will automatically label messages using content-based filtering. This approach is incapable of capturing more complex relationships at a deeper semantic level.

In[9] a research work by Massachusetts Institute of Technology a system to detect cyberbullying through textual context in YouTube video comments is developed. The system classifies the comment in a range of sensitive topics such as sexuality, culture, intelligence, and physical attributes and determining what topic it is. The system shows less precise classification outcome and increased false positives.

In[10] using a bag-of-words approach examined a baseline text mining system and improved by including sentiment and contextual features. Even with those models, vector machine learners produce a recall level of 61.9%.

In [11] bullying traces are identified using a variety of natural language processing techniques. Online and offline instances of bullying are traced. To identify the bullying they use sentiment analysis system and Latent Dirichlet Analysis to identify topics. In this method, the instances of bullying are not accurately detected.

Other interesting works[12] in this area performed harassment detection from comments and chat datasets provided by a content analysis workshop (CAW). Various features were generated including: TFIDF as local features; sentiment feature, which includes second person and all other pronouns like „you", „yourself", „him", „himself" and foul words; and contextual features[13]. Increased false positive is its limitation. There are also some software products available for fighting against cyberbullying like Bsecure, Cyber Patrol, eBlaster, IamBigBrother, Kidswatch[14][15][16][17][18].

Generally most existing systems are focusing on effects after cyberbullying incident and there is no system for online cyberbullying detection. The proposed system is to detect the cyberbullying activities and classify them as Flaming, Harassment, Racism and Terrorism, which helps to prevent the cyberbullying victims from facing effects of cyberbullying and take necessary actions like blocking, law enforcement or taking corresponding legal actions accordingly.

3. PROPOSED FRAMEWORK

In the proposed framework shown in Figure 1, the process of detecting cyberbully activities begins with input dataset from social network. Input is text conversation collected from formspring.me. Input is given to data pre-processing which is applied to improve the quality of the research data and subsequent analytical steps, this includes removing stop words, extra characters and hyperlinks. After performing pre-processing on the input data, it is given to Feature Extraction.

Feature Extraction is done to obtain features like Noun, Adjective and Pronoun from the text and statistics on occurrence of word (frequency) in the text.

The cyberbullying words are given as training dataset. With the training dataset the preprocessed online social network conversation is tested for bullying word presence. Lavenshtein distance algorithm detects the cyberbully words present in the conversation and displays it. For cyberbully Classification, Naïve Bayes classifier is used. A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions.

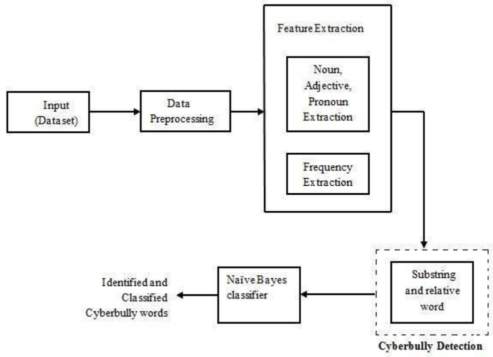


Figure 1. Proposed Framework of Cyberbully Detection and Classification system.

In the proposed framework for detecting cyberbully activities, following steps have been included

- Input
- Data Preprocessing
- Feature Extraction
- Cyberbully detection
- Naïve Bayes Classifier

3.1 Input

Textual conversation collected from social networks (Formspring.me, Myspace.com) is given as input.

3.2 Data Preprocessing

The data preprocessing is an important phase in representing data in feature space to the classifiers. Social network data are noisy, thus preprocessing has been applied to improve the quality of the research data and subsequent analytical steps, and this includes removing stop words.

The stop words add little semantic value to a sentence. Stop words are usually words like “to”, “I”, “has”, “the”, “be”, “or”, etc. Stop words bloat memory space and processing time without providing any extra value.

3.3 Feature Extraction

3.3.1 Noun Adjective and pronoun Extraction

Parsing noun, adjective and pronoun involves two steps such as part-of-speech tagging and extracting noun, adjective and pronoun from the tagged output. The part-of-speech tagging (POS tagging or POST), also called grammatical tagging, is the process of marking up a word in a text as corresponding to a particular part of speech.

The Part-of-speech tagging is carried out using the package provided by Stanford Natural Language Processing.

3.3.2 Frequency Extraction

The frequency extraction module involves extracting the occurrence count of the words parsed in the parser module. Using the frequency of the word the class probability is calculated using Naïve Bayes classifier.

3.4 Cyberbully Detection

Cyberbully detection is carried out using the Lavenshtein distance algorithm.

Lavenshtein distance Algorithm:

Given two strings a and b on an alphabet Σ, the edit distance d (a, b) is the minimum-weight series of edit operations that transforms a into b. Operations involved in edit distance algorithm are

- Insertion of a single symbol. If a = uv, then inserting the symbol x produces uxv. This can also be denoted ε→x, using ε to denote the empty string.
- Deletion of a single symbol changes uxv to uv (x→ε).
- Replacing a single symbol x for a symbol y ≠ x changes uxv to uyv (x→y).

In Levenshtein's definition, each of these operations has unit cost, so the Levenshtein distance is equal to the minimum number of operations required to change a to b.

The cyberbullying words are given as training dataset. With the training dataset the preprocessed online social network conversation is tested for bullying word presence. Lavenshtein distance algorithm detects the cyberbully words present in the conversation and displays it.

3.6 Naïve Bayes Classifier

For cyberbully Classification, Naïve Bayes classifier is used. A Bayes classifier is a simple probabilistic classifier based on applying Bayes' theorem (from Bayesian statistics) with strong (naive) independence assumptions.

In simple terms, a naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. The decision rule used for classification is,

$$\text{classify}(f_1, \text{dots}, f_n) = \text{argmax}_c\, p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

Where,

- C (upper case) represents the document class and c (lower case) is one of the possible classes. In this case the possible classes are Flaming, Insult, Harassment, and Terrorism.
- Fi, dots, Fn called features and fi, dots, fn are the values of the corresponding features. Features are the words occurring in the document and their value is the number of occurrences.
- argmaxc is the argmax function which means something like “The class c with the highest value for the following function.”
- P(C = c) is the probability that a given bullying word belongs to class c. Let‟s assume this is the same for all classes and hence ignore it as it is just a constant.
- P (Fi = fi | C = c) is the probability that the word F_i occurs in a document of class c.

3. EXPERIMENTAL RESULTS

4.1 Dataset

For this work, the datasets described below for the experiment on cyberbullying detection are considered, which are available from the workshop on Content Analysis for the Web 2.0 [10]. The dataset contains data collected from two different social networks: Formspring.me and MySpace. Formspring.me is a discussion-based site, users broadcast their message from which about 500 posts is selected randomly. MySpace is a popular social networking website and 600 posts are randomly selected from it. Datasets were provided in the form of text document for each conversation set between two users.

4.2 Evaluation Parameters

Precision: The total number of correctly identified true bullying posts out of retrieved bullying posts.

$$\text{Precision} = \frac{|(\text{relevant cyberbully words}) \cap (\text{retrieved cyberbully words})|}{(\text{retrieved cyberbully words})}$$

Recall: Number of correctly identified bullying cases from total number of true bullying cases.

$$\text{Recall} = \frac{|(\text{relevant cyberbully words}) \cap (\text{retrieved cyberbuully words}))}{(\text{relevant cyberbully words})}$$

F- measure: It is the harmonic mean of precision and recall.

$$\text{F} - \text{measure} = 2.\frac{\text{precision . recall}}{\text{precision} + \text{recall}}$$

The Precision, Recall and F- measure values obtained in the experimental results are shown in Table 1.

Table 1. Precision, Recall and F-measure values of Dataset used.

No. of cyberbully Posts	F-measure	
	Formspring.me	Myspace.com
100	79	77
200	83	84
300	71	74
400	94	92
500	87	89

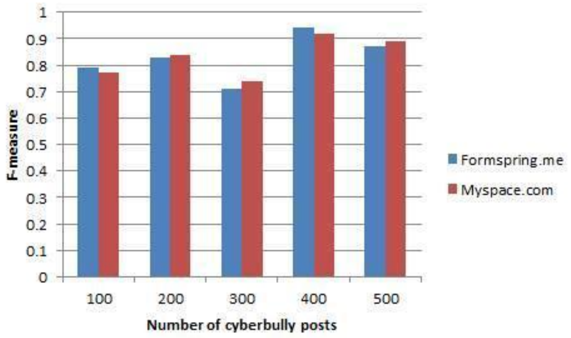


Figure 2. F–measure graph

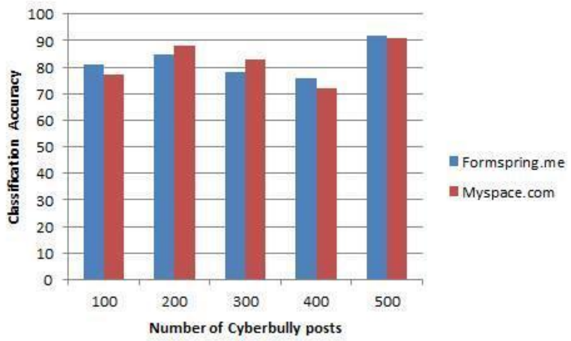


Figure 3. Classification accuracy graph

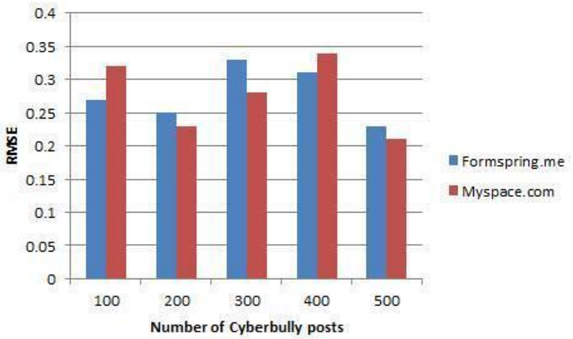
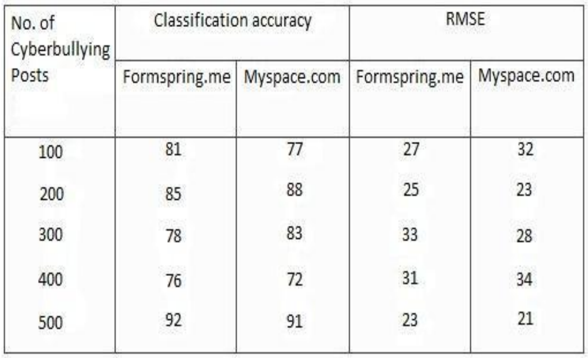


Figure 4. RMSE graph

Table 2. Classification accuracy and RMSE



No. of Cyberbullying Posts	Classification accuracy		RMSE	
	Formspring.me	Myspace.com	Formspring.me	Myspace.com
100	81	77	27	32
200	85	88	25	23
300	78	83	33	28
400	76	72	31	34
500	92	91	23	21

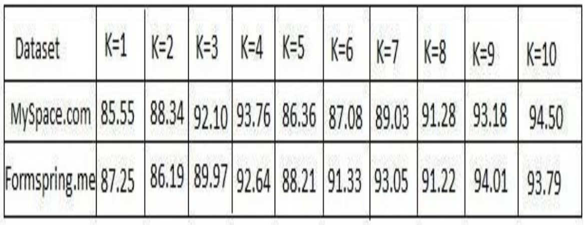
It is observed from Figure 2 that the proposed system shows best F- measure values.

Figure 3 showed the Classification Accuracy and Figure 4 shows Root Mean Squared Error [RMSE] graph obtained in FormSpring.me and Myspace.com datset.

4. DATA SET VALIDATION

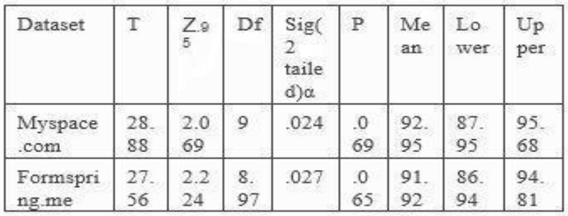
The data set was validated using 10 fold cross validation method. Split the data into 10 samples. Fit a model to the training for 9 samples and use the test 1 sample .Repeat the process for the next sample, until all samples have been used to either train or test the model. This table 2 shows the overall mean accuracy values for Formspring.me and Myspace.com dataset. A confidence interval is a range around a measurement that conveys how precise the measurement is .Table 3 shows the 10 fold cross validation performed on 95% confident Interval on Myspace.com and Formspring.me data set.

Table 3. Mean Accuracy values for different K values



Dataset	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=10
MySpace.com	85.55	88.34	92.10	93.76	86.36	87.08	89.03	91.28	93.18	94.50
Formspring.me	87.25	86.19	89.97	92.64	88.21	91.33	93.05	91.22	94.01	93.79

Table 4. 10 fold cross validation performed on convote dataset with 95% confidence interval



Dataset	T	$Z_{.95}$	Df	Sig(2 tailed)α	P	Mean	Lower	Upper
Myspace .com	28.88	2.069	9	.024	.069	92.95	87.95	95.68
Formspring.me	27.56	2.224	8.97	.027	.065	91.92	86.94	94.81

5. CONCLUSION

Cyber security is becoming an important concern with increase in the use of internet. A number of life threatening cyberbullying experiences among young people have been reported internationally, thus drawing attention to its negative impact. Though Government has framed laws and system for restricting cyberbully activities, taking action on the buller (person one who bullies other) is difficult since identifying the buller is critical. The proposed system focuses on detecting the presence of cyberbullying activity in social networks and to classify it using Lavenshtein algorithm and Naïve Bayes classifier, which helps government to take action before many users becoming a victim of cyberbullying.

6. REFERENCES

[1] B. Belsey. (6th June 2013). Cyberbullying.org. Available: http://www.cyberbullying.org/

[2] P. K. Smith, J. Mahdavi, M. Carvalho, S. Fisher, S. Russell, and N. Tippett, *"Cyberbullying: Its nature and impact in secondary school pupils,"* Journal of Child Psychology & Psychiatry, vol. 49, pp. 376-385, 2012.

[3] M. A. Campbell, *"Cyber bullying: An old problem in a new guise?"* Australian Journal of Guidance and Counselling, vol. 15, pp. 68-76, 2011.

[4] Sara Bastiaensens, Heidi Vandebosch, KarolienPoels, Katrien Van Cleemput , Ann DeSmet, Ilse De Bourdeaudhuij*," Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully,"* Elsevier, Computers in Human Behaviour 31 (2014) 259–271.

[5] Christina F. Brown, Michelle Kilpatrick Demaray, Stephanie M. Secord, *"Cyber victimization in middle school and relations to social emotional outcomes",* Elsevier, Computers in Human Behaviour 35 (2014) 12–21.

[6] Jin-Liang Wanga, Linda A. Jackson , James Gaskin , Hai-Zhen Wang, *"The effects of Social Networking Site (SNS) use on college student's friendship and well-being",* Elsevier, Computers in Human Behavior 37 (2014), pp.229–236.

[7] Victoria Lopez, Alberto Fernandez, Maria José del Jesus , Francisco Herrera, *"A hierarchical genetic fuzzy system based on genetic programming for addressing classification with highly imbalanced and borderline data-sets",* Elsevier, Knowledge-Based Systems 38 (2013), pp. 85–104.

[8] M. Dadvar, F. d. Jong, R. Ordelman, and D. Trieschnigg, *"Improved cyberbullying detection using gender information,"* In Proceedings of the Twelfth Dutch-Belgian Information Retrieval Workshop (DIR 2012), pp. 23-25, February 2012.

[9] K. Dinakar, R. Reichart, and H. Lieberman, *"Modeling the Detection of Textual Cyberbullying,"* in Proc. IEEE International Fifth International AAAI Conference on Weblogs and Social Media, Barcelona, Spain, 2011.

[10] K. Reynolds, A. Kontostathis, and L. Edwards, *"Using Machine Learning to Detect Cyberbullying,"* In Proceedings of the 2011 10thConference on Machine Learning and Applications Workshops, vol. 2, pp. 241-244, December 2011.

[11] McGhee, J. Bayzick, A. Kontostathis, L. Edwards, A. Mcbride, and E. Jakubowski, *"Learning to Identify Internet Sexual Predation, "*International Journal on Electronic Commerce 2011, vol.15, pp. 103-122, 2011.

[12] D. Yin, Z. Xue, L. Hong, B. D. Davison, A. Kontostathis, and L. Edwards, *"Detection of Harassment on Web 2.0,"* in Proc. Content Analysis of Web 2.0 Workshop, Madrid, Spain, 2009.

[13] Kontostathis, L. Edwards, and A. Leatherman, *"Chat Coder: Toward the Tracking and Categorization of Internet Predators,"* In Proceedings of Text Mining Workshop 2009 held in conjunction with the Ninth SIAM International Conference on Data Mining, 2009.

[14] Bsecure.Available: http://www.safesearchkids.com/BSecure.html.

[15] CyberPatrol.Available: http://www.cyberpatrol.com/cpparentalcontrols.asp

[16] eBlaster. Available: http://www.eblaster.com/.

[17] IamBigBrother. Available: http://www.iambigbrother.com/.

[18] Kid swatch. Available: http://www.kidswatch.com/.