Sentiment Analysis for Effective Detection of Cyber Bullying

Vinita Nahar[1], Sayan Unankard[1], Xue Li[1], and Chaoyi Pang[2]

[1] School of Information Technology and Electrical Engineering
The University of Queensland, Australia
[2] The Australian E-Health Research Center, CSIRO, Australia
{v.nahar,s.unankard}@uq.edu.au,xueli@itee.uq.edu.au,Chaoyi.Pang@csiro.au

Abstract. The rapid growth of social networking and gaming sites is associated with an increase of online bullying activities which, in the worst scenario, result in suicidal attempts by the victims. In this paper, we propose an effective technique to detect and rank the most influential persons (predators and victims). It simplifies the network communication problem through a proposed detection graph model. The experimental results indicate that this technique is highly accurate.

Keywords: Social Network, Cyber Bullying, Text Mining.

1 Introduction

With the proliferation of Web 2.0, cyber bullying is becoming an important issue. A number of life threatening cyber bullying experiences have been reported internationally, thus drawing attention to its negative impact [1]. Detection of online bullying and subsequent preventive measure is the main course of action to combat it.

To this end, we propose a detection method for identifying cyber bullying messages, predators and victims. Our methodology is divided into two phases. The first phase aims to accurately detect harmful messages. We present a new way of feature selection, namely common and sentiment features. The next phase aims to analyse social networks to identify predators and victims through their user interactions, and present the results in a graph model. A ranking algorithm is employed to detect the influential predators and victims. The proposed approach for anti-cyber bullying using a computed cyber bullying detection matrix and associated graphical representation of the results is unique.

Contributions. First, we propose a novel statistical detection approach, which efficiently identifies hidden bullying features to improve the performance of classifier. Second, we present a graph model to detect association between various users in the form of predators and victims. Besides identifying victims and predators, this graph model can be used to answer many important queries such as how many victims associate to the same predator. Furthermore, ranking methods are employed to identify the most influential persons on the social network. Third,

Q.Z. Sheng et al. (Eds.): APWeb 2012, LNCS 7235, pp. 767–774, 2012.

our experiments show that the proposed approach is statistically significant in terms of our test results.

The rest of this paper is organised as follows. Section 2 reviews the related literatures; Section 3 describes the proposed methodology; Section 4 explains how the experiments are designed and reports the results; Section 5 concludes this paper.

2 Related Work

Sentiment Analysis. It is a task of learning the semantic orientation of a document, sentence and phase. Sentences can also be classified as objective or subjective. Subjective sentences are ideal for sentiment classification. Extensive works have been done in this area [7]. However, the major application areas explored are product and movie reviews. Nevertheless, we employed sentiment analysis for bullying detection.

Cyber Bullying Detection. Current research to detect online sexual predators [5] correlates the theory of communication and text mining to discriminate predator and victim communication. It uses rule based approach applied on chat log dataset. Other interesting works in this area [9] and [8] mainly focused on the detection of harmful messages. [8] also performed affect analysis on inappropriate messages.

Ranking Methods. Page Rank [6] and Hyperlink-Induced Topic Search (HITS) [2] are ranking methods used extensively in many areas of network analysis and information retrieval to find appropriate hub and authority pages, where hub and authority pages are interlinked and effects each other. In both the methods, the subset of relevant web pages is constructed based on an input query, Page Rank proliferates search results through links from one to another web page to find the most authoritative page, whereas, HITS identifies authoritative as well as hub pages. Our approach uses the HITS to calculate potential predator and victim scores in the form of Eigen values and vectors.

3 Methodology

The proposed methodology is a hybrid approach. It employs sentiment analysis to classify given entry into 'bully' or 'non-bully' category and uses link analysis to find the most influential person. Each step is defined in detail as follows.

3.1 Feature Selection and Classification

The feature selection is a key step to represent data in a feature space as an input to the model. Social network's data are noisy, thus regressive preprocessing is employed. Also, the number of features grow with the number of documents. Thus, we propose following two types of feature selection methods:

A. Common Features. These are mixed features extracted from both bullying and non-bullying messages, which are generated using bag-of-word[1] approach.

B. Sentiment Features. These are generated by applying Probabilistic Latent Semantic Analysis (PLSA) [3] model on bullying posts only. Classification of text as positive, negative or neutral does not cater for the versatile nature of sentiment detection. To improve opinion identification, PLSA is used as a strong tool to deal with the hidden factors [4]. For our probabilistic approach to extract sentiment features, PLSA is applied on the messages described next to identify hidden bullying factors. Let B represent documents, which is a collection of messages $B = \{b_1,..., b_N\}$, with word $F = \{f_1,..., f_M\}$. Then the dataset can be defined as a matrix C of size N*M; $C = (n(b_i, f_j))_{ij}$, where $n(b_i, f_j)$ is the number of times f_j appears in a document b_i. Now we introduce unobserved data (latent variables) $Z = \{z_1,..., z_k\}$ within the documents. Latent variables deal with the hidden class in the documents. It is defined in the following steps: Say, $P(b)$ is a probability of a document b, $P(z|b)$ represents the probability distribution of the document over the latent class. The word generated with the probability $P(f|z)$ is conditional probability of the word given the latent class variable. The complete model in terms of word, topic and document can be defined as:

$$P(b, f) = P(b)P(f|b) \tag{1}$$

where, $P(f|b) = \sum_z P(f|z)P(z|b)$. By applying PLSA, which utilises iterative Expectation (E) and Maximization (M) steps to maximize the data likelihood, when it converges at the local optimal solution. Finally, Bayes method is applied to calculate posterior probability $P(z \rightarrow b)$ to generate feature and latent variable relationships. This information is utilized for the further classification task. These features show the stronger correlation between a word and an unobserved class; hence, a feature can be considered as a class.

Further, SVM[2] is learnt through both the features and tested for the classification of messages into one of two classes: bully or non-bully.

3.2 Victim and Predator Identification

The cyber conversation has appeared to a situation where the surge of many bullying messages toward a specific user. In a network, predators and victims are linked to each other via sent and received messages and identified by their usernames.

Scenario: We considered a subset (Figure 1) of the main network of the users. To identify the predator and victim, HITS algorithm is implemented by computing their respective scores. A predator can be identified by the highest predator score and victim by the highest victim score.
Assumption: Each user we considered is associated with at least one bullying message.

[1] http://en.wikipedia.org/wiki/Bag_of_words_model
[2] http://www.csie.ntu.edu.tw/ cjlin/libsvm

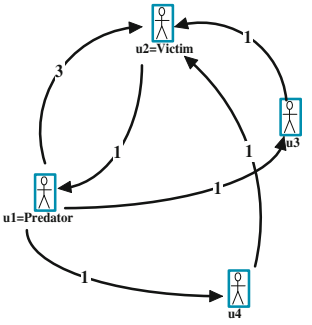

Fig. 1. Victim and Predator Identification graph

Predator: User sending out at least one bullying message.
Victim: A person who is receiving at least one bullying message.
Objective: Ranking on predators and victims to find the most influential person.
Graph model: We considered a subnet/subset (for example, four users) (Figure 1) of the main communication network of the users; which includes predators and victims. To identify the predator and associated victim, the Hyperlink-Induced Topic Search (HITS) algorithm is implemented in a victim and predator search by computing a predator and a victim score. The HITS method is based on the fact that the good hub pages point to good authority pages and good authority pages are linked by the good hub pages. The search query penetrates through web pages to identify potential hub and authority page based on the respective scores. Now, to find predator and victim, first we define following two equations:

$$p(u) \leftarrow \sum_{u \to y} v(y), v(u) \leftarrow \sum_{y \to u} p(y) \tag{2}$$

Where, $p(u)$ and $v(u)$ depicts the predator and victim scores respectively and $u \to y$ indicates the existence of the bullying massage from u to y and vice versa for $y \to u$. These two equations are an iteratively updating pair of equations for calculating predator and victim scores, which is based on our assumption defined above. In each iteration, scores are calculated through in degrees & out degrees and associated scores; this may results in large values. Thus to get the relative weights, each predator and victim scores are divided by the sum of all predator and victim scores respectively. Figure 1, delineates the identification of the predator and victim in a communication network. It is a weighted directed graph G = (U,A) where,

- Each node $u_i \in U$ is a user involved in the bullying conversation.
- Each arc $(u_i, u_j) \in$ A, is defined as a bullying message sent from u_i to u_j.
- The weight of arc (u_i, u_j), denoted as w_{ij}, is defined in the next section.

Victim and predator can be identified from the weighted directed graph G:

– The victim will be the nodes with many incoming arcs and the predator will be the nodes with many outgoing arcs. However, this paper attempts to find if a user is a potential victim or a potential predator.

3.3 Cyber Bullying Matrix

Now, to identifying predator and victim based on their respective scores, we formulate a cyber bullying matrix (w). Table 1, is a matrix w, which is a square adjacency matrix (which represents in degrees and out degrees of each node) of the subnet with entry w, which is a square adjacency matrix of the subset with entry w_{ij}, where,

$$w_{ij} = \begin{cases} n & \text{if there exist n bullying message from i to j} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Since each user will have victim as well as predator score, scores are represented as the vectors of n*1 dimension where i^{th} coordinate of vector represent both the scores of the i^{th} user, say p_i and v_i respectively. To calculate scores, equations $p(u)$ and $v(u)$ are simplified as the victim and predator updating matrix-vector multiplication equations. For the first iteration, p_i and v_i can initialize at 1. For each user (say, i = 1 to 4) predator and victim scores are as follows:

$$p(u_i) = w_{i1}v_1 + w_{i2}v_2 + ... + w_{i4}v_4, v(u_i) = w_{1i}p_1 + w_{2i}p_2 + ... + w_{4i}p_4 \tag{4}$$

These equations converge at a stable value to provide final predator and victim vector of each user. Finally, Eigenvectors are calculated to get both scores.

Table 1. Cyber Bullying Matrix (w)

	u_1	u_2	u_3	u_4	**...**
u_1	0	3	1	1	...
u_2	1	0	0	0	...
u_3	0	1	0	0	...
u_4	0	1	0	0	...
...	...	...	...	...	...

4 Experiments and Results

For our experiment, we used data available from the workshop on Content Analysis on the Web 2.0[3] and requested manually labelled data from Yin et al [9]. The dataset contains data collected from Kongregate, Slashdot and MySpace websites. Kongregate (online gaming site), provides data in the chat-log style. We have assumed that Kongregate players used aggressive words while playing.

[3] http://caw2.barcelonamedia.org/

Table 2. Classification performance based on common and sentiment features : MySpace and SlashDot datasets

	MySpace Dataset				SlashDot Dataset			
Features	**Cases**	**Bully**	**NonBully**	**Accuracy %**	**Cases**	**Bully**	**NonBully**	**Accuracy%**
	Sentiment features				**Sentiment features**			
500	740	15	725	97.97	2174	11	2163	99.49
1000	1034	28	1006	97.29	2868	18	2850	99.37
2000	1383	40	1343	97.11	3380	26	3354	99.23
4000	1657	50	1607	96.98	3808	34	3774	99.11
6000	1766	57	1709	96.77	3945	43	3902	98.91
14000	1934	65	1869	96.64	4067	48	4019	98.82
	Common feature				**Common feature**			
15632	1947	65	1882	96.61	4077	48	4029	98.85

Slashdot is a forum, where users broadcast messages and MySpace is a popular social network. Data from the Slashdot and MySpace were provided in the XML files, where each file represented a discussion thread containing multiple posts. Post and user information for each site were extracted and indexed through the inverted file index[4]; thus assigning an appropriate weight to each term.

4.1 Cyber Bullying Detection

LibSVM was applied for classification into bully or non-bully class using a linear kernel and 10 fold cross validation was performed. Using both the features, accuracy is reported as the evaluation criterion of three different datasets in Table 2 and 3. Table 2 defines classifier performance on Myspace and Slashdot while Table 3 shows classifier performance on Kongregate and Combined datasets.

4.2 Victim and Predator Identification

A predator and victim identification graph is developed for a given scenario. Only the messages identified as bullying are considered, as shown in Figure 1. The user information was extracted to examine victim and predator data in the format of the matrix described in Table 1. The rows depict message senders, and the columns delineate receivers. The matrix values indicate the number of messages sent and received. To analyse forum style data, we considered each user of a forum posting to be a sender and a receiver. However, we assumed that the users will not be sending message to themselves and hence assigned the message value as zero. The chat style dataset contains direct communication between two users, so there were one sender and one receiver for each message. Expert judgment is obtained by manually counting the number of sent bullying messages and ranking them. The performance of the HITS Algorithm is shown in Table 4. It shows that the HITS algorithm vs. expert judgment achieved similar predator identification accuracy, especially in regards to the top three ranked predators. Similarly, the HITS algorithm vs. expert judgment for victim identification results were also comparable. The datasets had many participants

[4] http://en.wikipedia.org/wiki/Inverted_index

and it was not specified whether messages were sent to anyone or not. Because of this, we assumed that all users who posted on the same topic by default were a receiver too. Hence, many users received the same victim ranking, which cannot be presented in a table form. Nevertheless, the results of HITS algorithm were closely aligned to the expert judgment results.

Table 3. Classification performance based on common and sentiment features: Kongregate and Combined datasets

	Kongregate Dataset				**Combined Datasets**			
Features	**Cases**	**Bully**	**NonBully**	**Accuracy %**	**Cases**	**Bully**	**NonBully**	**Accuracy%**
	Sentiment features				**Sentiment features**			
500	356	0	356	100.00	3240	26	3214	99.20
1000	575	2	573	99.65	4477	48	4429	98.69
2000	979	9	970	98.98	5742	75	5667	98.68
4000	1772	15	1757	99.10	7237	99	7138	98.63
6000	2276	23	2253	99.08	7987	123	7864	98.49
14000	3995	42	3953	99.22	9996	155	9841	98.54
	Common feature				**Common feature**			
15632	4292	42	4250	99.25	10316	155	10161	98.55

Table 4. Performance of Ranking Method : Predators

Expert Judgement		**Hits Algorithm**	
Rank	**User id**	**Rank**	**User id**
1	Android457	1	Android457
2	AmberPeace	2	AmberPeace
2	chance10149	2	chance10149
3	MS_143549239	3	YOUAREAHOMO
3	MS_226698859	3	mock03
4	mock03	4	im69
4	MS_3449431	4	Komedia
4	YOUAREAHOMO	4	LordShadow

4.3 Discussion

The performance of the classifier is evaluated on its accuracy based on both the features, as tested on the three different datasets. Accuracy is defined as the degree of measure by which classifier accurately identifies the true value. As shown in the graph (Figure 2), the results indicate a very high accuracy of the classifier with sentiment features, especially when the features are in the range of 500 to 4000, while it remains almost constant between 6000 to 14000. The classifier performs best with 500 features. We identified the top ranked 764 features with the best accuracy results as selected by the PLSA. We selected the top 500 features and calculated the accuracy for each case. We found that the F-1 measure is not significant for our evaluation, because of the large number of negative cases. We chose feature selection methods that optimised the accuracy of the classifier and achieved very high accuracy outcomes. We proposed a weighted directed graph model to critically analyse and answer user queries regarding victims and predators. Based on the weighted arcs between two users, the model iteratively assigns the victim and the predator score to each user and correctly identifies the most influential predator and victim.

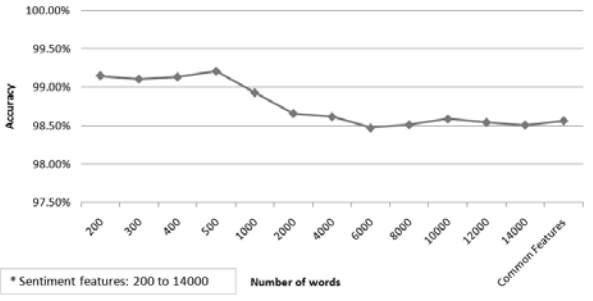

Fig. 2. Overall Performance based on Accuracy

5 Conclusions

We presented an effective sentiment analysis technique in this paper to detect cyber bullying messages by using PLSA for feature selection. The experimental results show that feature selection improves the accuracy of classifier, thus reducing resource usage significantly. In addition, we employs the HITS algorithm to calculate scores and rank the most influential persons (predators or victims). The proposed graph based model can be used to answer various queries about the user in terms of bullying. In future research, we plan to continue the in-depth analysis of indirect bullying and its emerging patterns, to help in its identification and prevention of cyber bullying.

References

1. ncpc.org, `http://www.ncpc.org/cyberbullying` (accessed September 15, 2011)
2. Easley, D., Kleinberg, J.: Link analysis using hubs and authorities. In: Networks, Crowds, and Markets: Reasoning about a Highly Connected World, pp. 399–405
3. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42, 177–196 (2001)
4. Liu, Y., Huang, X., Aijun, A., Yu, X.: Arsa: a sentiment-aware model for predicting sales performance using blogs. In: Research and Development in Information Retrieval, SIGIR, pp. 607–614 (July 2007)
5. McGhee, I., Bayzick, J., Kontostathis, A., Edwards, L., McBride, A., Jakubowski, E.: Learning to Identify Internet Sexual Predation. International Journal on Electronic Commerce 15(3), 103–122 (2011)
6. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999)
7. Pang, B., Lee, L.: A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In: Proc of the ACL, pp. 271–278 (2004)
8. Ptaszynski, M., Dybala, P., Matsuba, T., Masui, F., Rzepka, R., Araki, K.: Machine Learning and Affect Analysis Against Cyber-Bullying. In: Proceedings of the Thirty Sixth Annual Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2010), March 29- April 1, pp. 7–16 (2010)
9. Yin, D., Xue, Z., Hong, L., Davisoni, B.D., Kontostathis, A., Edwards, L.: Detection of harassment on web 2.0. In: Content Analysis in the WEB 2.0 (CAW2.0) Workshop at WWW 2009 (April 2009)