An Effective Approach for Cyberbullying Detection

Vinita Nahar[1], Xue Li[2], Chaoyi Pang[3]

[1, 2]School of Information Technology and Electrical Engineering,
the University of Queensland, Brisbane, Queensland 4072, Australia

[3]The Australian E-Health Research Center, CSIRO, Brisbane, Queensland 4029, Australia

[1]v.nahar@uq.edu.au; [2]xueli@itee.uq.edu.au; [3]Chaoyi.Pang@csiro.au

Abstract-The rapid growth of social networking is supplementing the progression of cyberbullying activities. Most of the individuals involved in these activities belong to the younger generations, especially teenagers, who in the worst scenario are at more risk of suicidal attempts. We propose an effective approach to detect cyberbullying messages from social media through a weighting scheme of feature selection. We present a graph model to extract the cyberbullying network, which is used to identify the most active cyberbullying predators and victims through ranking algorithms. The experiments show effectiveness of our approach.

Keywords- Social Networks; Cyberbullying; Text-Mining; Link Analysis

I. INTRODUCTION

With the proliferation of the Internet, cyber security is becoming an important concern. While Web 2.0 provides easy, interactive, anytime and anywhere access to the online communities, it also provides an avenue for cybercrimes like cyberbullying. A number of life threatening cyberbullying experiences among young people have been reported internationally, thus drawing attention to its negative impact. In the USA, the problem of cyberbullying has become increasingly evident and it has officially been identified as a social threat [1]. There is an urgent need to study cyberbullying in terms of its detection, prevention and mitigation.Traditional bullying is any activity by a person or a group aimed at a target group or individual involving repeated emotional, physical or verbal abuse. 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 behaviour by an individual or group, in order to harm others [2]. Cyberbullying is defined as an aggressive, intentional 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 [3].

Recent research has shown that most teenagers experience cyberbullying during their online activities including mobile phone usage [4], 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 victimised by cyberbullying [5]. The implications of cyberbullying [4] become serious (suicidal attempts) when the victims fail to cope with emotional strain from abusive, threatening, humiliating and aggressive messages [6]. The impact of cyberbullying is exasperated by the fact that children are reluctant to share their predicament with adults (parents/teachers), driven by the fear of losing their mobile phone and/or Internet access privileges [7]. The challenges in fighting cyberbullying include: detecting online bullying when it occurs; reporting it to law enforcement agencies, Internet service providers and others (for the purpose of prevention, education and awareness); and identifying predators and their victims.

In literature, cyberbullying has been studied extensively for understanding its various attributes and its prevalence from the social perspective. Prevention measures include:

- human interference,
- deletion of offensive terms,
- black list or scoring on the authors' cyber performance,
- educational awareness.

However, very little attention has been focussed on its online detection. Detection of cyberbullying and the provision of subsequent preventive measures are the main courses of action in combating cyberbullying. The aim of this paper is to propose an effective method to detect cyberbullying activities on social media. Our detection method can identify cyberbullying messages, predators and victims.

The proposed method is divided into two phases. The first phase aims to accurately detect harmful messages. We present a new way of feature selection, namely semantic and weighted features. Semantic features uses the Latent Dirichlet Allocation (LDA) [8] algorithm to extract latent features, whereas weighted features are the bullying-like features which include static badword vocabulary. LDA is a topic-modelling approach; it uses probabilistic sampling strategies to describe how words are generated within a document based on the latent topics. It finds the most suitable set of latent variables. Given a set of messages, we use LDA to understand the semantic nature of word usage to detect underlying topics ('bullying' and 'non-bullying') automatically. Words generated under both topics are extracted and ranked. Then, training samples are generated on

The second phase aims to analyse social networks to identify predators and victims through their user interactions, and to present the results in a graph model. A social network, in the form of connected graphs (nodes and edges) allows users (nodes) to be connected with other users (nodes) directly or indirectly. In an online community, users may publicize anything. The structure of the Web enables publicized material to penetrate the worldwide community quickly. The publicized material against the target user can be accessed repeatedly, over time, by the target as well as any number of other users. Moreover, those who post hurtful messages can be anonymous. The anonymity of the bully might create a sense of fear within the target user as they are dealing with an unknown identity. This handicaps the target user while at the same time advantages the anonymous bully with a feeling of power and control over the target victim [9].

In this study we find predators and victims by determining the most active users in the form of the most active predators and victims. A user can be a predator or a victim depending upon the number of bullying messages he/she posts or receives in the online community. The number of bullying messages a user posts and receives also incorporates the number of users in his/her network because a post published by a user can be read by all the users in the network. A user would be the most active predator if he/she publishes many bullying posts and the posted messages are received by many users. A user is the most active victim if he/she receives many bullying messages from many other active predators. To find the most active predators and victims, the users in the network need to be ranked. Thus, finding the most active predators and victims becomes a ranking exercise. We employ a ranking algorithm to detect the most active predators and victims. The proposed approach constructs a cyberbullying detection matrix and an associated graph representation.

Although cyberbullying is a common problem in online communities, the offensive material is not labelled or classified as such in any form, which makes the investigation of cyberbullying challenging. For our experiments, we collate three different datasets from the Workshop on Content Analysis for the Web 2.0 [10] and obtain the manually-labelled data from [11] as the ground truth for the evaluation. Through our methodology, we show improved cyberbullying detection results using the semantic and weighted features selection. Our work is a unique approach that deals with cyberbullying predators and victims as an identification graph using a ranking algorithm and a computed cyberbullying detection matrix.

We make three contributions in this paper. Firstly, we propose a novel statistical detection approach, which is based on the weighted TFIDF scheme on bullying-like features. It also efficiently identifies latent bullying features to improve the performance of the classifier. Secondly, we present a graph model to detect the most active predators and victims in social networks. Besides identifying predators and victims, this graph model can be used to classify users in terms of levels of cyberbullying victimization, based on their involvement in cyberbullying activities. Thirdly, our experiments demonstrate that the proposed approach is effective and efficient.

The rest of this paper is organised as follows. In Section 2, a literature review on cyberbullying detection is presented. Section 3 explains the proposed methodology. Section 4 describes how the experiments are performed. The results are discussed in this section. The conclusion and future work are presented in Section 5.

II. RELATED WORK

A. Cyberbullying Detection

In a recent study on cyberbullying detection [12], gender specific features were used and users were categorized into male and female groups. In other study [13], NUM and NORM features were devised by assigning a severity level to the badwords list (nosewaring.com). NUM is a count and NORM is a normalization of the badwords respectively. The dataset consisted of 3,915 posted messages crawled from the Web Site, Formspring.me. They employed replication of positive examples up to ten times and accuracy on the range of classifiers was reported. Their findings showed that the C4.5 decision tree and an instance-based learner were able to identify the true positives with 78.5% accuracy. Dinakar et al. in [14] considered detecting cyberbullying in the form of sexuality, race, intelligence and profanity label-specific comments. On 4,500 manually labelled YouTube comments,the accuracy was reported for binary and multiclass classifiers such as, naive Bayes, JRIP, J48 and SMO. Their results indicated that binary label-specific classifiers outperformed multiclass classifiers with 66.7% accuracy. Other interesting works [11] in this area performed harassment detection from forum and chat room 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. Contextual features are based on the similarity measure between posts, with the intuition that the posts which are dramatically different from their neighbors are more likely to be harassing posts. Research on online sexual predators detection [15], [16] associate the theory of communication and text-mining methods to differentiate between predator and victim conversations, as applied to one-to-one communication such as in a chat-log dataset. The formal method is based on the keywords search while the latter uses a rule-based approach.

There are also some software products available for fighting against cyberbullying e.g., [17], [18], [19], [20], [21]. However, filters generally work with a simple key word search and are unable to understand the semantic meaning of the text. While some filters block the webpage containing the keyword, some shred the actual offensive words themselves. Other software products exhibit a blank page on detection of the keywords. However, removal of the offensive word from the

sentence can totally distort the meaning and sense of the sentence. Moreover, Internet programmers can easily dodge filters. It can be argued that filters are not an effective anti-cyberbullying solution as there are many ways to express inappropriate, illegal and offensive information. Using chat rooms, mobile communication and peer-to-peer networking the blocked content can bypass central servers that maintain filters [9]. Another limitation is that filtering methods have to be set up and maintained manually.

B. Ranking Methods

PageRank [22] and Hyperlink-Induced Topic Search (HITS) [23] 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 affect each other. In both methods, the subset of relevant web pages is constructed, based on an input query. PageRank proliferates the 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 HITS to calculate potential predator and victim scores in the form of Eigen values and vectors.

III. METHODOLOGY

The proposed methodology is a twofold hybrid approach. It employs the classification of a given entry into a 'bullying' or 'non-bullying' category and then uses link analysis to find the most active users in terms of predators and victims. Each step is defined in detail as follows.

A. Classification Model for Harmful Posts Detection

The feature selection is an important phase in representing data in feature space to the classifiers. Social network data are noisy, thus pre-processing has been applied to improve the quality of the research data and subsequent analytical steps; this includes converting uppercase letters to lower case, stemming, removing stop words, extra characters and hyperlinks. Moreover, sparsity in feature space increases with the number of documents. Nevertheless, we propose the following types of features generated through the LDA [8] topic model and weighted TFIDF scheme.

At the first step, we apply semantic features for the detection of harassing, abusive and insulting posts. In harassment detection [11] the appearances of pronouns in the harassing post were illustrated. Similarly in this work we used three types of feature sets: i. all second person pronouns 'you', 'yourself', etc. are counted as one term; ii. all other remaining pronouns'he', 'she', etc., are considered together as another feature; iii. foul words such as 'f**k', 'bullshit', 'stupid', etc., which make the post cruel are grouped in another set of features. The list of bad words is available from *noswearing.com*

The intuition behind combining these features is that it will improve the effectiveness of the classification of bullying posts from non-bullying posts. Thus, we employed a weighted TFIDF scheme. We scaled the bad words counted by a factor of two, because we wanted to improve inductiveness in the bad posts. Also in this work, we used the LDA [8] model to generate features and a range of top features was selected and compared to improve the classification result as illustrared in the experiments.

B. Predator and Victim Identification

The cyber conversation has developed to the situation where there has been a surge of numerous bullying messages toward a specific user. In a network, predators and victims are linked to each other by means of posts and are identified by their usernames. In this paper, we extend the graph-model proposed in our previous paper [24] towards the identification of the most active predators and victims. We aim to find the most active person on the social networks and gaming sites in terms of cyberbullying. A user is the most active cyberbullying predator if he/she sends messages that contain harmful words and such messages are received by many other users with high victim scores. A user is the most active victim if he/she receives many cyberbullying messages from many other users with high predator scores. Predator and victim scores are calculated based on the number of hurtful messages posted and received respectively. To find the most active (e.g. top eight) cyberbullying persons, the users in the network are ranked in terms of their predator and victim scores.

Scenario: We considered a subset (Figure 1) of the main network of users. A subset is a cyberbullying network that is extracted from the main network. It contains only bullying posts and associated users. To identify predators and victims, the HITS algorithm is incorporated by computing their respective scores. A predator can be identified by the highest predator score and a victim by the highest victim score.

1) Graph Model:

We considered a communication network of the users, which includes predators and victims. We used Gephi [25], a graphical interface to visualise a user's connectivity based on the bullying posts in a network. Figure 1 depicts the bullying network; it is a user group extracted based on the bullying posts by applying modularity theorem [26, 27], to measure the strength of partition of a network into subgraphs or communities. For a given graph, modularity is defined as the summation of the weight of all the edges that fall within the given subgroups minus the expected fraction if edges were distributed at random.

As show in Figure 1, using modularity algorithm, nine groups or communities, depicted by different colours are formed by considering users that are densely linked within the community as compared to between community [28]. Density of messages is computed for each post, which represents the badness embedded within the post [13]. Density of a post is calculated as the total count of the badwords within the post divided by the total number of the words in the post. To identify the predators and associated victims, the Hyperlink-Induced Topic Search (HITS) algorithm is used in a victim and predator search to compute a predator's and a victim's scores. The reasoning behind the HITS method is that in a network, 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 pages based on the respective scores. Similarly, we utilised this concept to rank predators and victims in a communication network.

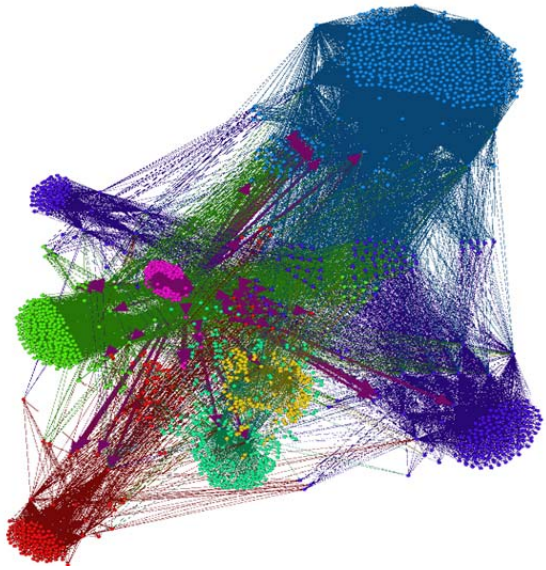

Fig. 1 Bullying network - User groups are formed based on the densely interconnected links (bullying post)

Assumption: Each user we considered is associated with at least one bullying message.

Predator: User posting at least one bullying message.

Victim: Person who is receiving at least one bullying message.

Objective: Ranking on predators and victims to find the most active person.

Now, to find predators and victims, we formally present a ranking module using the HITS method. A user can be a predator and a victim based on the messages he/she sends or receives. Therefore a user will be assigned a predator as well as a victim score. Following are two equations to compute predator and victim scores respectively:

$$p(u) \leftarrow \sum_{u \to y} v(y) \tag{1}$$

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

where $p(u)$ and $v(u)$ depict the predator and victim scores respectively. $u \to y$ indicates the existence of the bullying message from u to y, whereas, $y \to u$ indicates the existence of the bullying message from y to u. These two equations are an iteratively updating pair of equations for calculating predator and victim scores. They are based on our assumption that the most active predator links to the most active victims by sending bullying messages, and that the most active victim is linked by the most active predators by receiving bullying messages. In essence, if the user (u) is linked with another user with a high victim score, the user's predator score increases, and if the user (u) is linked through received messages to a user with a high predator score, the user's victim score increases. In each iteration, scores are calculated through indegrees and outdegrees, and associated scores; this may result in large values. Thus scores are normalized to unit length, i.e., each predator and victim scores is divided by the sum of all predator and victim scores respectively.

Now, we define to rank predators and victims in a bullying network shown in Figure 1. For simplicity and to explain a real scenario, we select only five users as shown in Figure 2 as an example. Figure 2 delineates the identification of the most active predators and victims in a bullying network. It is a weighted directed graph $G = (U, A)$ with $|U|$ nodes and $|A|$ arcs 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 as a summation of indegrees. It is discussed in detail in the next section.

Predators and victims 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. This paper attempts to find if a user is the most active user (a predator or a victim).

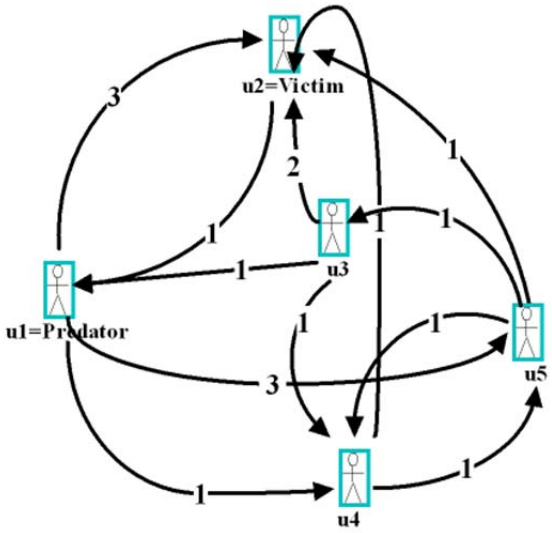

Fig. 2 Predator and victim identification graph

2) Cyberbullying Matrix:

To identify a predator and victim based on their respective scores, we formulate a cyberbullying matrix (w). Table I, is a matrix w, which is a square adjacency matrix (which represents indegrees and outdegrees of each node) of the subnet with entry w, which is a square adjacency matrix of the subset with entry w_{ij}, where,

$$w_{ij} = \begin{Bmatrix} n\ if\ \ there\ exist\ n\ bullying\ \ posts\ \ from\ u_i\ to\ u_j \\ 0\ otherwise \end{Bmatrix} \tag{3}$$

Since each user will have a victim as well as a predator score, scores are represented as the vectors of $n * 1$ dimension where i^{th} coordinate of the 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 are initialized at 1. For each user (say, i = 1 to N) predator and victim scores are as follows:

$$p(u_i) = w_{i1}v_1 + w_{i2}v_2 + \ldots + w_{iN}v_N \tag{4}$$

$$v(u_i) = w_{i1}p_1 + w_{i2}p_2 + \ldots + w_{iN}p_N \tag{5}$$

When these equations converge at a stable value (say k), it provides the final predator and victim vector of each user. Finally, we calculate the eigenvector to get the predator and victim scores.

TABLE I CYBERBULLYING MATRIX (W)

Receiver / Sender	U_1	U_2	U_3	U_4	U_5	….	U_N
U_1	0	3	0	1	3	….	….
U_2	1	0	0	0	0	….	….
U_3	1	2	0	1	0	….	….
U_4	0	1	0	0	1	….	….
U_5	0	1	1	1	0	….	….
….	….	….	….	….	….	….	….
U_N	….	….	….	….	….	….	….

Algorithm 1 gives a general framework of identification of the top ranked most active predators and victims. In the algorithm N is a total number of users and Top is a threshold value, which is set manually.

Algorithm 1: Predators and victims identification

Input: Set of users involved in the conversation with bullying post, N, Top

Output: Set of Top Victim and Top Predator

1: Extract senders and receivers from N;
2: Initialize predator and victim vector for each N;
3: Create adjacent matrix w using formula 3;
4: Calculate Predator and Victim vectors using iterative updating equations 4 and 5, and normalize, until converge at stable value k;
5: Calculate Eigen vectors to find Predator and Victim scores;
6: **Return** Top ranked Predators and Victims.

IV. EXPERIMENTS AND RESULTS

A. Dataset

For this work, we considered the datasets described below for the experiment on cyberbullying detection, which are available from the workshop on Content Analysis for the Web 2.0 [10] and we obtained the manually-labelled data from [11] as a ground truth dataset. The dataset contains data collected from three different social networks: Kongregate, Slashdot and MySpace. Kongregate is an online gaming site, which provides data in the chat-log style. Being gaming, the site players are likely to use aggressive words during their conversation. In Slashdot, a discussion-based site, users broadcast their message. MySpace is a popular social networking website. Datasets were provided in the form of XML files, where each file represented a discussion thread containing multiple posts. We extracted and indexed each post as one document. Each message is considered as one document and indexed through the inverted file index; thus assigning an appropriate weight to each term.

B. Cyberbullying Detection

LibSVM was applied to the two class classification problem using a linear kernel. Each post is an instance; positive classes contain bullying messages and negative classes contain non-bullying messages. Tenfold cross validation was performed in which the complete dataset was partitioned ten times into ten samples; in every round, nine sections were used for training and the remaining section was used for testing.

In our previous paper we chose the feature selection method which optimised the accuracy of the classifier as a performance measure [24]. The F-1 measure was not considered because of the large number of negative cases. Identifying bullying is a very critical issue because of false positive and false negative cases. Identifying non-bullying instance as bullying itself is a sensitive issue (false positive); on the other hand, system should not bypass bullying post as normal post (false negative). Therefore, false positive and false negative both are critical. Thus, precision,recall and F-1 measures are considerd for the performance evaluation metric. Also, we report identified false positive and false negative cases by the classifier.

In literature various strategies are proposed under imbalance text classification—we used oversampling of minority cases to improve the training of classifiers. We compare feature selection, namely, weighted TFIDF and semantic features on three different datasets. In our previous paper, TFIDF was calculated based on only the badwords list and performance was reported at top ranked PLSA features. Because of overfitting of the negative cases only accuracy of the classifier was compared. In this work we compare top ranked LDA features with Weighted TFIDF features for classification of cyberbullying posts.

C. Discussion

The performance of the classifier was evaluated on precision, recall and F-1 measure based on the top ranked features generated through LDA method against the truth set, as tested on the combination of three different datasets. ***Precision:*** the total number of correctly identified true bullying posts out of retrieved bullying posts. ***Recall:*** number of correctly identified bullying cases from total number of true bullying cases. ***F-1 measure:*** is the equally weighted harmonic mean of precision and recall. As shown in Figure 3, results indicate a very high precision, recall and F-1 measure on Kongregate. However, precision fell down at the top 2000 features. In most of the cases, the classifier performed almost similar, that is between 80 to 100 percent. Best performance is observed at 20000 features. On Myspace dataset (Figure 4) recall is very high, close to 1. However, precision varies between 77 and 86 percent except at feature value 18000 when it reaches 90 percent. Unlike Kongregate and Myspace datasets, performance is very low on Slashdot. From Figure 5, we can observe that although recall is very high, precision and F-1 measures degraded when feature sets were low. Also poor performance is observed at featurevalue 18000. Figure 6, depicts results on the combined dataset. Despite the fact that it is a combination of chat and discussion style community, performance at Weighted TFIDF is the best.

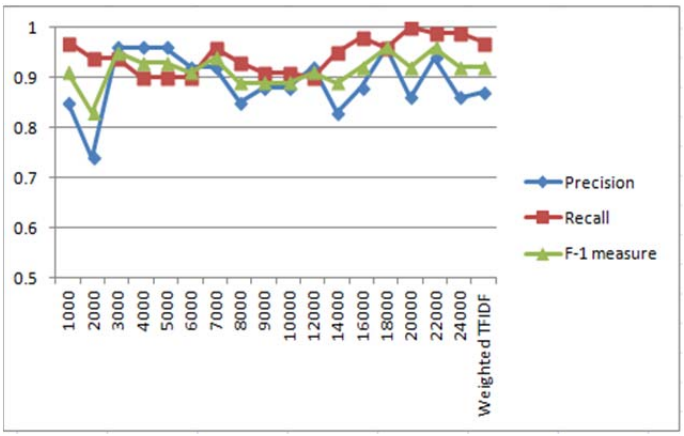

Fig. 3 Precision, Recall and F-1 measure on Kongregate

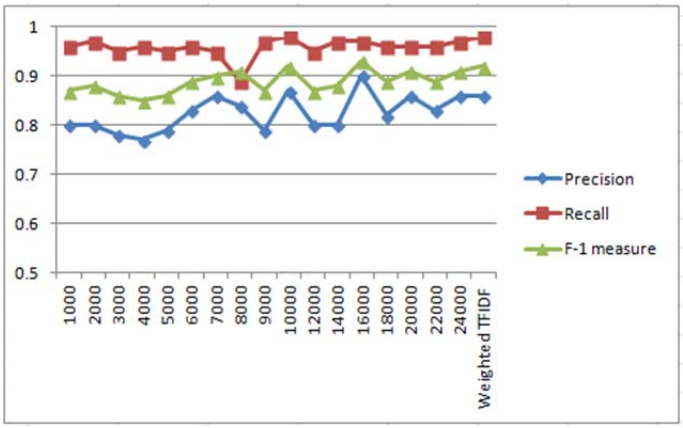

Fig. 4 Precision, Recall and F-1 measure on Myspace

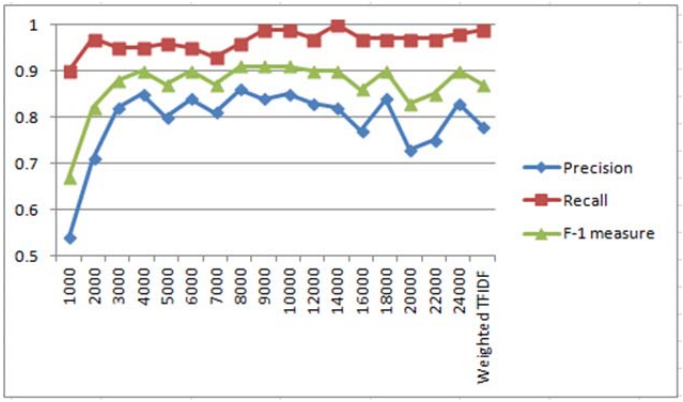

Fig. 5 Precision, Recall and F-1 measure on Slashdot dataset

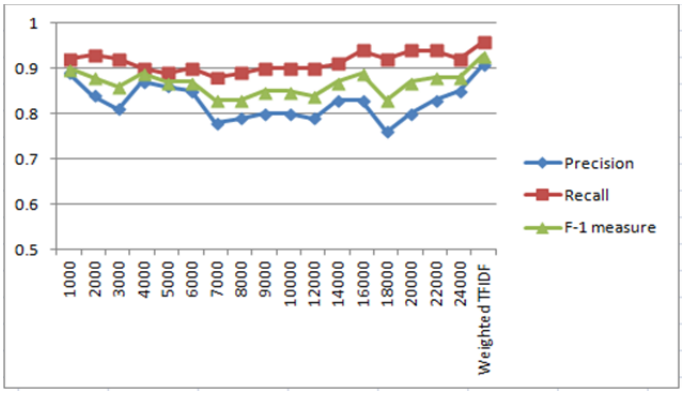

Fig. 6 Precision, Recall and F-1 measure on combined dataset

This work proposes a text classification model, which is helpful in identifiying suspecious: harmful and bullying-like posts from the online conversations. Therefore, it is significant to focus on recall [13], because it is important to reduce mislabelling bullying-like posts as normal posts i.e. reducing false negative. Thus, Table II compares false positive and false negative based on weighted features because of its better performace, on individual and combine datasets. Though dataset is imbalanced, reasonable performance was obtained on individual datasets. False negative cases are very low. It indicates that system is robust in identifying cyberbullying posts. However, number of false positive cases are still high, which is because of overfitting. Although oversampling of positive posts was adopted, high number of false positive result indicates more sophisticated learning methods need to be devised, which are able to deal with a few positive trainings. This is because in the real world problem, it is almost impossible to get a sufficient number of positive samples for training. Techniques like oversampling are subject to offline training.

TABLE II FALSE POSITIVE AND FALSE NEGATIVE CASES USING WEIGHTED TFIDF

Dataset	Actual positive	Actual negative	False positive	False negative
Kongregate	75	4339	11	2
Myspace	114	1828	18	2
Slashdot	76	4046	21	1
Combined	294	10184	28	12

D. Comparsion of Weighted TFIDF with Baseline Method

We compare the proposed weighted TFIDF method with the work done in [11] for harassment detection using TFIDF, sentiment and contextual features on three different datasets. The proposed feature selection method using weighted TFIDF, provided better performance as shown in Table III.

TABLE III COMPARISON OF PROPOSED APPROACH WITH BASELINE METHOD ON THREE DATASETS

		Kongregate	Slashdot	Myspace
Baseline	Precision	0.35	0.32	0.42
	Recall	0.60	0.28	0.25
	F-1 measure	0.44	0.30	0.31
Weighted TFIDF	Precision	0.87	0.78	0.86
	Recall	0.97	0.99	0.98
	F-1 measure	0.92	0.87	0.92

E. Victim and Predator Identification

Being a cyberbullying victim entails; being subjected to personal feelings. It is when a cyberbullying target is unable to defend oneself. Therefore, in identifying cyberbullying predators and victims we determine the most active predators and the most attacked users' 'victims' through the sent and received bullying messages, and the density of the badness of the message.

A predators' and victims' identification graph is developed for a given scenario. Only the posts identified as bullying were considered, as shown in Figure 1. In the experiments, each user is indexed and a userID is generated, which represents a node. Thus the username is represented by a user ID. The user information was extracted to analyse predators' and victims' data in the matrix form as depicted in Table I. The rows indicate message senders and the columns outline receivers of the post. The matrix values are the summation of bullying messages posted and received. To examine the data content from the forum-based website, we considered every user involved in a topic discussion as both a sender and a receiver of the post. However, we assumed that the individuals will not be posting messages to themselves. Therefore we excluded the self-loop and hence assigned the post value as zero. However, in future work, similarity measures between two posts will be considered to find the reply (or a receiver) of a particular post. The chatlog dataset consists of direct conversations between two users, so for every message there was only a sender and a receiver. In our previous paper, we have compared the identified top ranked predators and victims against expert judgement. However, in both cases, density of the post was not considered. In this paper, we identified the most active predators and victims, and compare the involvement of users in a bullying relationship as shown in the Table IV. Table IV shows that in some cases there are more than one user at the same rank. Therefore, users with the same rank are grouped together. We also noted that predators flagged at Rank I are also identified as a victim at Rank II. Similarly Rank II predators are Rank VII victims too.

TABLE IV PERFORMANCE OF GRAPH MODEL: PREDATORS AND VICTIMS IDENTIFICATION

	No of users	
Rank	**Predators**	**Victims**
I	4	8
II	2	4
III	1	7
IV	1	2
V	2	2
VI	7	1
VII	3	9
VIII	2	8

F. Discussion

We proposed a cyberbullying network, which is a weighted directed graph model. This graph model can be used to critically analyse and answer user queries regarding predators and victims. Based on the weighted arcs between two users, the model iteratively computes the predator and victim scores for each user, and accurately identifies the most active predator and its target. From Table IV, we observed that some of the users identified as predators are also identified as victims, with different ranks. This shows the involvement of a user in bullying activities as a predator and a victim. There could be several reasons for this. For example, suppose a user is involved in a discussion on a topic and that discussion may lead to an aggressive discussion, where users in a discussion thread started using aggressive language. Another reason could be that a receiver of the bullying message replied through a bullying message. The strategy of finding most active predators and victims can be adopted to classify users in various categories of victimization based on the predator and victim ranking of a user, for example, severe, moderate and normal bullying cases. The severe category could be the case when a user ranked high as a victim is not ranked (or ranked lower than threshold) as a predator. Thus it can be argued that the victim is unable to defend himself. Accordingly, victims identified at the Rank II may not be considered as victims because they are also the top ranked predators, which shows that these victims were able to defend themselves, hence cannot be considered to be victims. Therefore this case can be discarded for further investigation. Moreover, human interference can be employed; for example, consultation with social scientists to examine cases where users appear at a severe level.

V. CONCLUSION AND FUTURE WORK

In this paper we propose an approach for cyberbullying detection and the identification of the most active predators and victims. To improve the classification performance we employ a weighted TFIDF function, in which bullying-like features are scaled by a factor of two. The overall results using weighted TFIDF outperformed other methods. This captures our idea to scale-up inductive words within the harmful posts. However, bullying-like feature sets are limited to a static set of keywords. Therefore, dynamic strategies are required to be implemented to find emerging harmful and abusive words from the streaming text. To improve classifier's training in the absence of a sufficient number of positive examples, oversampling of positive posts is used. Also, throughout our experiments, we note that comparatively better performance was observed for false negative compared to false positive cases in individual and combined datasets. This is because of the fewer positive cases available for classifier's training. Therefore advance methods, which are capable of dealing with a few training sets in automatic cyberbullying detection, and to reduce false positive and false negative cases need to be developed, In addition, we proposed a cyberbullying graph model to rank the most active users (predators or victims) in a network. The proposed graph model can be used to answer various queries regarding the bullying activity of a user. It can also be used to detect the level of cyberbullying victimization for decision making in further investigations. Our future research in cyberbullying detection will continue to reduce false cases and train classifiers with fewer positive examples. We also plan to continue the in-depth analysis of cyberbullying victimization and its emerging patterns in stream text, to help the detection and mitigation of the cyberbullying.

REFERENCE

[1] Cyberbullying. Available: http://en.wikipedia.org/wiki/Cyberbullying

[2] B. Belsey. (6th July 2011). cyberbullying.org. Available: http://www.cyberbullying.org/

[3] 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, 2008.

[4] M. A. Campbell, "Cyber bullying: An old problem in a new guise?," Australian Journal of Guidance and Counselling, vol. 15, pp. 68-76, 2005.

[5] NCPC.org. Cyberbullying. Available: http://www.ncpc.org/cyberbullying

[6] NCPC.org. Stop Cyberbullying. Available: http://www.stopcyberbullying.org/what_is_cyberbullying_exactly.html

[7] NCH. (2005). Putting U in the picture - Mobile bullying survey 2005. Available: http://www.filemaker.co.uk/educationcentre/downloads/articles/Mobile_bullying_report.pdf

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993-1022, 2003.

[9] D. Butler, S. Kift, and M. Campbell, "Cyber Bullying In Schools and the Law: Is There an Effective Means of Addressing the Power Imbalance?," eLaw Journal: Murdoch University Electronic Journal of Law, vol. 16, 2009.

[10] CAW2. (April 2009, 10 November 2010). CAW 2.0 training datasets, in Fundacion Barcelona Media (FBM). Available: http://caw2.barcelonamedia.org/

[11] D. Yin, B. D. Davison, Z. Xue, L. Hong, A. Kontostathis, and L. Edwards, "Detection of Harassment on Web 2.0," In Proceedings of The Content Analysis In The Web 2.0 (CAW2.0) Workshop at WWW2009, 2009.

[12] 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.

[13] K. Reynolds, A. Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," In Proceedings of the 2011 10th International Conference on Machine Learning and Applications Workshops (ICMLA 2011), vol. 2, pp. 241-244, December 2011.

[14] K. Dinakar, R. Reichart, and H. Lieberman, "Modeling the Detection of Textual Cyberbullying," International Conference on Weblog and Social Media - Social Mobile Web Workshop, Barcelona, Spain 2011, 2011.

[15] A. Kontostathis, L. Edwards, and A. Leatherman, "ChatCoder: 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 (SDM 2009) 2009.

[16] I. 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.

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

[18] Cyber Patrol. Available: http://www.cyberpatrol.com/cpparentalcontrols.asp

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

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

[21] Kidswatch. Available: http://www.kidswatch.com/

[22] P. Lawrence, B. Sergey, M. Rajeev, and W. Terry, "The PageRank Citation Ranking: Bringing Order to the Web," Technical Report. Stanford InfoLab1999.

[23] D. Easley and J. Kleinberg, "Link analysis using hubs and authorities," in Networks, Crowds, and Markets: Reasoning about a Highly Connected World, ed, 2010, pp. 399–405.

[24] V. Nahar, S. Unankard, X. Li, and C. Pang, "Sentiment Analysis for Effective Detection of Cyber Bullying," In Proceedings of the 14th Asia-Pacific international conference on Web Technologies and Applications APWeb 2012, pp. 767-774, April 11-13 2012.

[25] M. Bastian, S. Heymann, and M. Jacomy, "Gephi: An Open Source Software for Exploring and Manipulating Networks," International AAAI Conference on Weblogs and Social Media, pp. 361-362, 2009.

[26] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," Physical Review E, vol. 69, 2004.

[27] M. E. J. Newman, "Modularity and community structure in networks," Proceedings National Academy of Sciences USA, vol. 103, pp. 8577–8696, 2006.

[28] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," Journal of Statistical Mechanics: Theory and Experiments, vol. P10008, pp. 1-12, 2008.