A Review of Cyber bullying Detection in Social Networking

Conference Paper · May 2017

CITATION
1

READS
1,491

2 authors, including:

Sonika Shrivastava
Maulana Azad National Institute of Technology, Bhopal
9 PUBLICATIONS **9** CITATIONS

SEE PROFILE

A Review of Cyberbullying Detection in Social Networking

Abstract- With the advancement of technology, craze of social networking platforms is proliferating. Online users now share their information with each other easily using computers, mobile phones etc. However, this has lead to the growth of cyber criminal acts for example, cyberbullying which has become a worldwide epidemic. Cyberbullying is the use of electronic communication to bully a person by sending harmful messages using social media, instant messaging or through digital messages. It has emerged out as a platform for insulting, humiliating a person which can affect the person either physically or emotionally and sometimes leading to suicidal attempts in the worst case. The main issue in preventing cyberbullying is detecting its occurrence so that an appropriate action can be taken at initial stages. To overcome this problem, many methods and techniques had been worked upon till now to control this problem. This paper is a survey covering cyberbullying and cyberbullying detection techniques. Next, we offer few suggestions for continued research in this area.

I. Introduction

Cyberbullying is a type of bullying that takes place using electronic technology including devices and equipments such as cell phones, computers through social media, text messages, chats etc [1].Examples of cyberbullying include mean text messages, rumors that can be very embarrassing to the concerned person. It can happen at any time and happens online and the text messages and images can be posted anonymously which can be distributed quickly to a very large audience.

The modern day technology is a boon and cannot be blamed for cyberbullying. Social media sites are used for positive activities, like connecting kids with friends and family, helping students with school, and for entertainment. But these tools can also be used to hurt other people. Whether done online or offline the effects of bullying are similar.

Cyberbullying is also defined as "willful and repeated harm inflicted through the medium of electronic text [2]".It mainly targets children and adolescents as they are most active on social networks. With Web 2.0 providing easy and ubiquitous online access, cyber security is becoming an important concern.

Some of the most common forms of cyberbullying are as follows [3]:

Flaming: Heated online arguments and fights using vulgar and abusive language.
Harassment: Repeatedly sending cruel, offensive or threatening messages.
Denigration: Exposing secrets of a person or gossips in order to damage reputation of a person.
Impersonation: Breaking into victim's account and sending mails.
Trickery: Tricking the victim into revealing sensitive information and passing it to others.
Interactive Gaming: Most gaming consoles allow people to connect and play online providing a chance to abuse using chats and comments.

Due to the lack of existing datasets, a very few studies have been done on detection of cyberbullying. At present whatever work that has been done in order to prevent cyberbullying are not accurate and reliable. In this paper we are going to review cyberbullying and the work that has been done to detect cyberbullying.

An article from The Times of India entitled "$188,776 Facebook grant for cyberbullying expert Sameer Hinduja" clearly states the increasing cyberbullying from the fact that Sameer Hinduja, an Indian-American and cyberbullying expert from Florida Atlantic University, has received a $ 188,776 grant from social networking site Facebook to study cyberbullying. The goal of the study is to study the nationwide prevalence and scope of cyberbullying [4].

According to Ipsos –a global market research company has found that 3 out of 10 parents in India say that their children have been victims of cyberbullying, majorly through Facebook and Orkut. The frequency of cyberbullying in India was found to be very high with 32% children having access to Internet or mobile phones.

An article from The Indian Express Alarming! 50% Indian youths have experienced cyberbullying found that most of the Indian parents don't find it important to talk to their children about online safety. Although there is age restriction on joining various social networking sites but 10-12 years teens report a very high access to these sites. The Global Youth Online Behavior Survey ranked India third in cyberbullying. Nishant (name changed), outshone his school seniors in basketball. But suddenly, he did not want to play and became withdrawn from the game. He finally confided in his father that his school seniors had been harassing him online, since they could not pull him down on the court. So his father went to a counsellor. Nishant's case is not an exception. A rising number of such cases are being reported, underlining the trend.

National Crime Prevention Council defines cyberbullying as sending text or images to hurt or embarrass another person by using Internet, mobile phones or other devices [5].
According to a research conducted by Symantec [6], only 25% of the parents were aware that their child was involved in cyberbullying incident. According to a survey [7], majority of cyberbullying is done through Facebook and around 55% of the youth exposed to cyberbullying committed suicide.

Cyberbullying incidents don't come as a surprise. School children confirm that cyberbullying is common. It is becoming the easiest way to get back at someone. A person can be knocked down in front of a large number of people online. Many cyberbullies think that bullying others online is funny. Cyberbullies may not realize the consequences for themselves of cyberbullying. The things teens post online now may reflect badly on them later in the future. Also, cyberbullies and their parents may face legal charges for cyberbullying. Teens may think that if they use a fake name they won't get caught, but there are many ways to track someone who is cyberbullying.

Cyberbullying can be very damaging to adolescents and teens. It can lead to anxiety, depression, and even suicide. Also, once things are circulated on the Internet, they may never disappear, resurfacing at later times to renew the pain of cyberbullying.

The diagram illustrates the percentage of people where they are bullied most [8]:

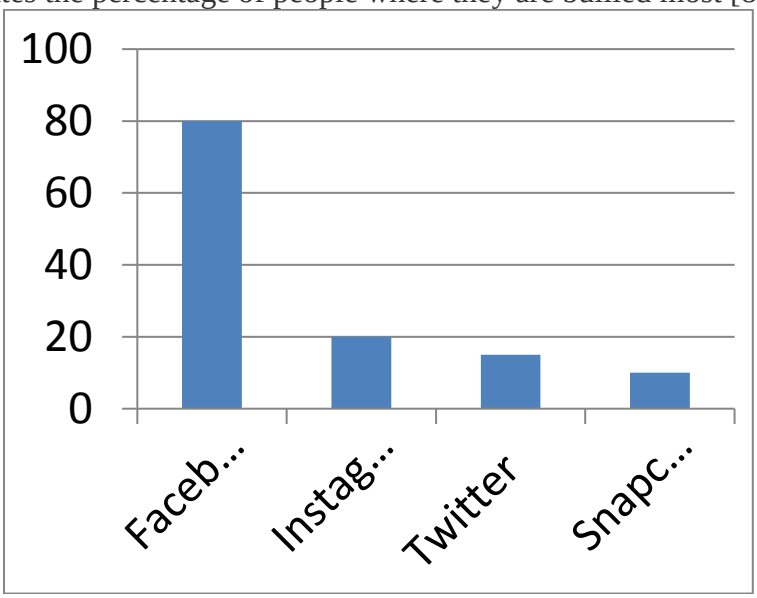

Figure 1. Percentage of people where they are bullied most

The motives that have been figured out behind cyberbullying are as follows: [9]

1. Revenge
2. Fear
3. Jealousy
4. Anger
5. Righteousness
6. Bigotry
7. To get attention of target or others

Youth between the ages of 12 and 17 spend more time with media than any other activity. One million children were harassed, threatened or subjected to cyberbullying on social media during the past year. According to a research study, eight out of ten teens using social media share personal information.

A misbehavior detection task was offered by the organizers of CAW 2.0 (Content Analysis in Web 2.0) in order to detect cyberbullying, but only one submission was received. In fact a very few research teams are working on the detection of cyberbullying. Cyberbullying is often very serious, including stalking and sometimes even death threat. Many incidents have been reported from all over the world through many new stories. Cyberbullying is very common among school going and college-going youth and since they are the assets of the society, a check on it is required to have a control over it. Sometimes the victim gets so depressed that they think of committing suicide or taking revenge on the attacker by cruel means. The victim may get physically harassed or may lose mental stability. So there is a need of detecting online harassment so that a necessary action can be taken against a cyberbullier before it becomes a serious offence. The Next section discusses the literature review of cyberbullying detection methods and techniques that have been adopted so far to counter cyberbullying.

II. Cyberbullying Detection

Various online tools, methods and techniques have been used so far in detecting cyberbullying. But only a very few techniques have provided accurate results in detection of cyberbullying. This section contains a review of techniques that have been adopted so far to detect this epidemic.

Using Computer Technology to address cyberbullying. The first method that was adopted to detect cyberbullying used computer technology by examining the incoming messages. Two approaches were used for this process [10].The first approach attached label of bullying to the suspected messages. For each message, users are able to vote positively or negatively on a message on the basis of a reputation score. For example, if a user X has messages m_1, m_2, m_3... m_n then each message has an overall message score denoted by r_1,r_2...r_n. To calculate user's reputation score, sum of all of his message scores is calculated and divided by the total number of messages. By calculating the average score, if the score comes out to be less than 1, then user is more likely to be malicious. The second approach used a filter mechanism to classify the messages as abusive or non-abusive. A filter can classify all the messages and discard insults and similar abusive messages. This filter is analogous to a spam filter. In this case, the objective is to exclude messages that can cause harm. But in a practical system, the filter will not be reliable. There will be false positives and false negatives.

In this research, a Cyber Bullying Reporting Platform (C.B.R.P.) is another critical step in the effort to combat cyberbullying by reporting instances of cyberbullying. It is a web portal where the victims or witnesses can report the incident. They are required to create a basic username on the website which would keep track of all the incidents related to that user. The incident gets submitted to the admin of the website and is being monitored by third party assistance for example a police department. Upon investigation, the account of the person reported will be blocked and appropriate legal action would be taken as per seriousness of matter.

Using online available applications and tools. eBlaster, Net Nanny, cloud9 and IamBigBrother are some available commercial tools available to protect children from Internet predators. They work on the basis of Packet sniffing. These tools scan all the outgoing and ingoing traffic in a network and then apply a filter. But the problem with these tools is that they are based on a simple keyword matching and hence its accuracy is questioned.

In order to overcome this limit, SafeChat was introduced. This software used the WinpCap library used by Wireshark. SafeChat supported Open System for Communication in Real time (OSCAR) protocol. But the documentation of this protocol failed as it was not compatible with other protocols.

Later SafeChat 2.0 was released as the new version of SafeChat. It is a third party plug-in for the Pidgin, an open source instant messaging system. It uses predator detection algorithms to classify chat participants as potential predators.

Using Psychological Perspective. Feinberg and Robey [11], worked on the psychological aspects of cyberbullying by preventing them using careful observation, monitoring, setting up school campaigns against cyberbullying as most of the affected individuals are teens, by hosting anti-bullying programs, counseling of individuals and by monitoring Internet traffic. Their paper mainly discusses the psychological aspects of preventing cyberbullying. Psychological perspective has given rise to the study of cyber profiling and so a criminological theory can be determined from it.

According to The Delete Cyberbullying Project [12], cyberbullying is best detected by simply observing the child. If a child's behavior changes, for example the kid stops using his/her cell phone or computer or any other communication device, if he/she gets upset after taking a call or receiving a text, it indicates that the child is being cyberbullied.

Using Semantic Analysis [3]. This technique overcomes the problem of Text Mining technique to detect cyberbullying. Two classes of data are used, one with the positive tone and other with the negative tone. These are then converted into vector form. The dataset used is of MySpace and used to train a supervised learning algorithm. The test data is classified into positive or negative based on the data that the supervised algorithm was trained with. The tool used is RapidMiner. The X-validation operator was used with a SVM (Support Vector Machine) with the following result:

Confidence (Positive)	Confidence (Negative)
0.615	0.385

Here, positive refers to presence of cyberbullying and negative refers to absence of cyberbullying data.

Using a Static dictionary of bad words. A computer software [13] known as Bully Tracer was designed to detect cyberbullying in a chat room conversation. This software has a dictionary of bad words like swear words and it matches the post to these bad words using rule-based algorithms to detect offensive text. This approach detected bullying content 85% of the time and innocent content 52% of the time. But such methods are not efficient as new swear words are created by bullies. Improvements to the existing tools are to be made to improve text mining and the bad word set needs to be updated on a daily basis.

Using Text Mining Approaches. Text Analytics play an important role in detecting the cyberbullying words. It involves applying data mining on text to extract useful text patterns by analyzing multiple word documents, social media data like comments and posts on social networks like Facebook, tweets on twitter, etc. The data that we gather have no specific structure and so called as unstructured data which are then refined using a multi-stage process.

The relevant documents are first collected to identify patterns in multiple documents. The data is then pre-processed which involves breaking up of a stream of text into tokens called as tokenization. Subsequently, cleaning up of the text, determination of the relationship of the

words with adjacent words to find their meaning [14, 15]. The next step deals with attribute generation where the text document is represented by words. Words and their occurrences are counted and a weight is assigned to each label using an in-built classifier. Then attribute removal is performed and data mining algorithms are applied to this data. A study [3] performed the above method to tackle cyberbullying on MySpace dataset to identify the occurrence of abusive words.

The text mining method is useful but it cannot detect bullying if it is done in non-curse words which when put together make up an offensive statement. For example, it will consider both "I hate you "and "I don't hate you" as offensive statements.

Using a Graph Model to detect and identify victim and perpetrator. An approach proposed by Nahar, Li and Pang [16], worked on detection of the victim and the perpetrator. The approach is divided into 2 phases. The first phase detects harmful messages by employing semantic and weighted features in the feature selection process using L.D.A. (Latent Dirichlet Allocation) algorithm [17]. It determines the word usage to identify the occurrence of bullying and weighted features provide a rough idea of severity of cyberbullying. In the second phase, the predators and victims are identified using HITS algorithm [18].The person sending the highest bullying content is considered as the bully and the person who receives at least one such message is considered as the victim and scores are assigned for ranking. The victim and the predator score is then calculated. This approach is based on the assumption that predator is a person who sends messages to multiple persons. A drawback of this technique is that a static bad-word set is used.

Using Semantic-Enhanced Marginalized Denoising Auto-Encoder [23]. In cyberbullying detection, one critical issue is robust and discriminative numerical representation learning of text messages. In this problem, a new representation learning method is developed via semantic extension of the popular deep learning model stacked denoising encoder. The semantic extension consists of semantic dropout noise and sparsity constraints, where the semantic dropout noise is designed based on domain knowledge and the word embedding technique. In addition, word embeddings have been used to automatically expand and refine a bullying word list that is initialized by domain knowledge. Experiments have been conducted over Twitter and MySpace which gave an accuracy of 70.53 % and 65.71 % respectively using different methods.

Dataset	Accuracy
Twitter	70.53%
MySpace	65.71%

Using Normative Agents to detect cyberbullying before it happens. It focuses on detecting cyberbullying before it happens unlike detecting after it happens. This approach employs normative agents which are physically present in the virtual network and support the victim against attacks [19]. The technique is based on BDI model [20] which detects the violation of predefined norms such as insulting or detection of bad words etc. These agents then take action

against this behavior. However, this method assumes that participants will learn to stick to the norms defined, which might not be a success in all cases.

Using individual topic sensitive classifiers. It has been found that using individual topic sensitive classifiers to detect cyberbullying is more effective after experimenting on 450,000 YouTube comments. Text classification comes into play by detecting sensitive topics. The comments are collected under a label based on sexuality, race and culture, mental capabilities of a person etc [21]. The datasets are subjected to binary and multi-class classifiers to detect comments referring to sensitive topics.

Using Machine learning to detect cyberbullying. Machine learning is the process of making a computer learn without being explicitly programmed. Through machine learning we can detect language patterns used by bullies and develop rules to automatically detect cyberbullying content. In a study that used machine learning to detect cyberbullying [6, 22], the dataset was downloaded from FormSpring.me, a social networking site, which contains a very high bullying content. The data was labeled using an Amazon Web Service called Turk. A list of bad words downloaded from www.noswearing.com was used to assign severity levels. The number of bad words was normalized and then used as feature to develop the model. Then various machine learning algorithms were applied such as decision tree using J48, a rule based algorithm using JRIP, SVM (support vector machine),Naive Bayes algorithm and k-nearest neighbor approach using IBK (Instance Based Algorithm) using a Weka tool kit.

SVM (Support Vector Machine) is a supervised learning algorithm. Its goal is to find a hyperplane which maximizes the margin of the training data. Initially the classifier is trained with labeled data. The data is then converted into a data matrix based on the values in the vocabulary. The values are then plotted and a hyperplane is chosen in such a way that it maximizes the margin of the training data. Once the classifier is trained the input data is passed to this classifier to get positive and negative instances of bullying.

The goal of Decision Tree Algorithm is to divide the data into classes by processing it through the decision tree. The internal nodes define various attributes and the branches give us all the possible values of these attributes. The leaf nodes give us the final classification of data. For each node, information gain is calculated. The feature that gives the best information for classification is said to have the highest information gain.

The defining characteristic of Rule Based Algorithm is the identification and utilization of a set of relational rules that collectively represent the knowledge captured by the system. Such algorithms identify a set of rules that comprise the prediction model, or the knowledge base.

Naïve Bayes is a conditional probabilistic classifier that works by applying Bayes Theorem with naïve independence assumptions between the different features. Bag of words model representation is used and words are assigned severity label. Class with the higher posterior probability is the class classified.

Comparison of Results obtained from Machine Learning Algorithms is as follows [6, 22]:

Machine Learning Algorithm Used	Accuracy Obtained
J48	78.5%
IBK with K=1 and K=3	78.5%
JRIP	73.77%
Naive Bayes	72.30%
SVM	60.49%

Using Time Series modeling to detect cyberbullying. In this research [24]**,** the proposed method utilized a dataset of real world applications that is a set of questions and answers between cyber predator and the victim. It is a sequential data modeling approach in which the predator's questions are formulated using a Singular Value Decomposition representation. The aim of this research is to study the accuracy of predicting the level of cyberbullying attack using classification methods and also to examine potential patterns between the linguistic styles of each predator. The whole question set is modeled as a signal whose magnitude depends on the degree of bullying content. Using feature weighting and dimensionality reduction techniques, each signal is parsed by a neural network that forecasts the level of result within a question. By applying Singular Value Decomposition on the time series data, it is observed that its plot is very similar to the plot of the class attribute. By applying a Dynamic Time Warping algorithm, the similarity of the above signals was proved to exist, providing a hint of cyberbullying within a dialogue.

III. Data Source

The lack of availability of datasets is the main reason behind very few researches on cyberbullying detection. Most of the research conducted as mentioned above, have mainly concentrated on the dataset available on CAW 2.0 consisting of datasets of Kongregate, MySpace, and Slashdot where the data can be divided into two communities; chat style and discussion style communities. In chat-style communities posts are contained of short messages. In discussion-style communities, posts are of long messages. The data is not labeled too. It was labeled using Amazon Turk. In predator communications there is a very little labeled data. The dataset in such research used chat logs transcripts from Perverted Justice [25].A large dataset from Formspring, a question and answer based format social networking site, has also been used in the research.

IV. Conclusion

With the rapid growth of the Internet, people interact with other people. However, the chance for misuse comes with any new technology. These techniques lead to cyberbullying. Our literature review illustrates the cyberbullying detection techniques adopted so far to address this problem. We discussed machine learning techniques to detect cyberbullying. These techniques make

automatic detection of bullying messages in social media and this could help to construct a healthy and safe social media environment. So these techniques can be worked upon by collecting new datasets and then applying various machine learning algorithms on them to obtained desired accuracy. Sentiment analysis can be performed on the social media data by extracting data from various social networking sites using the available datasets and tools to identify the presence or absence of cyberbullying by determining the tone of text at sentence level or document level. Weka, Rapidminer, R, Orange are some of the tools that can be used for this purpose.

Efficiency of cyberbullying detection will decrease due to a constrained word set of negative words. A research can be carried out in detecting new lingo used in bullying so that the bad word set can be updated on a daily basis to improve the learning ability of the classifier. As mentioned before, there is no labeled dataset and very few datasets are available, so future research can work on collecting new datasets for the future study. A large dataset is also available on ChatCoder which contains online bullying data. So the dataset can be downloaded from here and research can be carried out further to detect cyberbullying so that appropriate actions can be taken regarding generation of awareness and counseling in order to eradicate this evil.

References

[1] https://www.stopbullying.gov/cyberbullying/

[2] J.Patchin, & S. Hinduja, "Bullies move beyond the schoolyard; a preliminary look at cyberbullying." Youth violence and juvenile justice.4:2 (2006). 148-169.

[3] Sourabh Parime, Vaibhav Suri "Cyberbullying Detection and Prevention: Data Mining and Psychological Perspective", 2014 International Conference on Circuit, Power and Computing Technologies [ICCPCT]

[4]http://www.TIMESOFINDIA.com

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

[6] K. Reynolds, A Kontostathis, and L. Edwards, "Using Machine Learning to Detect Cyberbullying," In Proceedings of the 2011 1Oth international Conference on Machine Learning and Applications Workshops (ICMLA 2011), vol. 2, December 2011. pp. 241-244.

[7] http://www.statisticbrain.com/cyber-bullying-statistics/

[8] http://en.wikipedia.org/wiki/Cyber-bullying/

[9] A. M. Chandrashekhar, Muktha G S & Anjana D K, "Cyberstalking and Cyberbullying: Effects and prevention measures" Imperial Journal of Interdisciplinary Research (IJIR) Vol-2, Issue-3, 2016 ISSN: 2454-1362

[10] R. Cohen, D. Y. Lam, N. Agarwal, M. Cormier, J. Jagdev, T. Jin, M. Kukreti, J. Liu,K. Rahim,R. Rawat, W. Sun, D.Wang, M.Wexler, "Using Computer Technology to Address the Problem of Cyberbullying", SIGCAS Computers & Society | July 2014 | Vol. 44 | No. 2

[11] Ted Feinberg, Nicole Robey, "Cyberbullying: Intervention and prevention strategies", Helping children at Home and School volume 3, pp. S4H15-1-4, National association of school psychologists, 2010.

[12] www.deletecyberbullying.org

[13] Jennifer Bayzick, April Kontosthathis and Lynne Edwards, "Detecting the presence of cyberbullying using Computer Software", WebSci '11, Koblenz, Germany, National Science Foundation, Grant No. 0916152, pp.1-4, June 2011.

[14] Jim Stern," Text analytics for social media", S.A.S. Whitepaper, pp.1-13, 2010.

[15] Naveen Kumar, Saumesh Kumar and Padam Kumar," Parallel Implementation of parts of speech tags for Text mining using grid computing ", Advances in Computing and Communications in Computer and Information Science Volume 190, Springer Publications, pp. 461-470, 2011.

[16] Vinita Nahar, Xue Li and Chaoyi Pang. "An effective approach for cyberbullying detection", Volume 3, Issue 5, Communications in Information Science and Management Engineering, pp 238-247, May 2013.

[17] David M. Blei, Andrew Y. Ng and Michael I. Jordan. "Latent Dirichlet allocation", Volume 3, Journal of Machine Learning Research, pp. 993-1022, 2003.

[18] Ramesh Prajapati, "A Survey Paper on Hyperlink Induced Topic Search (HITS)Algorithms for Web Mining", Volume1, Issue 2, International Journal of Engineering Research and Technology, pp.13-20, 2012.

[19] Tibor Bosse and Sven Stam, "A Normative Agent System to Prevent Cyberbullying", in IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2011 © IEEE. Doi: 10.1109

[20] www.cs.drexel.edu/~greenie/cs510/bdilogic.pdf.

[21] Mohsen Arab and Mohsen Afsharchi, "A Modularity Maximization Algorithm for Community Detection in Social Networks with Low Time Complexity", presented at The IEEE/WIC/ACM International Conference on Web Intelligence, WI, 2012.

[22] Samaneh Nadali, Masrah Azrifah Azmi Murad, Nurfadhlina Mohamad Sharef, Aida Mustapha,Somayeh Shojae, A Review of Cyberbullying Detection . An Overview. 2013 13th International Conference on Intelligent Systems Design and Applications (ISDA)

[23] Rui Zhao and Kezhi Mao,"Cyberbullying Detection based on Semantic-Enhanced Marginalized Denoising Auto-Encoder", IEEE TRANSACTIONS ON AFFECTIVE COMPUTING

[24] Nektaria Potha, Manolis Maragoudakis, " Cyberbullying Detection using Time Series Modeling", 2014 IEEE International Conference on Data Mining Workshop

[25] [Online], http://www.Perverted-Justice.com .2008.