Automatic Monitoring and Prevention of Cyberbullying

Rekha Sugandhi
Department of Computer Engineering
MIT College of Engineering
Pune, India

Anurag Pande
Department of Computer Engineering
MIT College of Engineering
Pune, India

Abhishek Agrawal
Department of Computer Engineering
MIT College of Engineering
Pune, India

Husen Bhagat
Department of Computer Engineering
MIT College of Engineering
Pune, India

ABSTRACT

The digital age has given rise to a new form of bullying, termed cyberbullying. A majority of teens use some sort of social media service, thus leading to cyber bullying becoming quite rampant and in some extreme cases, also resulting in victim suicides. In this paper, we aim to show the results of the system we designed for the automatic monitoring and prevention of cyberbullying. The response grading system takes into account the severity of bullying and gives appropriate responses.

General Terms

Machine learning, Algorithms

Keywords

cyberbullying; social media; data pre-processing; support vector machine; multiclass SVM;

1. INTRODUCTION

Cyber bullying can be defined as “Willful and repeated harm inflicted through the use of computers, cell phones, and other electronic devices” [1]. The reason why cyber bullying has arisen to become such a major problem as opposed to conventional bullying is its sheer reach. Cyber bullying differs from traditional bullying in the fact that it extends beyond the physical confines of public places like schools, parks, etc. with the victim often experiencing no respite from it [7]. The perpetrator, in the case of cyber bullying, is capable of harassing the victim without pause as the bullying is done online and does not require the physical presence of the victim. Another major problem when it comes to cyber bullying is the lack of identifiable parameters which mark any post as a bullying instance. Even after identifying bullying, judging the severity of the instance is a challenge as it can be simple name calling leading to social exclusion, or uploading embarrassing pictures that might have even worse consequences[5]. A victim can be exposed to multiple instances of cyber bullying over various modes available online and the large audience which can witness these instances makes it even more shameful and embarrassing [5]. A recent study conducted by Microsoft Corporation to understand the global pervasiveness of cyberbullying states that India ranks 3rd in cyberbullying after China and Singapore [2]. According to recent studies 52 % of the youth in India have had some experience with cyberbullying and about 38 % of them have been bullied themselves[3]. Cyber bullying is basically of two categories, one containing abusive language and the other which is embarrassing for the intended target but does not use any cuss words outright. Posts containing abusive content or bad words are more likely to be labeled as cyber bullying [6]. According to [10], for the current young generation “Gay”, ”Bitch” and “Slag” are the most commonly used terms of abuse in school.

Examples:

“Screw that damn Jacques, he has the most girly voice ever.” (Openly abusive)

“Go jump off a cliff and die.” (No abusive language involved)

India has high occurrences of bullying instances. 79% Indians are aware and worried about cyber bullying in comparison to 54% worldwide. 53% Indians have been bullied compared to a worldwide average of 37%. In addition to this, 50% Indians have been involved in bullying someone online while worldwide only 24% of the population has been involved in similar instances. On an upside 63% Indians are educated about and 76% institutions have a formal policy on cyber bullying in comparison to a worldwide average of 23% and 37% [8].

The system work-flow consists of the following steps:

1. The first step towards detection of cyber bullying is to get raw data sets from various online sources. Data sets for cyber bullying usually consists of user comments, posts, images and videos on social networking sites and social media. It is quite easy to get access to tweets from Twitter using the Twitter API [4]. Labeled training data was gathered from Chatcoder[9], which included text from various sources like MySpace, Formspring etc.

2. The collected data is then preprocessed and passed on to the classifier. We tested the accuracies of various classification algorithms (Naive Bayes, SVM and KNN) specifically in the detection of cyberbullying on our training data. The analysis gave us SVM as the algorithm which is most consistent and has the highest accuracy.

3. The sentiment of the sentence is calculated in parallel with the SVM classification. This is used as a double check to make sure that any data classified by the SVM model is not misclassified. This sentiment analysis system employs a method in which it assigns polarity values to each sentence based on a certain formula.

4. The multi class SVM takes our bullying data and classifies it into three different classes namely high, medium and low depending on the severity of the post in question.

5. Once the post is put into its respective class a response grading system implemented by us is then executed. This system gives a response based on the class in which the post is. High level posts result in

a temporary ban while low level posts result in a pop up in the form of a reflective user interface.

We have divided our paper into multiple sections, each describing a step leading up to the recognition and mitigation of bullying. Starting from taking in datasets and pre-processing our data, we then look into how our various classification schemes classify this data and end with a view of how our response grading system will help in preventing such bullying attacks.

2. DATA COLLECTION AND PREPROCESSING

The data for testing our project were tweets obtained from Twitter via the Twitter API by using Tweepy which is a python library using the former. For training data, we obtained labeled datasets from the website chatcoder.com the link for which was provided by Professor April Kontostathis whom we had contacted for getting the data. Of the multiple datasets available, we made use of the DataReleaseDec2011 and the BayzickBullyingData datasets for training our classifier. We use tweets as our data because they are easily obtainable, courtesy the twitter API which gives us the tweet, user id, user's friends and other relevant data.

The data in DataReleaseDec2011 represented 50 ids from Formspring.me that were crawled in Summer 2010. For each id, the profile information and each post (question and answer) was extracted.Each post was loaded into Amazon's Mechanical Turk and labeled by three workers for cyberbullying content.

The data in Bayzick dataset contains a small subset of data from a crawl of MySpace groups. The data has been manually labeled for bullying content by three independent coders.

All this data is preprocessed by stop words removal, removal of white space, tokenizing, stemming and removal of special characters.

3. CLASSIFICATION OF DATA

3.1 Choosing a classifier

We tested various classifiers on our bullying dataset to see which classifier gives us the best accuracy. For this we split our labelled data in an 80-20 split, where 80% of the data was used for training and 20% was used for testing the classifier. The labelled data consisted of 393 bullying posts and 2886 non bully posts. The classifiers which we compared were SVM with a LinearSVC kernel, Multinomial Naive Bayes and KNN algorithm. The results are tabulated below:

Table 3.1: Comparison of classifiers

Algorithm	Precision	Recall	F1 Score	Accuracy
SVM	0.91	0.91	0.9	91.31%
Naive Bayes	0.89	0.88	0.82	87.65%
KNN	0.89	0.89	0.86	88.87%

Based on the above results we choose to go along with the SVM classifier.

3.2 SVM classifier

We used ScikitLearn's Linear SVC classifier for classifyingour data as either bullying or a non-bullying post. We extract features from the raw data using a TFIDF vectorizer which gives a matrix of the tf-idf features. These feature matrices are then used to train the classifier.

3.3 Senti-net

The SVM classifier classifies the data as either bully or non-bullying. However, SVM loses a certain amount of accuracy while trying to classify sentences that have no profanities used in them. This is because the training data provided for SVM mostly has bullying instances containing profanities or swear words which constitutes as explicit bullying. In case of subtle bullying or implicit bullying, SVM misclassifies it as non-bullying.

For example: “Tell Jim to remove his make-up and hide behind Daddy's skirt!”

This sentence will be misclassified by SVM as it doesn't have any profanities and the training data provided to it lacks such bullying cases. The lack of this kind of data is because of a dearth of labeled bullying data on the internet.

To prevent such misclassification, any data classified as non-bullying by SVM is then passed through the sentiment analysis system. Here senti-net uses the AFINN-111 dictionary which consists of 2477 words and phrases assigned a polarity value from -5 to +5 with +5 being the most positive sentiment and -5 being the most negative sentiment. Using the sentiment value of each word in a sentence we calculate the total polarity of a sentence by using the formula,

$$\mathbf{P = \frac{\Sigma WiFi}{n}}$$

P = polarity of the sentence

Wi = polarity of the word

Fi = frequency of the word in the sentences

If while using the sentiment analyzer, a previously non-bully classified post by SVM is found to be negative in sentiment, then that post will be re-classified as a bully post. Thus, using a combined approach of SVM and senti-net, even subtle cases of bullying are detected

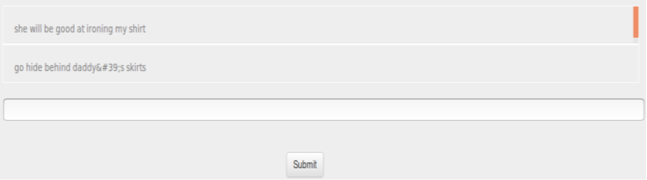
3.4 Multi-class SVM Classifier

The bully data needs to be further classified into three categories of high, medium and low depending on the severity of bullying. Thus, the bullying data is passed to the multi-class SVM classifier which uses the Linear SVC kernel. This classifier is trained by using only the bully data which was segregated into the three classes: high, medium and low based on the severity which is given in the labeled data by Amazon Mechanical Turk. It uses a One vs One approach in which one model is made for every class. The final classification decision is made by the class which has the highest match. The multi-class SVM classifier gives us an accuracy of 61.88%.

4. RESPONSE GRADING SYSTEM

The data is finally put into three classes namely, high, medium and low depending on the severity of bullying. Our system maps an appropriate response for each class taking into account the various parameters like the present social and political scenario, the severity, the overall sentiment against a particular issue, etc.

For posts in the low bullying category, a Reflective User Interface (RUI) is generated. A RUI is a pop-up which will inform the bully that his post is offensive and can hurt the victim's sentiments. It will also give him a count of the number of people who will be able to view this particular post along with a count-down timer of 60 seconds. Only after the 60 seconds are up will the bully be able to post the message. The framework for this design is obtained from principles espoused by DonaldSchönon reflective design, a Reflective User Interface is an array of solutions that might help stem or change the spread of hurtful online behavior.[11]Schön stated three notions of the reflective practitioner: "reflection in action", "reflection on action," and "ladders of reflections." Reflection in action is the notion we consider for our RUI as it would reflect on behavior asit happens so as to optimize the immediately following action. Through the interface, the end-user is encouraged (not forced) to think about the meaning of a given situation, and offered an opportunityto consider their options for reacting to it in a positive way. Reflection User Interfacesresist the urge to implement heavy-handed responses, such as direct censorship. Instead, the end-user is offered options to assist them in self-adjusting or seeking external help. The 60 second delay in posting is provided in the hopes that the delay will encourage reflection by the end-user. Such delays may not prevent severe cases of bullying. Alerting the end-user that their input might be hurtful and making them wait theircomment before actually submitting is very helpful. The end-user could decideto rephrase their comment or cancel it outright. This enforces a time for Schön's "reflection in action". Oftentimes the end-user does not realize that they are responding to the group's entiresocial graph, not just to the owner of the page they are commenting on, giving an indication of the number of people who will be able to view the post might make the end-user reflect upon his actions[7].

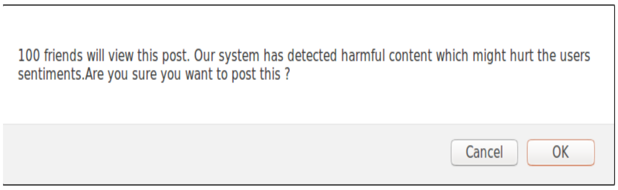


she will be good at ironing my shirt

go hide behind daddy&#39;s skirts

Figure 4.1: Sample conversation

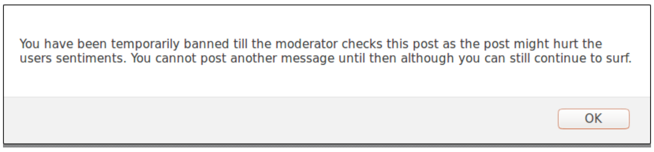
For medium severity of bullying, the end-user will be slapped with a 24 hour temporary ban during which time the moderator for the social network will ascertain the further action to be taken against the bully based on how offensive the end-user's post was. The moderator will have a separate page where he/she will get continuous alerts about medium-level bullying cases and where-in they can take further actions against the end-user as per their good judgment.The end-user will also be informed about the final action taken against him.



100 friends will view this post. Our system has detected harmful content which might hurt the users sentiments.Are you sure you want to post this ?

Figure 4.2: Reflective user interface

For high severity of bullying, the end-user will be banned outright from the social networking site and an alert will be sent to the moderator about this. The moderator will be able to decide on how long the ban will be enforced, be it life-long or for some months.



You have been temporarily banned till the moderator checks this post as the post might hurt the users sentiments. You cannot post another message until then although you can still continue to surf.

Figure 4.1: Temporary ban

All the victims in these bullying cases will also be provided various forms of online counseling by directing them to online groups where people anonymously help bully victims similar to an Alcoholic Anonymous online group, various YouTube videos, Depression chat-rooms, etc.

5. FUTURE SCOPE

As of now this system is implemented only on textual data. In the future we plan on extending the scope of our system by incorporating cross-media detection in the form of audio, video and images too. We also plan to try to make our system be context-aware with the help of deep learning in the future.

6. REFERENCES

[1] cyberbullying.org/about-us/ (Accessed 26th August)

[2] http://www.endcyberbullying.org/india-ranks-third-on-global-cyber-bullying-list/ (Accessed 28th August)

[3] http://indianexpress.com/article/technology/technology-others/alarming-50-indian-youths-have-experienced-cyberbullying/ (Accessed 29th August)

[4] L. Hon and K.Varathan, "Cyberbullying Detection System on Twitter", IJABM, Vol.1, No.1, April 2015.

[5] R. Sabella, J. Patchin and S. Hinduja, "Cyberbullying Myths and Realities ", Elsevier Transaction on Computers in Human Behaviour, Vol. 29, August 2013, Page No. 2703-2711

[6] K. Reynolds, A. Kontostathis and L. Edwards, "Using Machine Learning To Detect Cyberbullying", ICMLA, Vol. 2, December 2011, Page No. 241-244

[7] K. Dinakar, B. Jones, C. Havasi, H. Lieberman and R. Picard, "Common Sense Reasoning for Detection, Prevention and Mitigation of Cyberbullying", ACM Transactions on Interactive Intelligent Systems, Vol.2, No. 3, September 2012, Article 18

[8] http://download.microsoft.com/download/E/8/4/E84BEEAB-7B92-4CF8-B5C7-7CC20D92B4F9/WW%20Online%20Bullying%20Survey%20-%20Executive%20Summary%20-%20Singapore_Final.pdf (Accessed 31st August)

[9] http://www.chatcoder.com/DataDownload

[10] H. Sanchez and S. Kumar, "Twitter Bullying Detection", UC Santa Cruz

[11] SCHON D. 1983. The Reflective Practitioner. Basic Books, New York.