Methods for detection of cyberbullying: A survey

Conference Paper · December 2015

DOI: 10.1109/ISDA.2015.7489220

CITATION
1

READS
240

5 authors, including:

Rekha Sugandhi
Maharashtra Institute of Technology College of Engineering
12 PUBLICATIONS **30** CITATIONS

SEE PROFILE

Methods for Detection of Cyberbullying: A Survey

Rekha Sugandhi, Anurag Pande, Siddhant Chawla, Abhishek Agrawal, Husen Bhagat
Department of Computer Engineering
MIT College of Engineering
Pune, India
rekha.sugandhi@mitcoe.edu.in,anurag.r.pande@gmail.com,sidz2500@gmail.com,
abhi1994.aa@gmail.com,bhagathusen@gmail.com

***Abstract* - The advent of the digital age has paved the way for a new form of bullying which often leads to social stigma. With an increase in the use of social media platforms by adolescents, cyber bullying has become quite rampant and in some extreme cases it has also resulted in suicides by the victims. Very little efforts have been taken to curb this social menace and hence, this paper tries to address this issue by reviewing the steps that can be undertaken to detect cyber bullying on online social networks. This paper aims to review the different methods and algorithms used for detection in cyber bullying and provide a comparative study amongst them so as to decide which method is the most effective approach and provides the best accuracy.**

Keywords-cyberbullying; social medium; data pre-processing; support vector machine;co-reference resolution;

I. 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 [11]. 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 labelled as cyber bullying [10]. According to [21], for the current young generation "*Gay*", "*Bitch*" and "*Slag*" are the most commonly used terms of abuse in school.

Examples:
"*Kevin is a faggot.*" (Openly abusive)

"*Rohan looks good in a mini skirt.*" (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% [15].

The detection of cyberbullying involves the following steps:

1. The first step towards detection of cyber bullying is to get data sets from various online networks. 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]. Data from websites like YouTube [16], Facebook [17], Myspace [18], Instagram [19], etc. are also used for detection of cyber bullying along with ready-made datasets from websites like UCI, sentiment140.com and SNAP.
2. The next step involves data pre-processing which is used to process the data set so that the data set contains only relevant information. Removal of white spaces, stop words and special characters is a part of data pre-processing after which tokenization and

978-1-4673-8709-5/15/$31.00 ©2015 IEEE

lemmatization takes place. There are various other techniques which also can be used at this step to clean your data set.

3. The third and final step for detection of cyber bullying is classification of data. The data is classified into positive or negative instances of cyber bullying i.e. the data which definitely has cyber bullying content versus the data which has no significant cyber bullying content. In this survey, the CAW 2.0 dataset is the primary context that has been focused on, wherein classification is done for finding the subjectivity and polarity of the data. Classification algorithms could predict the label of an input but, to do this it needs a training set consisting of labelled examples before classifying new data [21]. Various algorithms and methods can be used for classification of data like pattern matching using bag of words, support vector machines(SVM), Naive Bayes algorithm, logistic regression, etc.

Each of these methods have their own merits and demerits pertaining to various aspects of cyber bullying.

In this paper, a few data pre-processing techniques are discussed along with the majorly important classification algorithms. As these classification algorithms have their various pros and cons and there has been no comparative study made for these algorithms pertaining to cyber bullying, this paper aims to provide just the same. Algorithms like SVM, Naive Bayes and J48 have been studied in this paper. In addition to this, data pre-processing techniques like tokenization, stop words removal, stemming, lemmatization, case folding and replacement of special symbols have been studied.

We have divided our paper into multiple sections each which describes an important aspect of cyber bullying. Starting from taking in datasets and pre processing our data, we then look into how various algorithms classify this data and end with a comprehensive comparison of these algorithms and their performance measures in response to various datasets.

II. DATA PREPROCESSING

Data preprocessing consists of a set of the following steps which serve to clean the text to facilitate easy processing by further analysis.

A. Tokenization:

This involves conversion of the large blob of unstructured text into a set of tokens divided by white spaces and/or punctuation marks, classified into categories such as words, phrases and sentences.

Example: "*I voted for Sammy because he was most inclined and considerate towards my ideals*", she said.

By applying tokenization we get the following set {[I],[voted],[for],[Sammy],[because],[he],[was],[most],[inclined],[and],[considerate],[towards],[my],[ideals],[she],[said]}

B. Stop words Removal:

Some very common words such as "*a*", "*and*", "*are*" etc add very little to the meaning of the text and are of little value in helping classify text. These 'Stop Words' are dropped from the text to ensure easier analysis of the text in further steps.

Using our above example the new set we get is {[I],[voted],[Sammy],[because],[he],[most],[inclined],[considerate],[towards],[my],[ideals],[she],[said]}

The words 'for', 'and' and 'was' are removed.

C. Replacement of Special Characters:

This step involves the replacement of characters such as "@" with "*at*", and removal of "#". This is of special importance in tweets due to the large number of occurrences of special characters.

D. Stemming and lemmatization:

Stemming is a heuristic method that simply truncates prefixes and suffixes to obtain the root of a word. Lemmatization is a refinement of the stemming technique that makes use of a dictionary-based approach for morphological analysis of words to obtain the base form of a word, called a lemma. Porter's algorithm [6] is widely used for lemmatization, as it has been consistently shown to be effective. In reference to the above example, simple stemming would transform it into:

{[Me],[vote],[Sammy],[cause],[he],[most],[incline],[considerate],[toward],[me],[ideal],[she],[say]}

E. Co reference Resolution:

Co reference resolution is the task of finding all expressions that refer to the same entity in a text. Repetition is one of the most common co referential devices in written text, which in turn makes string-match features more important to all co reference resolution systems [20]. "*It is an important step for a lot of higher level NLP tasks that involve natural language understanding such as document summarization,*

question answering, and information extraction" [20]. For example, consider the following sentence: "*I voted for Sammy because he was most inclined and considerate towards my ideals*", she said. Here "*I*", "*my*" and "*she*" refer to the same entity, as do "*Sammy*" and "*he*".

III. ALGORITHMS

A. Support Vector Machine

SVM (Support Vector machine) is a supervised learning algorithm, and is one of the most efficient and universal classification algorithms. Its goal is to find the optimal separating hyperplane which maximizes the margin of training data. Initially the classifier is trained with labelled data before being used to classify the data to test accuracy. Before the data can be used to train our classifier, it is imperative to process it. This consists of the following steps:

- Labelling of data
- Generation of vocabulary
- Creation of document-term matrix

Once the labelled data is converted into a data matrix based on the values in the vocabulary, the values are then plotted and optimal hyperplane is chosen based on the convex hull. The optimal 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 segregate it into positive and negative instances of bullying. This input data for testing purposes is also converted into data matrix and this data matrix is passed to the classifier. SVMs use sophisticated statistical learning theory to overcome the curse of dimensionality [8].

Instead of specifying the feature vector, kernel functions can be used to provide similarity between data points. There are various kernels that can be used with SVM namely,

- RBF kernel (Radial basis function)
- Linear kernel
- Gaussian kernel

Linear kernel is a special case of the RBF kernel, and works best when the number of features is very large. The linear kernel on data sets acquired from Myspace, Kongregate and Slashdot datasets were used. The datasets are available from the workshop on Content Analysis for the Web 2.0 [12]. The datasets contain manually-labeled data from [13], which is used as a ground truth dataset. Data from 3 different social networking sites are included in the dataset: Slashdot (496 files, 140,000 comments total (one for each article)), Kongregate (12 files, 150,000 comments total (one for each chatroom)) and MySpace (16346 files, 380,000 comments each (one for each thread)). Kongregate, an online gaming site, provides user messages from chat logs. Due to inherent frustration when playing online games, as well as a textual way to reach opponents, aggression is common in the posts. Slashdot is a discussion-based social networking site wherein users broadcast messages to others. MySpace is a popular social networking website. Datasets are in the form of XML files each containing and describing a discussion thread with multiple posts. Each post was extracted as a singular data element. Each data element is considered as one document and indexed through the inverted file index, assigning an appropriate weight to each individual term [7]. Applying LibSVM using a linear kernel, followed by ten-fold cross-validation gives a false positive of 28 in 294 instances and false negative of 12 in 10184 instances [7]. The model used in this case is the weighted TF IDF model. Over sampling of the training cases was used to improve the training [7].SVM with linear kernel using unigrams gives an accuracy of 79.6% while with bigrams it gives 81.3% leading to the conclusion that bigrams should be used with the SVM linear kernel. This conclusion was obtained after testing the above on twitter corpus data (1762 tweets – 39% labeled as bullying traces). Taking all this data into consideration our conclusion is that linear SVM in combination with bigrams gives the best possible accuracy [9].

B. J 48

This algorithm is derived from the C 4.5 algorithm. It's a univariate decision tree algorithm. The goal of this method is to divide data into classes by processing it through the decision tree. The decision as to which class the specified data goes into is taken based on previously classified dummy data sets.

Some basic steps are given below to construct the tree

- First, check whether all cases belong to the same class, then the tree is a leaf and is labelled with that class.
- For each attribute, calculate the information and information gain
- Find the best splitting attribute (depending upon current selection criterion)

The internal nodes of the decision tree 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. The feature which can give us the best information for classification is said to have the highest information gain. These are the features which we choose for faster and easier classification. Once the data is separated, in our case, into bullying and non-bullying instances, it is assigned to the appropriate classes. Thus the input tweets we have can be passed through the classifier and segregated successfully. The advantage of this algorithm is that it can take tables of

data with a huge number of columns and classify them into simple decision trees.

A Twitter corpus from the CAW 2.0 dataset was obtained. This corpus contains around 900,000 posts collected from amongst 27,135 users. After filtering this dataset, based on social connection and available chat history between persons involved in a post, a final result was 4865 messages was obtained. This data was pre-processed using a dictionary of bad words, consisting of 713 curse words. In addition to this, a dictionary of hieroglyphs (eg. @SS, 5hit) and emoticons was also used [14]. To filter the data even further, it was scanned for upper case letters and exclamations as they signify instances of shouting.

For classification, 70% of the data was used as training set with the remaining being used as the testing set. This resulted in a true positive rate of 0.259. It is calculated by using the formula:

$$\frac{mean(bully)-mean(nonbully)}{mean(bully)}$$

This ratio has been calculated taking into account textual as well as social features. The social features are derived using ego networks. The true positive rate using only textual features is 0.081. This shows that by including social features, the true positive rate increases by a huge margin.

C. *Naïve Bayes*

The Naive Bayes family of classifiers are simple conditional probabilistic classifiers that work by applying Bayes theorem with naive independence assumptions between the different features. All features are assumed independent given label Y:

$$P(X_1,\ldots,X_n|Y)=\prod_{i=1}^{n}P(X_i|Y)$$

A very simple document representation is used here, usually bag of words. Words important to the meaning of the text, and thus imperative in its classification, are considered, and given weight according to meaning, or in this case, severity. For instance, "*faggot*" would receive a higher weight than "*bitch*", due to the former being sexually discriminatory and abusive.
Thus, given a document 'd' and class 'c':

$$P(c\,|\,d)=\frac{P(d\,|\,c)P(c)}{P(d)}$$

The maximum posterior class, or the most likely class, being in our case either bullying or not, would be:

$$c_{MAP}=\underset{c\in C}{\operatorname{argmax}}\,P(c\,|\,d)$$

$$=\underset{c\in C}{\operatorname{argmax}}\,\frac{P(d\,|\,c)P(c)}{P(d)}$$

$$=\underset{c\in C}{\operatorname{argmax}}\,P(d\,|\,c)P(c)$$

The corpus of data obtained to experiment with is the same as that used for J48. In this case, a true positive rate of 0.723, taking into account both textual and social features, was obtained. Without taking into account social features, the rate was 0.584 once again proving, as with similar tests performed with J48, that social features help improve the result [14].

IV. Comparison

The Support Vector Machine algorithm implemented with linear kernel using bigrams is known to be a definite improvement in accuracy over general datasets over other algorithms [9]. The results can be summarized as follows:

1. **SVM** was used to analyze a dataset of 10487 messages from the CAW 2.0 dataset. It has been shown to have a false positive ratio of 0.0952 and a false negative ratio of 0.0012. Also, using a bigram model rather than a unigram model gives a 1.7% boost to accuracy. Furthermore, studies by Bellmore et al have shown that SVM when implemented using a Part of Speech in addition to the bigram model, gave an accuracy of 81.6% [9].
2. **J48** was used to analyze the twitter corpus from CAW 2.0, consisting of 4865 messages. It has been shown to have a true positive rate of 0.259 when social as well as textual features were taken into account, as compared to 0.081, when considering only textual features [14].
3. **Naïve Bayes** algorithm was run on the same dataset as J48. It was similarly shown to have an accuracy of 0.723 when social features were also taken into account, and 0.584 when considering only textual features [14].

V. Conclusion

After comparing the above mentioned algorithms, we realize support vector machines have given the best result. We plan to implement SVM in our project as the primary classifier for our base dataset. In addition to the above mentioned algorithm, we also take social features into consideration to increase accuracy. To further optimize our results, we would like to introduce Hidden Markov models to first classify the data into a few predefined categories. We would also take the aid of common sense reasoning to do the same. In

addition to this, the implementation of support vector machines to distinguish bullying traces from non-bullying ones will give us a better result. This knowledge would be used to prevent bullying at its source. Finally, we would like to introduce a response grading system which would categorize the bullying instances, thus making it easier to deal with each instance as deemed appropriate.

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] P. Willett, "The Porter Stemming Algorithm: Then and Now", Electronic Library and Information Systems, Vol. 40, No. 3, April 2006, Page No. 219-223

[7] V. Nahar, X. Li and C. Pang, "An Effective Approach For Cyberbullying Detection", Communications in Information Science and Management Engineering, Vol. 3, Iss. 5, May 2013, PP. 238-247

[8] E. Greevy and A. Smeaton, "Text Categorisation of Racist Texts Using a Support Vector Machine ", JADT 2004

[9] J. Xu, K. Jun, X. Zhu and A. Bellmore, "Learning from Bullying Traces in Social Media", Conference of The North American Chapter of the Association for Computational Lingustics: Human Language Technologies, June 2012, Page No. 656-666

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

[11] 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

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

[13] 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.

[14] Cyberbullying Detection Using Social and Textual Analysis - Qianjia Huang, Vivek Singh, Pradeep Atrey

[15] 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)

[16] www.youtube.com

[17] www.facebook.com

[18] www.myspace.com

[19] www.instagram.com

[20] http://nlp.stanford.edu/projects/coref.shtml(Accessed 30th August)

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