International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 – 5161

Available at http://inpressco.com/category/ijcet

Research Article

A Framework for Cyberbullying Detection in Social Network

Krishna B. Kansara†* and Narendra M. Shekokar†

†Department of Computer Engineering, D J Sanghvi College of Engineering, Vileparle (west), Mumbai-56, Maharashtra, India.

Accepted 25 Feb 2015, Available online 28 Feb 2015, **Vol.5, No.1 (Feb 2015)**

Abstract

Due to tremendous growth and popularity of current web technologies and social networking, online presence of the person has become permanent now. Through social networking connections, people usually express feelings, opinions, emotions and the most commonly seen is discussing the daily incidences which may include private or public talks. Especially in teenagers and young adults, expansion of social connections in terms of friendships has become fast without careful consideration of what is being shared and with whom. Such environment has become an avenue for cyber bullying, where abusive contents in terms of text messages or images are posted on the social network. For victims, this kind of behavior can lead to depression and other severe life-threatening problems. Therefore, measures need to be put in place to monitor and detect potentially harmful online activities. Manual check for all posts on social network is highly impossible because of large number of users. Therefore, in this paper, we have proposed a framework deployed for the detecting negative online interactions in terms of abusive contents carried out through text messages as well as images. The combination of text & image analysis techniques is considered as a suitable platform for the detection of potential cyber bullying threats.

Keywords: *Social Network, Cyberbullying, Abusive Image and Text Detection, BoW and BoVW model.*

1. Introduction

Every day, billions of people communicate on the online social network. Facebook, with more than a billion of users, is currently the largest and most popular OSN in the world. Other known OSNs are Google+, with over 235 million of users; Twitter, with more than 200 million users; and LinkedIn, with around 160 million users (Fire M, Goldschmidt R, Elovici Y. *et al*, 2014). Usage of social network is growing rapidly for sharing private and/or intimate information by various applications that assist users to get in close contact with others without considering cyber security breaches. These kinds of communication may lead towards some hazardous outcomes in terms of injecting various kinds of security attacks in social network. Message posts can contain the sharing of some kinds of abusive or offensive contents which can emerge the threats like cyberbullying. Usually, Adults can be able to put a line of secure communication and are having better awareness of often curious to explore new fields without the ability the existing dangers in social networks bring along. By contrast, children or teenagers often have wrong threat perception and are to weigh up potential risks (Rybnicek M, Poisel R, Tjoa S, *et al*, 2013). Cyberbullying is an online attack based on deliberately insulting, threatening, embarrassing or harassing people on the internet. It also indicates the Communication with the victim through the abusive texts and images via social media. Fig.1 shows the various ways of cyberbullying generally occurs on the social network. As insults, rumors and misinformation can be immediately disseminated to a large audience, cyberbullying in social networks is particularly painful for victims (Rybnicek M, Poisel R, Tjoa S, *et al*, 2013).

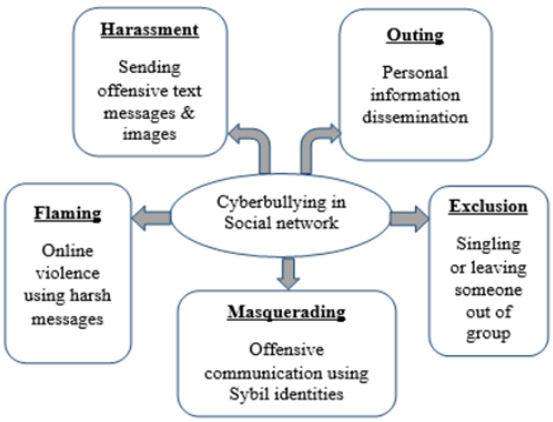

Fig. 1 Various ways of Cyberbullying in Social Network

Cyberbullying gives a bully the power to embarrass or hurt a victim before an entire community online (Dinakar K, Reichart R, Lieberman H *et al* ,2011). The results of cyberbullying in terms of physical or mental

*Corresponding Author: **Krishna Kansara**

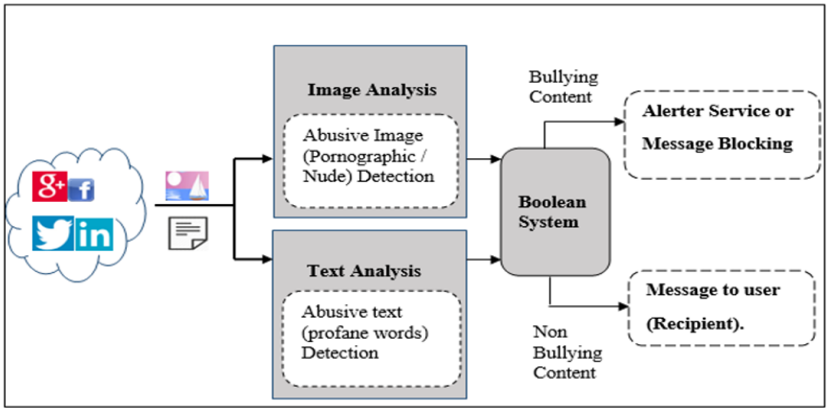

Fig.2 Framework for the implementation of detection

offences to victims can turn into the very serious and life threatening incidents which include depression and committing suicide. Research has shown that at least 11% of all youngsters, between the age of 12 and 18, have been a victim of cyber bullying (Deselaers T, Pimenidis L, Ney H *et al*, 2008). For example, Phoebe Prince, around 15 years student of a high school, committed suicide as a result of being cyberbullied by negative comments on the one of the social network (T Vanhove, P Leroux, T Wauters, *et al*,2013). Traditionally, bullying was considered to be a face-to-face encounter between children and adolescents in school yards, but now it has also found its way into the cyberspace. Protection platform against these kinds of harmful acts on the social network is required. So, we have proposed an automated system to detect cyberbullying activities like abusive text messages or images on the social network. In image analysis, Bag of Visual Words (BoVW) concept has been used with the SVM classifier for classification of adult images and for analysis of text messages, Bag of words (BoW) is incorporated with the Naïve Bayes classifier to classify the abusive text messages. Finally, Boolean system categorize the bullying content by considering the analysis results obtained by the text & image classification modules. The remainder of this paper is organized as follows: In Section II, we review similar research for analysing cyberbullying in social network. The proposed architecture is described in Section III and Section IV concludes this paper.

2. Related Work

As the cyberbullying is the emerging attack and the research is still carrying out, detection of the cyberbullying using combined approach of image and text is rare. Survey of the literatures based on the detection of cyberbullying emphasis on different aspects of detection. Dinakar and Birago Jones *et al*. (2011) proposed the study of various aspects for social network and solutions to overcome the problem of cyberbullying by giving suggestions on the theoretical basis. None of the specific approach has been discussed Maral Dadvar *et al*. (2012) incorporated the users' characteristics information and post-harassing behavior. Cross-system analyses of the users' behavior-monitoring their reactions in different online environments can lead to more accurate detection of cyberbullying. This suggested technique for detection of cyberbullying performed by text analysis using few words occurrence in the comments or post in social network. Comment made by male or female is also one of the analysis factor included in this research. This research on detection of bullying content using text analysis with suggested techniques has been experimented with only few words. Thomas Vanhove *et al*. (2013) proposed a pluggable architecture with components having features of reusability, mostly able to detect harmful content quickly. The platform uses text, image, and audio and video-based analysis modules to detect inappropriate content. Analysis where none of the specific technique has been proposed for detection of cyberbullying. Rybnicek M. *et al*. (2013) proposed a research application providing the platform for detection of cyberbullying. Research in this paper mentioned suggestions in the research platform is including the text, audio, video, images and addition to that social media analysis has been covered but for the implementation, the domain services aggregate this data and flag user profiles if necessary. Moderators on social network check the validity of the flagged profiles. The advance topic covered in this research that it provides the efficient knowledge upon key requirements of the platform, the essential components for architectural design and challenges to deal with cyberbullying using audio, video, and text and images but it only the survey of suggested techniques cannot guarantee for the efficient result as implementation strategies have not proved the result analysis.

3. Proposed System

Most of the above mentioned systems provides the research platform for analyzing the areas of cyberbullying detection in much generalized way. There are many ways in which bullying activities can be performed like threatening, harassment, black mailing and use of offensive language and vulgar

Cyberbullying Detection images, audio, video are used to provoke it. Some of the above researches have included effective techniques for detection of bullying activities using text analysis only. The combination of image and text analysis can yield an efficient result for cyberbullying detection. To incorporate that we have proposed an automated system framework as shown in fig.2 to detect cyberbullying using abusive image detection and abusive text detection. Any ill-minded user can post abusive image or text, which can be detected by the system to block immediately before that abusive content get broadcasted to the social network. Designed framework has two main modules: (1). Abusive image detection. (2). Abusive text detection.

3.1 Abusive image detection

It includes the detection of abusive – nude or pornographic image. As a growing number of people specially the minors, are prone to pornographic images, it is necessary to prevent access to these kinds of contents.

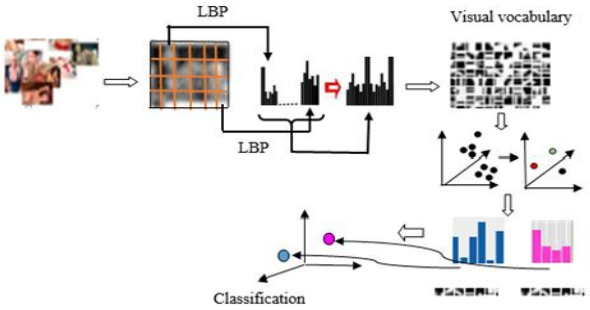

Fig. 3 process of abusive image detection using BOVW and classifier

As shown in fig.3, the process of abusive image detection started from the feature extraction where in the detection and description of interest points of an image is detected using Local Binary Pattern (LBP). The extracted features using LBP is assigned or mapped to the existing visual word in vocabulary. Visual vocabulary or codebook is generated using K-means algorithm where clustering of similar features forms a one visual word. Assignment of each local descriptors or features to its nearest visual word is performed which is a description of the image as vector of visual word frequencies (i.e. number of occurrences). The occurrence of certain visual words provides powerful hints for the presence of offensive material in image. Finally, a discriminative classifier is used for classification thereafter.

3.1.1 Feature extraction

An image has local interest points usually around the corners and edges. The detected points are described by local descriptors. Local binary pattern (LBP) which is mainly used for feature extraction using texture analysis, LBP labels the pixel of an image by thresholding the neighborhood of each pixel with the value of the center pixel and considers the result as a binary number. For preparing LBP feature vector, input image is divided into cells (e.g. 16 x16 pixels for each cell). For each pixel in a cell, comparison of the each pixel to each of its 8 neighbors (on its left-top, left-middle, left-bottom, right-top, etc.)is done following and including all the pixels along a circle, i.e. in the direction of clock and anti-clock. Where the center pixel's value is greater than the neighbor's value, the value is 1. Otherwise, 0. This gives an 8-digit binary number (which is usually converted to decimal for convenience). Histogram is computed, over the cell, of the frequency of each number occurring (i.e., each combination of which pixels are smaller and which are greater than the center). Optionally after normalization of the histogram, concatenation of histograms is processed for each cell. This gives the feature vector for the input image using local binary pattern (LBP). Since the BoVW model represents an image as order less collection of local features, it discards the information about the spatial layout of the features and thus has limited descriptive ability. Specifically, it cannot describe shape or segment an object from its background. To overcome the disadvantage of the basic BoVW model, Spatial Pyramid Matching (SPM) that repeatedly subdivides the image and computes histograms of local features at increasingly fine resolutions. Therefore, the spatial information can be encoded in the BoVW model. The combined technique of BoVW model with LBP (Local Binary Pattern) results efficient accuracy for feature extraction.

3.1.2 Visual vocabulary - (BoVW) model

BoW (Bag of visual word) model is a histogram representation based on independent features. It is applied to image classification as Bag of visual word (BoVW) model. For feature detection, each image is represented by several local patches. Numerical vectors which are generated by the feature representation methods to represent the patches. They are called as feature descriptors. After this proceesing, now each image is a collection of vectors of the same dimension where the order of different vectors is of no importance. Final step for the BoVW model is to map vector represented by patches to visual words which belongs to visual vocabulary (visual word dictionary). The number of visual words represents the visual vocabulary size. Visual vocabulary has many visual words where a visual word can be considered as a representative of several similar image patches. Thus, each patch in an image is mapped to a certain visual word through the clustering process (K-means) and the image can be represented by the histogram of the related visual words of an image. Finally, a classifier classify the image as abusive or not.

3.1.3 Machine Learning Classification

Support vector machines (SVM) are supervised learning models with associated learning algorithms

that analyze data and recognize image patterns. Given a set of training images to the classifier, each image marked as belonging to abusive class by the classifier if given image contain abusive or pornographic contents. SVM training algorithm builds a model that assigns new image into the abusive class or categories by applying the learned rules to identify the abusive image. Support Vector Machines are based on the concept of decision planes that define decision boundaries. A decision plane separates set of objects having memberships of different classes.

3.2 Abusive text detection

For detection of abusive text messages, system will perform pre-processing of the text messages and then bag of words is applied for extracting the features of text message. From the extracted features, the matrix is generated which is used by the classifier to categorize the abusive text messages.

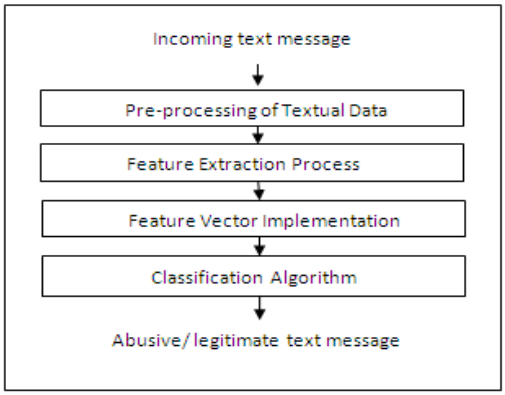

Fig. 4 Process Map for Abusive text detection

3.2.1 Pre-processing

The message field needs to be pre-processed because it contains unstructured text. The purpose of pre-processing is to transform messages into a uniform format that can be understood by the learning algorithm. In pre-processing, the process of tokenization, stemming and stop word removal, number removal and strip whitespaces is performed. For each sentence in the dataset, an automatic spelling checking and correction of grammar leads to the introduction of the dataset to the classifier. With the help of WordNet corpus and spell-correction algorithm, correction of spelling and grammar mistakes in the raw sentences occurs by tasks such as deleting repeated words, meaningless symbols and splitting of long words, substituted letters transposition ,replacing the incorrect and missing letters in words.

3.2.2 Bag of word (BoW) model

One of the most important subtasks in abusive text classification is feature extraction. By using machine learning algorithms for training the classifier, representation of text as a feature vector is required. For that a commonly used model in Natural Language Processing is Bag of Words (BoW) model. The primary stage of this model is the creation of vocabulary of words which is in our approach indicates the vocabulary or the collection of abusive words. Among the baseline approaches for text classification, the BoW approach has the highest recall rate 66% (Chen, Ying, *et al* 2012). In BoW model, each word is associated with a count of occurrences. This vocabulary can be understood as a set of non-redundant words where the order doesn't matter. The BoW approach disregards grammar and detects offensive sentences by checking whether or not they contain abusive or offensive words. In our approach, Each text message is represented as a feature vector composed of binary attributes for each word that occurs in that message Let *{w1,...,wm}* be a predefined set of *m* features (vocabulary of words)that can appear in a message. Let *ni(d)* be the number of times w*i* occurs in a message *d*. Then each message*d* is represented by the message vector *d:=(n1(d), n2(d),...,nm(d))*. If a word present in the vocabulary appears in a given text message, its corresponding attribute is set i.e. 1, else it is set to 0.

3.2.3 Classification

Using the multivariate Bernoulli Naïve bayes model, where in the features are independent Booleans (binary variables) describing inputs. Binary term occurrence features are used rather than term frequencies. If F_i is a Boolean expressing the occurrence or absence of the i'th term from the abusive words vocabulary, then the likelihood of a textual message belongs to an abusive class C is given by

$$p(F_1,..,F_n|C) = \prod_{i=1}^{n} [F_i\, p(w_i|C) + (1 - F_i)\ (1 - p(w_i|C))]$$

Where $p(w_i|C)$ is the probability of class C generating the term w_i. Bernoulli's naïve bayes, explicitly perform modelling of the absence of terms. Naive Bayes classifier with a Bernoulli is not the same as a multinomial NB classifier with frequency counts truncated to one.

3.3 Boolean system

Results from the image and text analysis is fed to the Boolean system. For that results are converted to binary 1 or 0 values as per the image and text analysis results and then it is processed further by the Boolean system for finally determining the contents are bullying or not. Boolean system signals true for at least any of either image or text analysis results as true. In summary, the proposed automated system can identify the abusive image in messages by the combined approach of bag of visual word (BoVW) model, local binary pattern (LBP) and SVM classifier. Text message

abusiveness detection includes bag of word (BoW) model with naïve bayes classifier. At the final stage the Boolean system is used to categorize the bullying content by analyzing the result obtained by both the image and text classification.

Conclusions

Up till now, a very few implementation has been done to detect the cyberbullying activities by analyzing the combination of text and image data. Only the text analysis have gained the majority as a detection mechanism. We have proposed an automated system that can identify the abusive kinds of images and text messages. Our system analyzes the abusive image using local binary pattern & bag of Visual word concept whereas text analysis is conducted using Bag of Word model. We are using a Boolean system to detect the presence of bullying content by evaluating both image and text analysis results. Our Proposed approach for analysis of incoming messages in terms of image or text on the social network is feasible for the detection of cyberbullying.

Future work

Other than Image and Text messaging, Cyberbullying using audio and video contents can also provoke the victims to step forward for some serious life threatening decision. Victims can be harassed by any kind of horror or violent audios or videos. As the future work, we can implement cyber bullying detection techniques for audio and video messages. In the text analysis also we can include sentiment or context related knowledge to enhance the detection of bullying text contents.

References

Rybnicek M, Poisel R, Tjoa S,(2013) Facebook Watchdog: Research Agenda for Detecting Online Grooming and Bullying Activities, *Systems, Man, and Cybernetics (SMC), 2013 IEEE International Conference*.

Lieberman H, Dinakar K, Jones B,(2011) Let's Gang Up on Cyberbullying, *Computer*.

Maral Dadvar, F De Jong (2012) Cyberbullying Detection – A step towards a safer internet Yard Proceedings of the 21st international conference, dl.acm.org.

T Vanhove, P Leroux, T Wauters, (2013)Towards the Design of a Platform for Abuse Detection in OSNs using Multimedia Data Analysis , *Integrated Network Management (IM 2013), 2013 IFIP/IEEE International Symposium*.

Deselaers T, Pimenidis L, Ney H (2008) Bag-of-visual-words models for adult image classification and filtering. In: Proceedings of the 19th International Conference on Pattern Recognition.

BoW-based Image/Scene Classification with Naive BayesClassifiers/SVMs,[online]https://sites.google.com/site/yorkyuhuang.

Dinakar K, Reichart R, Lieberman H,(2011) Modeling the detection of textual Cyberbullying, International Conference on Weblog and Social Media - Social Mobile WebWorkshop,Barcelona.

Hosseinmardi, Homa, (2014) Towards understanding cyberbullying behavior in a semi-anonymous social network. *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE.

Munezero, Myriam, (2013) Antisocial behavior corpus for harmful language detection. *Computer Science and Information Systems (FedCSIS), 2013 Federated Conference on*. IEEE.

Reynolds, Kelly, April Kontostathis, and Lynne Edwards.(2011) Using machine learning to detect cyberbullying. *Machine Learning and Applications and Workshops (ICMLA), 2011 10th International Conference on*. Vol. 2. IEEE.

Fahrnberger, Gunter,(2014) Safe Chat: A tool to shield children's communication from explicit messages. *Innovations for Community Services (I4CS), 2014 14th International Conference on*. IEEE.

Chen, Ying, (2012) Detecting offensive language in social media to protect adolescent online safety. *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Conference on Social Computing (SocialCom)*. IEEE.

Li, Mingmei, and Atsushi Tagami.(2014) A Study of Contact Network Generation for Cyber-bullying Detection. *Advanced Information Networking and Applications Workshops (WAINA), 2014 28th International Conference on*. IEEE.

Fire M, Goldschmidt R, Elovici Y ,(2014) Online Social Networks: Threats and Solutions, *Communications Surveys & Tutorials, IEEE*.