International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 4 Issue 05, May-2015

Approaches for Mining YouTube Videos Metadata in Cyber bullying Detection

Mr. Shivraj Sunil Marathe
M. E. Student
Vidyalankar Institute of Technology
Mumbai, India

Prof. Kavita P. Shirsat
Assistant Professor
Vidyalankar Institute of Technology
Mumbai, India

***Abstract* - Recent years have witnessed the evolution of Web 2.0 such as social-networking sites, video sharing sites, wiki & blogs, etc. which has drastically increased the volume of community-shared textual & media resources including posts, comments, videos and images.**

Moreover, today Web 2.0 has become an effective communication platform for people to promote their ideas, share resources, and communicate among each other. As a result, various malignant users have also been attracted towards video social networks. YouTube, Yahoo! Screen, Dailymotion, Vimeo, Vube are some of the popular video sharing sites on web. Among them YouTube is the largest and most popular free video social network.

A significant percentage of data uploaded on YouTube contains objectionable content and violates YouTube community guidelines. Currently YouTube contains several copyright violated videos, spam's, hate and extremism promoting videos, cyber bullying content along with vulgar and pornographic material. Out of these, presence of cyber bullying & harassment related videos is one of the major problem. This is primarily due to the anonymity and low publication barriers to content uploaded.

In this paper we have reviewed the existing approaches for detecting cyber bullying promoted through video social networks like YouTube.

Keywords – YouTube; video social network; video sharing sites; cyber bullying; harassment

I. INTRODUCTION & RESEARCH MOTIVATION

YouTube[1] & other video social networks allow users to upload and watch unlimited number of videos for free. In addition, they support number of social networking features. For example, a user can like or dislike a video, share it on other social networking websites like Facebook[2], Twitter[3], LinkedIn[4], etc. Also user can post comments in textual form, subscribe for a particular channel[5], search any video using keywords & category and interact with other users by comments and replying on the comments.

According to YouTube statistics[6], YouTube has 1 billion unique users. Every day people watch hundred millions of hours of videos on YouTube and generate billions of views. YouTube has several community guidelines posted on its website in order to stop users to upload inappropriate content. However, despite of these community guidelines [13], large number of objectionable content is present on YouTube [1]. Anonymity, low publication barriers and high reachability of videos worldwide has led users to upload many cyber bullying and malicious content on the website. This shows cyber bullying & harassment as major concern for YouTube.

In the context of YouTube and other video social networks, we can define cyber bullying as unauthorized shooting of a person's video and uploading it on website. If the scenes in the video are negative (such as vulgarity, violence and abuse) then public disclosure of such content can be regarded as harassment of the person in the video. Cyber bullying can be of two types intentional and unintentional. Sometimes users post videos on a website in order to threat and disturb one or more people. For example, violent, abusive and humiliating act that violates the claimant's dignity. And sometimes users take a clip of some incident and share it on a website without any intention to hurt that person involved in the video.

Table I shows statistics of few videos posted on YouTube. These statistics are collected in January, 2015. Table I reveals that the cyber bullying videos are also very popular and have a large number of views & likes. We notice several key terms present in the title and description showing that the videos are objectionable according to YouTube policy.

Many papers have been published on the detection of cyber bullying in video social networks. But so far no review paper has been published in this field which consolidated the existing research. Our paper aims to provide a review of the academic research and work done in this field by various researchers. This paper is structured as follows: Section II describes methodology used to carry out this review; followed by cyber bullying theory which have been briefed in Section III & Section IV; Features distinguishing cyber bullying videos have been covered in Section V; Section VI covers the existing approaches for cyber bullying detection; Further, Research directions and future challenges are noted in Section VII; finally Section VIII concludes the review.

II. METHODOLOGY

This survey of existing approaches for detecting cyber bullying in YouTube has been done after a systematic review with principled approach in which major digital libraries for Computer Science have been searched like IEEE Xplore, ACM Digital Library, Springer, Google Scholar, and Science Direct for concerned topic. We focused on papers after year 2009 only; as the concept of Web 2.0 started evolving since 2005 and became popular later.

[1]https://www.youtube.com
[2]https://www.facebook.com
[3]https://www.twitter.com
[4]https://www.in.linkedin.com
[5]http://smallbiztrends.com/2009/05/5-reasons-youtube-social-marketing strategy.html
[6]http://www.youtube.com/yt/press/statistics.html

TABLE I. SAMPLE OF FEW CYBER BULLYING & HARASSMENT VIDEOS (POSTED ON YOUTUBE) AND THEIR STATISTICS. # = NUMBER OF

Sr. No.	Video ID	Category	#Views	Duration (Sec)	#Likes	#Comments	Key Terms	Uploaded Date
1	Fz_AxXR9W7c	Entertainment	81,205	344	70	20	Ragging, kissing	27/01/2014
2	jtnLlExFbqY	Entertainment	667	39	10	1	ragging, seniors	29/12/2014
3	Brg9vZaJqu4	Entertainment	1,823,025	547	468	125	ragging, girls, hostel	01/12/2010
4	kxQwb5nW-lo	People & Blogs	2,737,650	80	1,106	41	girls, hostel, mms	17/12/2013
5	RwN9OYgfMOc	People & Blogs	252,019	772	91	54	ragging, juniors	05/09/2012

YouTube was launched in February, 2005 which became very popular after it was bought by Google in November, 2006. So it took some time for people to get familiar with YouTube & other video social networks for communication and hence making inappropriate use of it.

Papers reviewed for this study were selected after reading titles and abstracts of all the papers. Only those papers were chosen that were found suitable for the present study. Papers with titles and abstracts regarding cyber bullying & its detection counting a total of 12 papers have been selected for review. Mainly the papers have been categorized on the basis of features used to detect cyber bullying content. Through this paper we are trying to compile a list of papers in detection of cyber bullying content. After going through this survey paper, new researchers can easily evaluate what work has been done, and how the present work can be extended to make cyber bullying detection more accurate.

III. CYBER BULLYING DEFINED

What makes cyber bullying so dangerous - is that anyone can practice it without having to face the victim. Person doesn't have to be strong or fast, simply equipped with a cell phone or computer and a willingness to bully someone.

Cyber bullying is the form of bullying which takes place with the help of electronic technology. Electronic technology means devices and equipment such as cell phones, computers, and tablets as well as communication platforms including social media sites, text messages, chat, and websites. Examples of cyber bullying are mean text messages or emails, rumors sent by email or posted on social networking sites, posting or sharing embarrassing pictures, videos, websites, or creating fake profiles.

Patchin and Hinduja [7] define cyber bullying as "willful and repeated harm inflicted through the medium of electronic text." Willful harm excludes sarcasm between friends comment(s) meant to criticize or disagree with an opinion but not meant to attack an individual.

IV. TYPES OF CYBER BULLYING

Cyber bullying can be as simple as continuously sending emails or text messages, harassing someone in response to some action. It may also include public actions such as repeated threats, sexual remarks, pejorative labels (i.e., hate speech) or false accusations, ganging up on a victim in online forums, posting false statements as fact aimed discrediting or humiliating a targeted person.

Nine different types of cyber bullying were identified from the literature [10]; these are listed below:

1. *Flooding* consists of the bully that takes control of the media so that the victim cannot post a message.
2. *Masquerade* involves the bully logging in to a website, or program using another user's name to either bully a victim or damage the victim's reputation.
3. *Flaming*, or *bashing*, involves more than two users attacking each other on a personal level. The conversation consists of a heated, short lived argument, and there is bullying language in all of the users' posts.
4. *Trolling*, also known as *baiting*, involves intentionally posting comments that disagree with other posts for the purpose of provoking a fight, even if the comments don't necessarily reflect the poster's actual opinion.
5. *Harassment* most closely mirrors traditional bullying with the stereotypical bully victim relationship. This type of cyber bullying involves repeatedly sending offensive messages to the victim over an extended period of time.
6. *Cyber stalking* and *cyber threats* involve sending messages that include threats of harm, are intimidating or very offensive, or involve extortion.
7. *Denigration* involves discussing about someone online. Writing vulgar or untrue rumors about someone to another user or posting them to a public community or website falls under denigration.

V. FEATURES DISTINGUISHING CYBER BULLYING VIDEOS

Table II lists the category of features used for detection of cyber bullying content on YouTube. Features on the basis of which cyber bullying content differentiated are user based content based and contextual. User based features are the properties of the profile & the behavior of user in social network and content based features are the properties of the text posted by users.

A. *User based features consists of metadata like:*

1. *Subscribers:* Cyber bullying promoters have less number of subscribers.
2. *Subscriptions:* Cyber bullying promoters tend to have large number of subscriptions.
3. *Reputation:* It is the ratio of subscribers to the sum of subscribers and subscriptions. Cyber bullying promoters have very less reputation.
4. *Age of account:* It is obtained from current date and account creation date. Cyber bullying promoters have generally new accounts so this feature has less value for cyber bullying promoters.

TABLE II. FEATURES USED FOR DETECTION OF CYBER BULLYING CONTENT

User based features	Include demographic features like user profile details, number of subscribers, number of subscriptions, reputation, age of account, etc [4] [8]
Content based features	Include linguistic features, temporal features, popularity features and category feature [1] [4] [7] [9]
Contextual features	Include similarity feature and contextual post feature [2] [9]

B. Content based features divide this metadata into four main categories:

1. *Linguistic Features:*

a. *Percentage of X-Terms present in Title and Description [1]:* The presence of certain X-terms (from Bag-of-words) in title and description is a good indicator to recognize cyber bullying videos.

b. *Percentage of People Type present in Video's Title and Description [1]:* Presence of people type (like boy, girl, student, senior, fresher etc.) along with some cyber bullying terms in title and description of the videos show the presence of cyber bullying related actions in the video content.

2. *Popularity Features:*

a. *Ratio of number of likes by numbers of views [1]:* Number of likes on cyber bullying videos are less as compared to number of views.

b. *Ratio of number of comments by numbers of views [1]:* Usually, due to the privacy reason, users don't post frequent comments on positive class videos resulting very less number of comments in comparison to number of views.

3. *Temporal Features [1]:* Temporal based feature is related to time, like uploading time of video, time of comment and duration of video.

4. *Category Feature [1]:* These are the user defined category of video, like entertainment, sports, music, news, education etc.

C. Contextual features are divided in two categories:

1. *Similarity Feature [9]:* In this feature, each post compared with its k neighbors, also compared it with the first post of a comments thread. This is based on the heuristic that the first post defines the topic of the thread. Additionally, it also compared with the average information of a thread. Posts which are different from the thread average have the potential to be cyber bullying positive.

2. *Contextual Post Feature [9]:* In this feature, each post is represented as the vector sum of its neighboring posts. This feature was defined based on the assumption that there will be a reaction in posts which are near a cyber bullying post. So, the cluster of posts near a cyber bullying post should look different from the cluster of posts which are near normal posts.

VI. APPROACHES FOR CYBER BULLYING DETECTION

Different techniques have been used by researchers to find out the cyber bullying in various online social networks. We are focusing on work that has been done to identify cyber bullying in video social networks & few popular online social networks. Table III shows the summary of some papers reviewed regarding the detection of cyber bullying in these social networks.

Significant work has been done by Nisha Aggarwal et. al. [1] in year 2014, which used one class classifier approach and characterization study on series of sub-problems: vulgar video detection, abuse and violence in public places and ragging video detection in school and colleges to identify privacy invading harassment and misdemeanor videos by mining YouTube video metadata. The result indicates that identified content based discriminatory features can be used to exploit the harassment detection on YouTube up to a reasonable accuracy. Whereas, linguistic features and temporal based features seems to be more influential for accuracy.

Vidushi Chaudhary et. al. [2] formulates the problem of video response spam detection as a one-class classification problem (a recognition task) and divides it into three sub-problems: promotional video recognition, pornographic or dirty video recognition and automated script or botnet uploader recognition. The empirical analysis for each classifier of sub-problems based on certain linguistic features, temporal features, popularity based features, time based features opt-out with the 80% accuracy showing only the metadata of the YouTube video is discriminatory enough to recognize spam on YouTube.

Maral Dadvar et. al. [4] utilizes comments feature for detection of cyber bullying in MySpace corpus with the help of content based and user based features. One limitation of this approach is the limited size of the dataset. A larger and more diverse dataset can help in automatic cyber bullying detection.

Close analysis of the language used in cyber bullying has been done by April Kontostathis et. al. [7] and extending his work by using supervised machine learning approach in cyber bullying detection.

It is much clearer that cyber bullying occurs via offensive messages posted on social media. Since, the textual contents on online social media are highly unstructured, informal, and often misspelled; Ying Chen et. al. [8] propose the Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users in online social media like YouTube

Identification of online harassment is feasible when Term Frequency Inverse Document Frequency (TFIDF) is supplemented with contextual feature attributes; this has been proved in the research by Dawei Yin et. al. [9].

There are several other approaches which are quite effective in detecting online cyber bullying, one of which follows conventional key frame based methods with statistical

TABLE III. TECHNIQUES USED FOR DETECTION OF CYBER BULLYING CONTENT

Author	Metrics/ Features Used	Methodology Used	Dataset Used	Results
Nisha Aggarwal et. al. [1]	content based	One class classification algorithm for Vulgar Video Detection (VVD), Violence and Abuse Video Detection in School & Colleges (VAVDS), Violence and Abuse Video Detection in Public Places (VAVDP) and Ragging Video Detection in School & Colleges (RVDC), individually.	For VVD 960 videos, For VAVDS 1256 videos, For VAVDP 1561 videos, For RVDC 1396 videos in YouTube	Overall accuracy for VVD, VAVDS, VAVDP and RVDC classifiers is 83%, 84% , 90% and 97% respectively
Vidushi Chaudhary et. al. [2]	Contextual	One class classification algorithm for pornographic video response detection (PVRD), botnet video response detection (BVRD) and promotional or commercial video response detection (CVRD), individually.	For PVRD 10018 videos, For BVRD 3389 videos, For CVRD 9256 videos in YouTube	Overall accuracy for experimental dataset is 80%
Maral Dadvar et. al. [4]	content based and user based	Supervised learning approach to train a classifier for detecting online harassment and Support Vector Machines (SVM) model in WEKA as classification tool.	100,000 randomly selected posts in MySpace	Accuracy measures for basic approach has 31% precision, 15% recall, 20% f-measure
April Kontostathis et. al. [7]	content based	Supervised machine learning called Essential Dimensions of LSI (EDLSI) approach for detection of cyber bullying.	Training dataset consists of 13,652 Formspring.me posts, Testing dataset consists of 10,482 unjudged posts	The average precision for the top 1000 documents with the highest scores is 67.1%
Ying Chen et. al. [8]	content based and user based	The Lexical Syntactic Feature (LSF) architecture to detect offensive content and identify potential offensive users.	Dataset includes comments from 2,175,474 distinct YouTube users	Achieves precision of 98.24% and recall of 94.34% in sentence offensive detection, as well as precision of 77.9% and recall of 77.8% in user offensive detection.
Dawei Yin et. al. [9]	content based and contextual	A supervised learning approach including Term Frequency Inverse Document Frequency (TFIDF) approaches.	Dataset includes 4,802 posts in Kongregate, 4,303 posts in Slashdot, 1,946 posts in MySpace.	Achieves precision of 35.2%, 32.1% and 41.7% for Kongregate, Slashdot, and MySpace respectively.

analysis of MPEG-4 motion vectors, proposed by Christian Jansohn et. al. [11]. And the other one proposed by Nilesh J.Uke et. al. [12] consists of segmentation phase, for extracting the key frames for nude images detection, and classification phase for segregation of objectionable video which will be marked porn or non-porn depending upon the judgment criteria.

VII. RESEARCH DIRECTIONS AND FUTURE CHALLENGES

During survey it became quite apparent that a lot of work has been done for detecting cyber bullying in video social networks. Still improvements can be made to get better detection rate by using a different technique and covering more and robust features as deciding parameter. So following are the few conclusions drawn from survey:

1. Since YouTube has millions of active users and this number is constantly increasing. And almost all the researchers have used small testing dataset to see the performance of their approach. So there is a need to increase the testing dataset to see the performance of any approach.

2. When utilizing the social media as training data, how to remove the noise in the tags and comments or how to handle the noise in the learning process is an important issue to tackle.

3. Need to develop an approach that can detect all kinds of cyber bullying.

VIII. CONCLUSION

Many approaches have been developed and used by various researchers to find out cyber bullying in different video social networks. From the study it can be concluded that most of the researches and techniques for cyber bullying, harassment and other objectionable content detection follow the classifier approach like SVM, One Class Classification and Best First Search. Present approach(s) divide single detection problem into sub-problems, so there is need to develop unique approach in order to detect all types of cyber bullying. On other hand, combining variety of features for detection of cyber bullying has shown better performance in terms of accuracy, precision, recall etc. as compared to using user based, content based or contextual features, individually.

IX. REFERENCES

(1) Nisha Aggarwal, Swati Agrawal, Ashish Sureka, "Mining YouTube Metadata for Detecting Privacy Invading Harassment and Misdemeanor Videos," Twelfth Annual International Conference on Privacy, Security and Trust (PST), IEEE, pp. 84 – 93, 2014.

(2) Vidushi Chaudhary, Ashish Sureka, "Contextual Feature Based One-Class Classifier Approach for Detecting Video Response Spam on YouTube," Eleventh Annual International Conference on Privacy, Security and Trust (PST), IEEE, pp. 195 – 204, 2013.

(3) Swati Agarwal, Ashish Sureka, "A Focused Crawler for Mining Hate and Extremism Promoting Users, Videos and Communities on YouTube," 25th ACM conference on Hypertext and social media, pp. 294-296, 2014.

(4) Maral Dadvar, Franciska de Jong, "Cyberbullying Detection; A Step Toward a Safer Internet Yard," 21st international conference companion on World Wide Web, ACM, pp. 121-126, 2012.

(5) Maral Dadvar, Dolf Trieschnigg, Roeland Ordelman, Franciska de Jong, "Improving Cyberbullying Detection with User Context," 35th European Conference on IR Research, Springer, pp. 693-696, 2013.

(6) Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, Sidharth Chhabra, "Mining YouTube to Discover Extremist Videos, Users and Hidden Communities," 6th Asia Information Retrieval Societies Conference, Springer, pp. 13-24, 2010.

(7) April Kontostathis, Kelly Reynolds, Andy Garron, "Detecting Cyberbullying: Query Terms and Techniques," 5th Annual ACM Web Science Conference, pp. 195-204, 2013.
(8) Ying Chen, Sencun Zhu, Yilu Zhou, Heng Xu, "Detecting Offensive Language in Social Media to Protect Adolescent Online Safety," ACM, 2012.
(9) Zhenzhen Xue, Dawei Yin, Liangjie Hong, Brian D. Davison, April Kontostathis, Lynne Edwards, "Detection of Harassment on Web 2.0," CAW2.0, 2009.
(10) Paridhi Singhal, Ashish Bansal "Improved Textual Cyberbullying Detection Using Data Mining", International Journal of Information and Computation Technology, pp.569-576, 2013.
(11) Christian Jansohn, Adrian Ulges, Thomas M. Breuel, "Detecting Pornographic Video Content by Combining Image Features with Motion Information," ACM, pp. 601-604, 2009.
(12) Nilesh J.Uke, Dr. Ravindra C. Thool, "Detecting Pornography on Web to Prevent Child Abuse – A Computer Vision Approach," International Journal of Scientific & Engineering Research, pp. 1-3, 2012.
(13) YouTube Community Guidelines, Available at: https://www.youtube.com/t/community_ guidelines, Last accessed on: 6th January, 2015 2:30 PM
(14) Harassment and cyber bullying, Available at: https://support.google.com/youtube/answer/ 2802268, Last accessed on: 6th January, 2015 3:00 PM