Mining Youtube Videos Metadata For Cyberbullying Detection

[1] Mr. Shivraj Sunil Marathe[2] Prof. Kavita P. Shirsat
[1] M. E. Student,[2] Assistant Professor
[1][2] Vidyalankar Institute of Technology, Wadala (E), Mumbai, India
[1] shivraj.marathe@vit.edu.in, [2] kavita.shirsat@vit.edu.in

***Abstract*— Recent years have witnessed the evolution of Web 2.0, which has drastically increased the volume of community-shared textual resources (posts, comments) and media resources (videos, images) over the web. Moreover, today the internet has become an effective communication platform for people. Because of which many cyberbullying content promoters have also been attracted towards online social networks. Online video sharing websites such as YouTube contain large number of videos and users promoting cyberbullying and harassment. Due to immense popularity, anonymity and fewer restrictions for publication, YouTube is misused by some users to promote cyberbullying and online harassment.**

Our research presents an approach to identify misdemeanor, harassment resulting in cyberbullying by mining the video metadata. We conduct a study on a training dataset obtained by extracting several videos metadata using YouTube API. We formulate the problem of identifying cyberbullying videos as a search problem and present Shark Search algorithm based approach for cyberbullying detection. We present the result using standard information retrieval metrics such as f-measure, precision and recall. The accuracy of the proposed solution on the sample dataset is 83.65%. Our result favors the requirement of several contextual meta-data like, terms present in the title of the videos, description and comments, video subscribers and likes, number of views, YouTube category, length of videos and content focus in cyberbullying detection.

***Index Terms*— Cyberbullying, Information Retrieval, Mining User Generated Content, Online Harassment Detection, Text Mining, YouTube, Video Sharing Sites**

I. INTRODUCTION & MOTIVATION

With the proliferation of the Internet, online security of an individual has become an important concern. While Web 2.0 provides easy, interactive, anytime and anywhere access to the online communities, it also provides an avenue for cybercrimes like cyberbullying.

Now-a-days, online social networking websites such as Facebook, Twitter, YouTube, Instagram, Flickr etc. are experiencing a huge growth in popularity. In particular, video content is becoming a most important part of user's daily life. By allowing users to generate and distribute their own multimedia content to public, the Web has transformed into a major platform for the delivery of multimedia supporting different types of interactions among users such as discussions, debates, educational tips, how-to's etc.

YouTube is one of the most popular and widely used video sharing websites which allows users to publicize and share their independently generated content over the internet. Research shows that YouTube has become a convenient platform for many users to share offensive and malicious information and promote their ideologies. This led YouTube to become a repository of large amounts of malicious and offensive videos. For example, harassment And insulting videos [9], video spam [2], pornographic content [2] [12], hate and extremism promoting videos [6].

Despite several community guidelines[1] and administrative efforts made by YouTube, an old troubling problem with a new face, i. e. cyber bullying has found its way into the web.

Cyber bullying is defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact, repeatedly and over time, against a victim who cannot easily defend himself or herself [4]. Initially, though cyberbullying may not seen to cause any physical damage, but there are some potentially disturbing examples like bullied person has gone through depression, low self-esteem, suicide ideation, and even suicide [15]. Hannah Smith, a 14-year-old, hanged herself after negative comments were posted on her Ask.fm page, a popular social network among teenagers[2]. According to

Cyberbullying Research Center[3], cyberbullying was a contributing factor in her death.

Detection of cyberbullying and finding preventive measures are the main areas of focus in combating cyberbullying. In the context of YouTube cyberbullying can be an unauthorized shooting or uploading negative video of a claimant on website. The aim of this paper is to counter and combat cyberbullying and misdemeanor activities on YouTube, since the instances of cyberbullying on YouTube have become an increasing concern. YouTube is a dynamic website and hence identifying cyberbullying content on YouTube is a challenging problem. Therefore our work presented in this paper is motivated by the following facts:

a) YouTube's popularity, anonymity and low publication barriers allow users to upload cyberbullying and misdemeanor promoting content.
b) Limitations in solution capabilities being used by YouTube for cyberbullying detection.
c) Developing a systematic framework for automatic identification and to combat and counter cyberbullying videos on YouTube.
d) The research aim of the work presented in this paper is following:
e) The identification and characterization of such videos and users, promoting cyberbullying (Focus of this paper) on YouTube.
f) To investigate the effectiveness a Shark Search algorithm based approach for detecting YouTube videos and users promoting cyberbullying. Our aim is to examine the significance of the proposed approach.
g) To investigate the effectiveness of several contextual features in cyberbullying detection.
h) To discover users and/or communities playing central role in spreading cyberbullying and online harassment.

II. RELATED WORK & RESEARCH CONTRIBUTIONS

We conduct a literature survey in the area of cyberbullying, offense, online harassment and abusive content detection on popular social networking websites. However, based on our review of existing literature, we conclude that most of the researches for cyberbullying detection are performed in the field of mining media like, images & video frames and user generated content like, comments & messages.

For identifying privacy invasion and misdemeanor on YouTube, Nisha Aggarwal et. al. [1] proposed one class classifier approach and performed a characterization study on several sub problems: detection of vulgar video, abuse & violence in public places and detecting videos of ragging in school and colleges.

Vidushi Chaudhary et. al. [2] has recognized promotional videos, pornographic or dirty videos and automated scripts or botnet responses in YouTube corpus by formulating the video response spam detection problem as a one-class classification problem.

For detecting cyberbullying in MySpace corpus, Maral Dadvar et. al. [4] investigated a gender-specific text classifier approach. They have utilized content based and user based features along with Vector Machine model for training text classifier using WEKA.

Analysis of the language used for bullying has been done by April Kontostathis et. al. [7] and the research has extended by using supervised machine learning approach on labeled data, in conjunction with techniques provided by the WEKA Tool Kit to train the computer to recognize cyberbullying content.

Ying Chen et al. [8] investigated existing text mining methods for detection of offensive contents to protect adolescent's online safety, using proposed Lexical Syntactic Feature (LSF) architecture. Dawei Yin et. al. proposed supervised learning approach for detecting harassment. They determined that identification of online harassment is feasible when Term Frequency Inverse Document Frequency (TFIDF) is supplemented with N-gram and contextual feature attributes [9].

Jun-Ming Xu et. al. introduced social media as a large-scale, near real-time, dynamic data source for the study of bullying. They formulated cyberbullying detection as Natural Language Processing (NLP) tasks [17].

In a recent study on cyberbullying detection, Homa Hosseinmardi et. al. [16] investigates approaches for automatic detection of cyberbullying over Instagram. They device Naïve Bayes and linear SVM classifiers on a sample data set from Instagram consisting of manually labeled images and their associated comments.

Lots of previous work in cyber bullying detection has mostly concentrated on the media key frame based analysis & related solutions. Christian Jansohn et. al. [11] proposed conventional key frame based methods with statistical analysis of MPEG-4 motion vectors. Whereas, Nilesh J. Ukeet. al. [12] proposed an approach which consists of segmentation and classification phases for extracting the key frames in nude images, and segregation of objectionable videos, respectively. Later, the videos were marked as porn or non-porn depending upon the judgment criteria.

In context to existing work, the study presented in this paper makes the following unique contributions:

a) In comparison to previous work, the work presented in this paper is first step towards application of Shark Search algorithm based approach for detecting cyberbullying content on YouTube, using video metadata and contextual features.

b) We conduct a series of experiments on real-world data fetched from YouTube to demonstrate the effectiveness of the proposed solution.

c) We perform a characterization study of the cyberbullying promoting videos based on terms present in the video title & description, YouTube category, content focus and average length of videos.

III. PROPOSED SOLUTION APPROACH

Our goal is to identify cyber bullying (intentional or un-intentional) on YouTube, one of the most popular video sharing website today. We purpose a mechanism for identifying videos and users promoting cyber bullying. This proposed system incorporates a set of video attributes, discriminatory features and classification algorithm. To achieve our goal, we collected required dataset from YouTube. We evaluate the effectiveness of proposed approach using a test dataset, which was then built from a sample of the collected data.

Our proposed framework is described in section A, whereas; section B presents implementation details.

A. Proposed System

Fig. 1 presents a general research framework for proposed approach. As shown in Fig. 1, the proposed approach is a multi-step process consisting of three phases; namely, training & testing profiles collection, dynamic model building and implementation based on Shark Search algorithm.

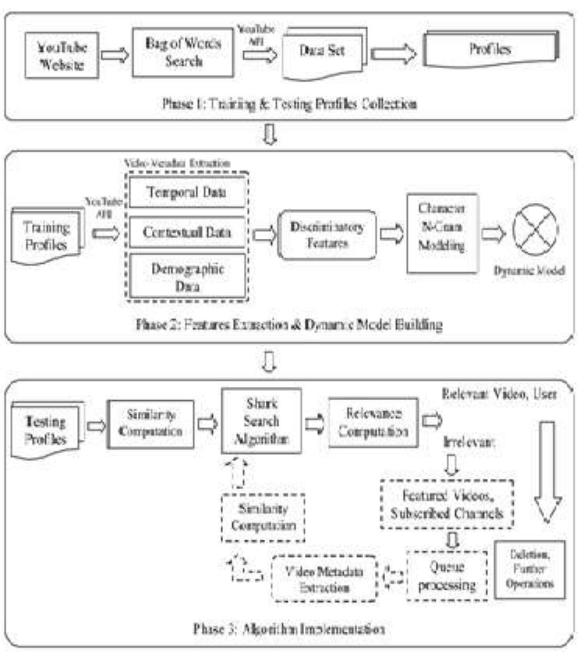

Fig. 1. General Research Framework for Proposed Solution Approach

In phase 1, we first perform a manual analysis along with visual inspection of YouTube videos and its contextual metadata. Using YouTube API[4], we build & collect our training dataset by retrieving all the available meta-data of several relevant (positive class) videos. In the training dataset, we observe several terms relevant to cyberbullying which contributes to next step of characterization and identification of discriminatory features. The retrieved meta-data of training dataset serve as a base for various discriminatory features, like, temporal and popularity based features (no. of subscribers, likes, dislikes, views and comments posted in response to the video), linguistic features (title and description of the video) and time based features (duration of video, upload time-stamp).

In phase 2, we build a dynamic model from these training profiles. For this purpose, we use character n-gram based approach, as it does not require extensive language specific pre-processing.

In phase 3, we build a system based on Shark Search algorithm which is a recursive process operates on testing dataset. It takes YouTube video as a seed and finds the textual similarity between seed video meta-data and training data. We implement a binary classifier to classify a video as relevant or irrelevant. A video is said to be relevant i.e. cyber bullying promoting video if its computation score is above the predicted threshold. Irrespective of the relevance, we further extend video frontiers like links to other YouTube videos; subscribers of the channel; featured channels or videos, suggested or related videos. We extract these frontiers by using HTML parser library[5] and YouTube API features.

B. Solution Implementation

In this section, we present the methodology and solution implementation details for the general research framework articulated in the previous section. In proposed solution we use Shark Search Algorithm (SSA). The goal of SSA is to first classify a video to be relevant (positive class) or irrelevant (negative class) and then explore the frontiers of both types of videos.

Inputs to this algorithm are seed (a video) *U*, threshold *th* for classification, n-gram value *Ng* for similarity computation. We compare each training profile meta-data with all n-gram values in bag-of-words collected using standard dictionary[6]and compute their similarity score for each seed in testing dataset.

C. Shark Search Algorithm

We propose a Shark Search Algorithm (Algorithm 1), which is an adaptive version of the algorithm introduced in

[6] http://www.noswearing.com/dictionary

[1]. The proposed method (Algorithm 1) explores frontiers of both relevant and irrelevant videos to seed input. Steps 1 and 2 extract all contextual features for training profiles using Algorithm 2 and build a training data set. Algorithm SSA is a recursive function which takes *U* as a seed input. Steps 3 and 4 extract all features for seed user *U* and compute its similarity score with training profiles using character n-gram and probability of maximum likelihood. Steps 5 to 8 represent the classification procedure and labeling of videos as relevant or irrelevant.

Steps 10 to 14 extract frontiers of a user channel using Algorithm 3 and repeats steps 3 to 14 for each linked video.

D. Features Extraction

In Algorithm 2, we retrieve contextual metadata of a YouTube user channel and video using YouTube API. Step 1 extracts the profile summary of the user. Steps 2 to 5 extract the titles of videos uploaded, commented, shared and marked favorite. The result of this algorithm can be stored in a text file containing all video titles and user profile information.

E. Frontiers Extraction

In Algorithm 3, we extract all external links of a YouTube video to other YouTube videos. These links could be the subscribers, featured videos or channels, and related videos. YouTube API does not allow users to retrieve the information of other users which is why we use HTML parser library to fetch all frontiers.

```
Data: Seed Video U, Threshold th, N-gram N_g
Result: List of Relevant and Irrelevant Videos
1.  for all u Ɛ U do
2.      D.add(ExtractFeatures(u))
    end
   Algorithm SSA (U)
   3. videofeeds U_f ← ExtractFeatures(U)
4. score s ← LikelihoodProbability(D, U_f, N_g)
5.    if (s < th) then
6. U.newclass ← Irrelevant
7.    else
8. U.newclass ← Relevant
      end
9. Hashmap U_sorted.InsertionSort(U, s)
10.   for all U_g do
11.   fr = Extract_Frontiers(U_g)
12.   Hashmap U_crawler.add(fr)
      end
13.   for all U_fr Ɛ U_crawler do
14.   SSA(U_fr)
      end
    end
```

Algorithm 1. Shark Search Algorithm

```
Data: User Video u
Result: Video Information
   Algorithm ExtractFeatures(U)
1. u_Profile ← u.getSummary()
2. u_Uploaded ← u.getUploadedVideo()
3. u_Commented ← u.getCommentedVideo()
4. u_Shared ← u.getSharedVideo()
5. u_favorited ← u.getFavoritedVideo()
```

Algorithm 2. Features Extraction Algorithm

```
Data: User Video u
Result: Frontiers of a Video
   Algorithm Extract_Frontiers(U)
1. u_subs ← u.getSubscribers()
2. u_fc ← u.getFeaturedChannels()
3. u_rv ← u.getRelatedVideos()
```

Algorithm 3. Frontiers Extraction Algorithm

IV. ANALYSIS & PERFORMANCE EVALUATION

In this section we present the characterization and empirical analysis of cyberbullying videos. We describe the experiments and analysis set up, calculate performance and the effectiveness of our proposed solution approach.

A. Experimental Dataset

1. Training Dataset

Our proposed solution needs to classify a given video is relevant or not with respect to cyber bullying. The SSA requires sample documents or training dataset to learn the specific characteristics and properties of videos, users promoting cyber bullying. We perform a survey and manual analysis on YouTube and query for several cyber bullying and harassment keywords. Table I shows 41 keywords from a list used for building training profiles. These keywords help to collect 997 videos using YouTube API. Initially we collect 151 relevant videos, and later we extract 846 videos. Hence in total we collect a testing data set of 997 videos for cyber bullying detection. We make sure that there is no redundancy in the training dataset. We identify discriminatory features from videos of training dataset. We believe that the discriminatory features of such videos reacts user interests and can be used for building a predictive model.

2. Test Dataset

We create a test dataset of 997 videos by extracting the positive as well as negative class videos on YouTube. Table II shows the size of training and test dataset we collected for cyber bullying detection.

Our training dataset includes positive class videos and the test dataset includes both positive and negative class videos. Therefore, the size of training dataset is smaller than the test dataset. We annotate the dataset and label each video as relevant or irrelevant.

TABLE I: A SAMPLE LIST OF KEYWORDS PRESENT IN CYBERBULLYING AND MISDEMEANOR PROMOTING VIDEOS

Terms	People Type	Examples of Video Title
hot, private, removing, MMS, sexy, kiss, nipple, breast, ass, removing, porn, boob, hottest, naked, sex, f**k, smooching, secret, sexy, seduce, lovers, scandal, kissing, sexually, harassed	girls, girl, boy, gay, student, boyfriend, female, girlfriend, guys, lover, classmate, men, ladies, aunty, people, student, couple	CCTV footage Girl sexually harassed in metro MMS of girl in car MMS Kand in School
Fights, fighting, fight, brutally, violence, killed, beats, domestic, beaten, mess, scolding, beaten	Student, kids boy, girl, girls gay, dudes, boy, student, teacher, friends, police, aunty, man, women, people, women, girls	Teacher Slaps Student In Class BRUTAL BEATING High School Boy VICIOUSLY BEATS A Girl In The Hallway
Ragging, ragged, horrible, shocking	Girl, seniors, student, juniors, students, junior, senior, fresher	Bully Ragging An Innocent Boy.wmv SHOCKING VIDEO OF RAGGING EMERGES

TABLE II: SIZE OF THE EXPERIMENTAL DATASET

Training Dataset	Testing Dataset
151	997

TABLE III: CONFUSION MATRIX

		Predicted	
		Relevant	Irrelevant
Actual	Relevant	187	78
	Irrelevant	85	647

TABLE IV: PERFORMANCE RESULTS

FPR	TNR	Precision	Recall	F-Score	Accuracy
0.1161	0.8838	0.6875	0.7056	0.6965	0.8365

B. Evaluation Metric

To evaluate the effectiveness of the proposed solution approach, we have used a standard confusion matrix with each column of the matrix representing the predicted class instances while each row of the matrix representing the actual class instances. Each position in the confusion matrix represents the number of elements belonging to that particular class.

Table III shows the confusion matrix for proposed solution approach. We execute SSA classifier for test dataset of 997 videos and it classifies 272 (187+85) videos as relevant and 725 (78+647) videos as irrelevant. Table III reveals that 85 and 78 videos are misclassified as relevant and irrelevant respectively. There are several reasons of this misclassification. Few of them are listed below:

Presence of noisy data such as, misleading information, misspelled words and lack of information. For example, “Brutal pelea campal todos contra todos Brutal pitched battle all against all” video is misclassified as irrelevant because of noisy data. The video “Sexual Harassment” is misclassified as relevant because of lack of information.

a) Presence of commercial, news and advertising videos on YouTube. For example, a video titled “Workplace Issue Sexual Harassment” is posted on YouTube for the public awareness but due to the presence of terms sexual and harassment in title, this video is misclassified as a positive class video.

b) We evaluate the performance of our proposed solution approach in terms of False Positive Rate (FPR), True Negative Rate (TNR), Precision, Recall, F-score and Accuracy. Table IV shows the performance evaluation of our proposed solution approach. Table IV reveals that overall accuracy for cyberbullying detection is 83.65%.

C. Empirical Analysis

We identify all discriminatory contextual features to detect cyberbullying promoting videos, users on YouTube. We characterize each video by its meta-data. This set of contextual features is divided into 4 categories: linguistic, YouTube basic, temporal and popularity based, and time based features.

a) Linguistic features: We hypothesize that percentage of cyberbullying terms in title (PCTT) and description (PCTD) is an indicator to recognize relevant videos. Our hypothesis is based on the observation that more than 72% positive class videos contain some cyberbullying and misdemeanor terms in title while more than 69% negative class videos contain negligible amount of relevant terms in title (Fig. 2) and around 76% relevant videos contain some cyberbullying terms present in description while around 52% negative class videos does not contain any relevant terms in description of that video (Fig. 3) which shows the discriminatory behaviour of the feature. We use a standard dictionary of cyberbullying & negative words to match predefined terms in title and description of video.

YouTube Basic Features: Category of the video (CV) is

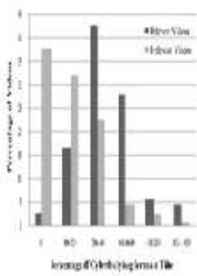
Fig. 2. PCTT

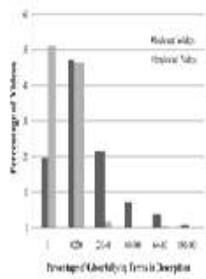
Fig. 3. PCTD

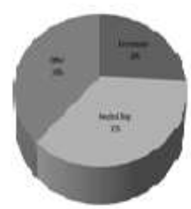
Fig. 4. CV

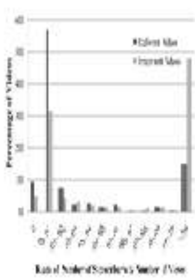
Fig. 5. RSBV

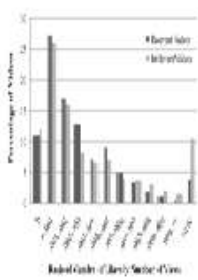
Fig. 6. RLBV

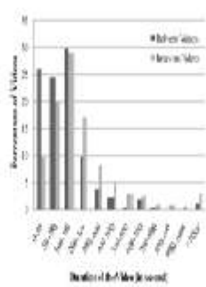
Fig. 7. DYTV

the feature which shows the type of the video (education, entertainment, news etc.) to which it belongs. A visual inspection of multiple cyber bullying videos across their category shows the discriminatory behaviour of the feature as out of total 32 YouTube video categories, 26% cyber bullying videos fall under the category entertainment and 35% cyber bullying videos are of category people & blogs (Fig. 4).

a) Temporal and Popularity Based Features: Change in features value with time shows the popularity of the videos on YouTube. Features like number of subscribers, likes, dislikes and views comes under the category Temporal and Popularity. We fetch the number of subscribers, likes and views of each video response present in our training dataset and compute the ratio of number of subscribers by number of views (RSBV) and number of likes by number of views (RLBV). These values can be a good indicator to detect cyberbullying videos. We hypothesize that low value of RSBV and RLBV signals pornographic behaviour. We confirm the effectiveness of this phenomenon in the evaluation dataset wherein the RSBV and RLBV exhibit a low value as around 70% positive class videos have RSBV value less than 0.01 while around 65% negative class videos have RSBV value greater than 0.01 (Fig. 5) and around 40% positive class videos have RLBV value less than 0.001 while 62% negative class videos have RLBV value greater than 0.001 (Fig. 6).

b) Time Based Feature: We observe the pattern of duration of multiple YouTube videos (DYTV). It clearly shows that duration of the video is a good indicator for relevant video detection as around 26% cyberbullying videos have duration less than 50 seconds and more than 51% videos have duration less than 100 seconds while around 71% irrelevant videos have duration greater than 100 seconds (Fig. 7).

V. CONCLUSION

Cyber bullying is a serious problem in online social networks and becoming a major threat to teenagers and adolescents. In this paper, we presented Shark Search algorithm based classification approach for automatic identification of users, videos promoting cyber bullying on YouTube. Our findings and performance evaluation result reveals that, the proposed solution approach correctly able to identify cyber bullying promoters with 83.65% accuracy.

Our results showed that, incorporation of various discriminatory features like linguistic features, popularity based features, temporal features, time based features and other reliable contextual meta-data significantly improves cyber bullying and misdemeanor detection accuracy.

FUTURE WORK

In future stages this work could be extended by considering performance of our proposed system on larger and more diverse training & testing dataset. One future direction along the proposed line of research could be to find the performance of text classifier considering language independence. Additionally, further future work requires an investigation of techniques for cyberbullying videos that do not contain any relevant meta-data and on which text classification cannot be applied.

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]. Sara Owsley Sood, Elizabeth F. Churchill, Judd Antin, "Automatic Identification of Personal Insults on Social News Sites," Journal of the American Society for Information Science and Technology, ACM, pp. 270-285, 2012.

[14]. Laura P. Del Bosque, Sara E. Garza, "Aggressive text detection for cyberbullying," 13th Mexican International Conference on Artificial Intelligence, Springer, pp. 221-232, 2014.

[15]. Hinduja, S., and Patchin J. W., "Cyberbullying research summary, cyberbullying and suicide," 2010.

[16]. Homa Hosseinmardi, Sabrina Arredondo Mattson, Rahat Ibn Rafiq, Richard Han, Qin Lv, Shivakant Mishra, "Detection of Cyberbullying Incidents on the Instagram Social Network," Association for the Advancement of Artificial Intelligence, ARXIV, 2015.

[17]. Jun-Ming Xu, Kwang-Sung Jun, Xiaojin Zhu, Amy Bellmore, "Learning from bullying traces in social media," NAACL HLT '12 Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, ACM, pp. 656-666, 2012.