Contextual Features Based Naïve Bayes Classifier for Cyberbullying Detection on YouTube

[1]Shivraj Sunil Marathe, [2]Kavita P. Shirsat

Vidyalankar Institute of Technology, Wadala (E), Mumbai, India

Abstract— With the growth of the Web 2.0, online communication and video sharing websites has started emerging. This evolution on internet is now allowing users to share their information and collaborate with each other easily. In addition, video sharing websites are helping users to establish new connections between people and promote their views, ideas, etc. As a result, various malignant users are getting attracted towards these social networks. Among several video sharing websites (with social networking features), YouTube is the most popular & widely used website. Due to anonymity of content uploaded and low publication barriers, YouTube is misused by some users as a platform to post videos promoting cyberbullying, harassment and online abuse.

In this research work, we employ a Naïve Bayes Classifier to identify cyberbullying videos, users on YouTube by mining video metadata. We frame the problem of YouTube cyberbullying detection as a search problem. We conduct study of training dataset by downloading considerable videos & related discriminatory features using YouTube API. Our evaluation of performance results on test dataset reveal that accuracy of proposed approach is more than 67% which significantly demonstrate the effectiveness of the proposed approach.

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

I. INTRODUCTION

The web 2.0 has built up enormously which led to evolution of search engines, social networking sites, video sharing and photo sharing websites. Specially social networking websites such as Facebook, Twitter, YouTube, Flickr, Instagram have increased a lot since last few years which specializes in discussion forums, micro-blogging, and multimedia sharing.

YouTube is one of the most popular and widely used video sharing website over the Internet. YouTube provide us several social networking functionalities such as to like or dislike a video, make video as favorite, post a textual comment, share a video, subscribe channel etc. YouTube statistics[1] states that, every month 1 billion unique users visit YouTube and over 6 billion hours of videos are watched. Since YouTube has very low publication barriers, anonymity of content uploader provide a safer environment for user. There are increasing evidences that YouTube have been used by people for uploading offensive and malicious contents. For example, harassment and insulting videos [9], video spam [2], pornographic content [2] [11], hate and extremist promoting videos [6]. Drawbacks of YouTube and high reachability of videos has led to an old troubling problem with a new face under new circumstances, i. e. cyberbullying. Traditionally bullying was considered to be a face to face encounter between people, but now it has also found its way into the web.

Cyberbullying is defined as an aggressive, intentional act carried out by a group or individual, using electronic forms of contact (e.g. email and chat rooms) repeatedly or over time against a victim who cannot easily defend himself or herself [4]. In the context of YouTube cyberbullying can be defined as unauthorized shooting or uploading negative video (e.g. vulgar, violence and abuse) on website. Public disclosure of such content can be said as harassment of the claimant. Cyberbullying can be of two types intentional or unintentional. Sometimes user posts a video on website with a motive to threat or disturb other users. For example, an abusive or 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 the person in the video.

Presence of cyberbullying promoting videos degrades the reputation of users who are involved in those videos. It also causes the bandwidth wastage for the users who are not willing to watch such videos. YouTube has its own community guidelines[2] available on website in order to prevent users from uploading inappropriate content. However, despite of these community guidelines YouTube has become a repository of cyberbullying and

[1] http://www.youtube.com/yt/press/statistics.html

[2] http://www.youtube.com/yt/policyandsafety/en-GB/communityguidelines.html

harassment promoting videos [1]. Detecting such videos and users is technically challenging problem. 100 hours of videos are uploaded every minute, which makes YouTube a very dynamic website. Metadata of these videos is a user generated data, which is available in free form text. That data is highly unstructured and can have lots of noisy content. So, identifying such videos by keyword based search is highly impractical. Therefore our work presented in this paper is motivated by the fact that, YouTube's popularity, anonymity and low publication barriers allow users to upload cyberbullying and harassment promoting content.

The research aim of the work presented in this paper is to investigate the effectiveness of contextual features based Naïve Bayes Classifier approach for detecting YouTube videos and users promoting cyberbullying.

II. RELATED WORK & RESEARCH CONTRIBUTIONS

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

Nisha Aggarwal et. al. propose one class classifier approach and perform a characterization study on vulgar video detection, abuse & 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 [1]. Vidushi Chaudhary et. al. formulates the problem of video response spam detection as a one-class classification problem for promotional video recognition, pornographic or dirty video recognition and automated script or botnet response recognition [2].

Maral Dadvar et. al. [4] utilizes content based and user based features for detection of cyberbullying in MySpace corpus, but one limitation of this approach was the limited size of the dataset. Analysis of the language used in cyberbullying has been done by April Kontostathis et. al. and he extended his work by using supervised machine learning approach in cyberbullying detection [7]. It is much clearer that offensive messages on social media lead to cyberbullying. Ying Chen et. al. detect offensive content and identify potential offensive users in social media like YouTube using proposed Lexical Syntactic Feature (LSF) architecture [8].

Dawei Yin et. al. proposed that identification of online harassment is feasible when Term Frequency Inverse Document Frequency (TFIDF) is supplemented with contextual feature attributes [9].

There are several other approaches which stand quite effective in online cyber bullying detection, one of which proposed by Nilesh J. Uke et. al. consists of segmentation and classification phases for extracting the key frames in nude images, segregation of objectionable videos, respectively. The videos were marked as porn or non-porn depending upon the judgment criteria [11].

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

Homa Hosseinmardi et. al. investigate approaches for understanding and automatic detection of cyberbullying over images in media-based mobile social network, Instagram. They device two classifiers, Naïve Bayes and linear SVM classifier separately on a sample Instagram data set consisting of manually labeled images and their associated comments [14].

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

a) In comparison to previous work, the work presented in this paper makes its own contribution towards detection of cyberbullying content on YouTube using video metadata and contextual features. We develop a Naïve Bayes classifier based approach for detecting videos based on video's contextual features.

b) We conduct a series of experiments and perform an analysis on real world dataset (fetched from YouTube) to evaluate the effectiveness of the proposed system.

III. PROPOSED SOLUTION APPROACH

Our goal is to design a mechanism for identification of videos and users promoting cyberbullying, using a set of discriminatory features and classification algorithm. To achieve this goal, we collected required dataset from YouTube. The effectiveness of our proposed approach is evaluated using a test dataset, which was then built from a sample of the collected data. Section 3.1 describes our proposed framework, whereas Section 3.2 presents the solution implementation.

A. Proposed System

Fig. 1 represents a research framework for the proposed approach. As shown in Fig. 1, our proposed solution is a multi-step process primarily consists of three phases; cited as, training & testing profiles collection, dynamic model building, and an implementation based on Naïve Bayes algorithm.

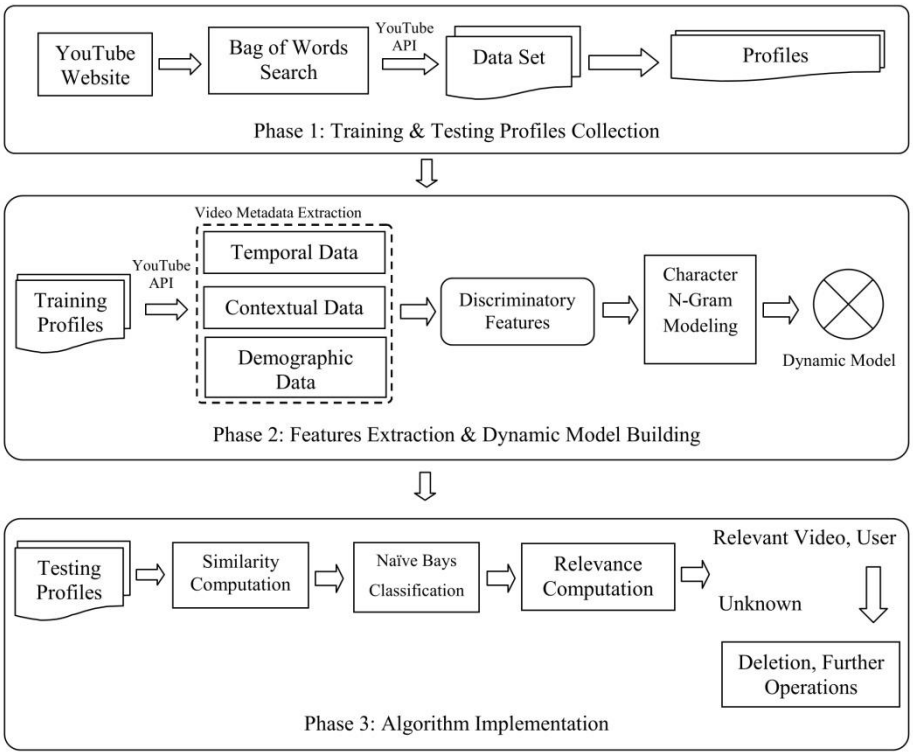

Fig. 1. General Research Framework for Proposed Solution Approach

In phase 1, we perform manual analysis and inspection on various YouTube videos and its contextual metadata. For the purpose of training the classifier we collect positive class training dataset (promoting cyberbullying). Using YouTube API[3], we download all the available meta-data of several relevant (positive class) videos and their related videos. We build our training dataset by extracting this meta-data. In the training dataset, we observe several terms relevant to cyberbullying for the purpose of identifying discriminatory features, such as: temporal and popularity based features (number of subscribers, likes, dislikes, views and comments), linguistic features (title and description of the video), and times based features (duration, time-stamp).

In phase 2, we use character n-gram based approach to build a dynamic model from these training profiles.

In phase 3, we build a system based on Naïve Bayes algorithm. It is a simple probabilistic model based on Bayes rule for text classifications. It takes one YouTube video as an input, finds an extent of textual similarity between this video metadata and training data. Based on the probability score a video is classified as relevant i.e. cyberbullying promoting or irrelevant.

B. Solution Implementation

In this section, we present the methodology and solution implementation details for the proposed classification algorithm. In proposed method we use Naïve Bayes algorithm to classify a video to be relevant (positive class) or irrelevant (negative class).

Inputs to this algorithm are a seed (a video) u, n-gram value Ng, bag-of-words. We compare each training profile with all positive & negative n-gram values in bag-of-words and compute their likelihood score for each seed.

1) Naïve Bayes Algorithm

The Naïve Bayes model involves a simple conditional independence assumption, i.e. given a class which may be positive or negative; the words are conditionally independent of each other. This assumption doesn't much affect the accuracy of text classification but makes really fast classification applicable for the problem.

In our case, the Maximum Likelihood Probability (MLP) of words x_i belonging to a particular class c is given by (1),

$$P\left(\frac{x_i}{c}\right) = \frac{\text{Count of } x_i \text{ in documents of class c}}{\text{Total number of words in documents of class c}} \quad (1)$$

We store the frequency counts of words in hash tables during the training phase itself. According to the Bayes Rule, the probability of a particular video u belonging to class c_i is given by (2),

[3] https://developers.google.com/youtube/

$$P\left(\frac{c_i}{u}\right) = \frac{P(c_i)\,P\left(\frac{u}{c_i}\right)}{P(u)} \qquad (2)$$

Steps 1 and 2 in the proposed method (Algorithm 1) extract all contextual features for training profiles using Algorithm 2 and build a training data set. Naïve Bayes Algorithm 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. Step 5 stores the frequency counts of the words in hash tables. Steps 6 to 9 represent the classification procedure and labeling of videos as relevant or irrelevant.

Data: Seed Video *u*, N-gram N_g, Bags-of-words, Class *c - pos, neg*
Result: List of Relevant and Irrelevant Videos
1. for all *u* Ɛ *U* do
2. *D.add(ExtractFeatures(u))*
end
Algorithm Naïve Bayes *(u)*
3. videofeeds $u_f \leftarrow$ *ExtractFeatures*(*u*)
4. score *s* ← *MLP(D, c, N_g)*
5. Hashmap U_{sorted}.*InsertionSort*(*u*, *s*)
6. if p(pos/u) ≥ p(neg/u) then
7. *u.newclass* ← Relevant
8. else
9. *U.newclass* ← Irrelevant
end
end

Algorithm 1. Naïve Bayes Classification Algorithm

2) Features Extraction

In Algorithm 2, we retrieve contextual metadata of a user 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 favourite. The result of this algorithm can be stored in a text file containing all video titles and user profile information.

Data: User Video *u*
Result: Video Information
Algorithm *ExtractFeatures*(*U*)
1. $u_{Profile}$ ← u.getSummary()
2. $u_{Uploads}$ ← u.getUploadedVideo()
3. $u_{Commented}$ ← u.getCommentedVideo()
4. u_{Shared} ← u.getSharedVideo()
5. $u_{favorited}$ ← u.getFavoritedVideo()

Algorithm 2. Features Extraction Algorithm

IV. MATHEMATICAL MODEL

Our proposed method is based on a multinomial model, in which attribute values are independent of each other for the particular class $\gamma(ß|\alpha) = \gamma(\omega_1 \ldots \omega_n|\alpha)$. In this approach a document is an ordered sequence of word events, drawn from the same vocabulary V. Since, the lengths of documents are independent of class; each document is drawn from a multinomial distribution of words with as many independent trials of the length. This yields the familiar bag-of-words representation for documents. The BOW model is commonly used in methods of document classification, where the (frequency of) occurrence of each word is used as a feature for training a classifier.

We consider W as the set of all terms or words (dictionary) that occur at least once in a collection of documents D. The BOW representation of document d is a vector of weights $(\omega_1 \ldots \omega_n)$. Here, unigram feature easily helps in finding presence or absence of a single word within a text. We estimate the conditional probability $(\omega|\alpha)$ using (3) as the relative frequency of term ω in documents belonging to a class including multiple occurrences of a term in a document.

$$(\omega|\alpha) = \frac{\text{count}(\omega|\alpha)+1}{\text{count}(\alpha)+|V|} \qquad (3)$$

Where count $(\omega|\alpha)$ is the number of occurrences of ω in training documents from class α, count (α) is the number of words in that class and |V| is the number of terms in the vocabulary.

Table I takes the example of video title. Naïve Bayes classifier classifies video as a relevant or irrelevant based on terms present in video title. We calculate priori probability of pos and neg by using (1).

$$\gamma\ (pos) = 3/4$$

$$\gamma\ (neg) = 1/4$$

Then we calculate maximum likelihood smoothing Naïve Bayes estimate by using (3).

γ (hits | pos) = (2+1) / (15+30) = 1/15 = 0.06

γ (hilarious | pos) = (1+1) / (15+30) = 2/45 = 0.044

γ (hostel | pos) = (0+1) / (15+30) =1/45 = 0.022

TABLE I: EXAMPLE

Dataset	Video Document	Keywords	Class
Training	1	teacher hits student in class	Pos
	2	man hits a women in a brutal fight	Pos
	3	hilarious ragging	Pos
	4	chennai hostel girls mp4	Neg

Testing	5	hilarious hostel bully hits sleeping friend then fat kid get drop kicked	?

γ (hits | neg) = (0+1) / (4+30) = 1/34 = 0.029

γ (hilarious | neg) = (0+1) / (4+30) = 1/34 = 0.029

γ (hostel | neg) = (1+1) / (4+30) = 1/17 = 0.058

Then we calculate posteriori probability for the video document d5,

γ (pos | d5) = 3/4 * 1/15 * 2/45 * 1/45 = 0.00004356

γ (neg | d5) = 1/4 * 1/34 * 1/34 * 1/17 = 0.0000121945

Here, γ (pos | d5) $\geq \gamma$ (neg | d5). Maximum value of γ (pos | d5) means probability of positive words in video title document d5 is maximum hence fifth video is relevant.

V. PERFORMANCE EVALUATION

In this section we discuss the experiments and analysis set up, calculate performance and the effectiveness of our proposed solution approach.

A. Experimental Dataset

Our proposed solution needs to classify a given video, user is relevant or not with context to cyberbullying. The Naïve Bayes classification algorithm requires exemplary training dataset to learn the specific properties of videos, users promoting cyberbullying. We perform a manual search on YouTube and query for several cyberbullying and harassment keywords. We collect a training data set of 80 videos for cyberbullying detection. We identify discriminatory features from videos of training dataset that shows user interests and can be used for building a proposed model.

We build a test dataset by extracting 965 random positive and negative class videos on YouTube. Table II shows the size of training and test dataset we select for cyberbullying detection.

TABLE II: SIZE OF THE EXPERIMENTAL DATASET

Training Dataset	Testing Dataset
80	965

Our training dataset includes positive class videos and the test dataset includes both positive and negative class videos. We annotate each video as relevant or irrelevant.

B. Evaluation Metric

To evaluate the effectiveness of the proposed solution approach, standard confusion matrix is used. Table III shows the confusion matrix with each column of the matrix representing instances of predicted class while each row of the matrix representing instances of actual class. We execute Naïve Bayes classifier on test dataset of 965 videos and it classifies 265 (196+69) videos as relevant and 700 (243+457) videos as irrelevant. Table II reveals that considerable amount videos are misclassified as relevant and irrelevant respectively. The reasons of misclassification can be presence of noisy data (misleading information, misspelled words and lack of information), ambiguity in the title or description of the video, presence of commercial and advertising videos on YouTube.

We evaluate the performance of our proposed solution approach in terms of precision, recall, accuracy and f-score.

TABLE III: CONFUSION MATRIX

		Predicted	
		Relevant	Irrelevant
Actual	Relevant	196 (α_1)	69 (β_1)
	Irrelevant	243 (γ_1)	457 (δ_1)

a) Precision is the proportion of predicted relevant videos that were correct, calculated using the equation:

$$\text{Precision} = \alpha_1 / (\alpha_1+\gamma_1) \quad (4)$$

b) Recall is the proportion of relevant videos that were correctly identified, calculated using the equation:

$$\text{Recall} = \alpha_1 / (\alpha_1+\beta_1) \quad (5)$$

c) Accuracy is the proportion of the total number of predictions that were correct, calculated using the equation:

$$\text{Accuracy} = (\alpha_1+ \delta_1) / (\alpha_1+\beta_1+\gamma_1+\delta_1) \quad (6)$$

d) F-Score is the weighted harmonic mean between precision and Recall, calculated using the equation:

$$\text{F-Score} = \text{Precision+Recall} / 2 \quad (7)$$

TABLE IV: PERFORMANCE RESULTS

Precision	Recall	F-Score	Accuracy
0.4464	0.7396	0.5930	0.6766

Table IV shows the performance results of our proposed solution approach. Table IV reveals that overall accuracy for cyberbullying detection is 67.66%

VI. CONCLUSION

We present an approach based on Naive Bayes classification algorithm to detect users, videos promoting cyberbullying on YouTube. Our experimental results reveal that the proposed solution approach correctly able to identify users, videos promoting cyberbullying with more than 67% accuracy. This proves that, various discriminatory features like linguistic features (presence of k-terms in the title, description and comments of the videos), popularity based features, temporal features, category of videos and other reliable contextual meta-data can be significantly used for automatically recognizing cyberbullying on YouTube.

VII. FUTURE WORK

In future studies, evaluating the performance of our proposed system with large number of training & testing dataset could be taken into account. Some cyberbullying videos do not contain any relevant meta-data. So, further future work requires investigating techniques for such cases in 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] 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.

[12] 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.

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

[14] 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.

[15] Shivraj Sunil Marathe, Prof. Kavita P. Shirsat, "Approaches for Mining YouTube Videos Metadata in Cyber bullying Detection," International Journal of Engineering Research & Technology (IJERT), pp. 680-684, 2015.

[16] K. Nalini, Dr. L. Jaba Sheela, "A survey on Datamining in Cyber Bullying," International Journal on Recent and Innovation Trends in

Computing and Communication, pp. 1865-1869, 2014.

[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.

[18] Samaneh Nadali, Masrah Azrifah Azmi Murad, Nurfadhlina Mohamad Sharef, Aida Mustapha,Somayeh Shojaee, "A review of cyberbullying detection: An overview," 13th International Conference on Intelligent Systems Design and Applications (ISDA), IEEE, pp. 325-330, 2013.