Mining YouTube Metadata for Detecting Privacy Invading Harassment and Misdemeanour Videos

Student Name: Nisha Aggarwal

IIIT-D-MTech-CS-IS-12-012

March 13, 2014

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Dr. Ashish Sureka (Chair)
Prof. Pravesh Biyani
Prof. Vikram Goyal

Submitted in partial fulfillment of the requirements
for the Degree of M.Tech. in Computer Science,
with specialization in Information Security

Keywords: YouTube, Privacy, Invading, Harassment, Misdemeanour, Social Media Analytics, One-Class Classification, User Generated Content(UGC), Vulgar, Ragging, Violence.

Certificate

This is to certify that the thesis titled **"Mining YouTube Metadata for Detecting Privacy Invading Harassment and Misdemeanour Videos"** submitted by **Nisha Aggarwal** for the partial fulfillment of the requirements for the degree of *Master of Technology* in *Computer Science & Engineering* is a record of the bonafide work carried out by her under my guidance and supervision in the Security and Privacy group at Indraprastha Institute of Information Technology, Delhi. This work has not been submitted anywhere else for the reward of any other degree.

Professor Ashish Sureka
Indraprastha Institute of Information Technology, New Delhi

Abstract

YouTube is one of the most popular and largest video sharing websites (with social networking features) on the Internet. A significant percentage of videos uploaded on YouTube contains objectionable content and violates YouTube community guidelines. YouTube contains several copyright violated videos, commercial spam, hate and extremism promoting videos, vulgar and pornographic material and privacy invading content. This is primarily due to the low publication barrier and anonymity. We present an approach to identify privacy invading harassment and misdemeanour videos by mining the video metadata. We divide the problem into sub-problems: vulgar video detection, abuse and violence in public places and ragging video detection in school and colleges. We conduct a characterization study on a training dataset by downloading several videos using YouTube API and manually annotating the dataset. We define several discriminatory features for recognizing the target class objects. We employ a one-class classifier approach to detect the objectionable video and frame the problem as a recognition problem. Our empirical analysis on test dataset reveals that linguistic features (presence of certain terms and people in the title and description of the main and related videos), popularity based, duration and category of videos can be used to predict the video type. We validate our hypothesis by conducting a series of experiments on evaluation dataset acquired from YouTube. Empirical results reveal that accuracy of proposed approach is more than 80% demonstrating the effectiveness of the approach.

Acknowledgments

First and foremost I would like to express my sincere gratitude to my advisor Dr. Ashish Sureka for his continuous support in my M.Tech study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. He always inspired me to do things perfectly. I respect the confidence and trust he showed in me. Without his encouragement, inspiration and guidance this thesis would not have been possible.

Besides my advisor, I would like to thank the rest of my thesis committee: Prof. Pravesh Biyani and Prof. Vikram Goyal, for their encouragement and support.

I would also like to thank my fellow mates Shilpi Jain, Swati Agrawal and Lovey Agrawal for their encouragement, insightful comments and suggestions.

In addition, I would like to thank my family for supporting and encouraging me throughout my life and believing me. A special thanks to my loving sister and my father who are always there for helping me and their blind trust on me.

Last but not the least, I am grateful to everyone who has directly or indirectly helped me to achieve this goal. No amount of words written on this page would be sufficient to quantify their advice, prayers, love and support for me. This thesis would never be successful without your support and love.

Contents

List of Figures

List of Tables

Chapter 1

Research Motivation and Aim

Due to the availability of the broadband high speed internet, Web 2.0, such as social-networking sites, video sharing sites, wikis and blogs have acquired a significant popularity in recent years [8]. As growth of the social media has been increased upto a peak level, people are spending significant amount of time on these websites[1]. Specifically some social networking sites such as Facebook[2], Twitter[3], YouTube[4], Flickr[5] have become the way of interaction among the people in the worldwide [4]. Therefore various malignant users have also been attracted to those sites. Today, Web 2.0 has become an effective communication platform for extremists to promote their ideas, share resources, and communicate among each other. YouTube, *Yahoo! Screen*[6], Vimeo[7], Dailymotion[8] are popular video sharing sites on web. Among them YouTube is the largest and most popular free video sharing site and has reached a level of ubiquity in the video-sharing market because of popularity of the smartphones (percentage of YouTube traffic from mobile phone is 40%)[9] and tablet.

1.0.1 YouTube

YouTube is one of the largest video sharing websites on the Internet. Other than this, Today YouTube is more popular than cable television[10]. YouTube reaches more adults than any cable network. In the United States, the number of people who watch television has fallen behind the number of people who watch YouTube on a regular basis. This makes it clear that televised content is undergoing a decline, online consumption of video is on the incline. The reason behind this is YouTube is available on phones, tablets, game consoles and smart TVs, allowing us to watch all our favorite videos on the go or on the best screen anywhere. Many companies

[1] http://www.jeffbullas.com/2014/01/17/20-social-media-facts-and-statistics-you-should-know-in-2014/

[2] http://www.facebook.com

[3] https://twitter.com

[4] https://www.youtube.com/

[5] https://www.flickr.com/

[6] https://screen.yahoo.com/

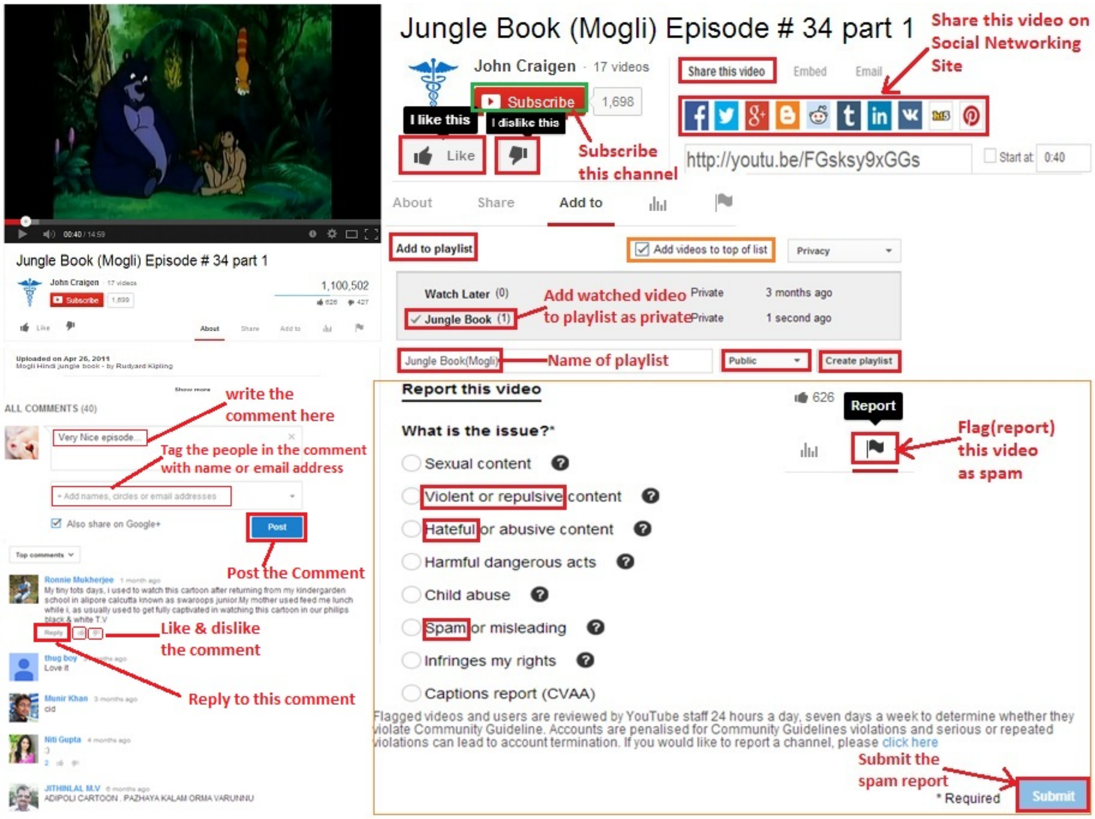
[7] https://vimeo.com/

[8] http://www.dailymotion.com/in

[9] http://expandedramblings.com/index.php/youtube-statistics/#.Uw2urkjAnl8

[10] http://youtubetvnow.com/

have taken advantage of this by releasing their ads or marketing campaigns on YouTube first before they debut on TV. YouTube allows its users to upload a video, watch the videos freely, share videos on other social netwoking sites etc. Videos can be any type like songs, movies, recording clips, animation etc. YouTube have no restriction on the number of videos a user can watch, upload and share. Today YouTube become main e-learning (educational purpose) and entertainment source. According to the YouTube statistics[11]: Over 1 billion unique users visit YouTube each month; about 6 billion hours of video are watched each month on YouTube and 100 hours of video are uploaded to YouTube every minute. This statistics shows the enormous popularity of YouTube on web.



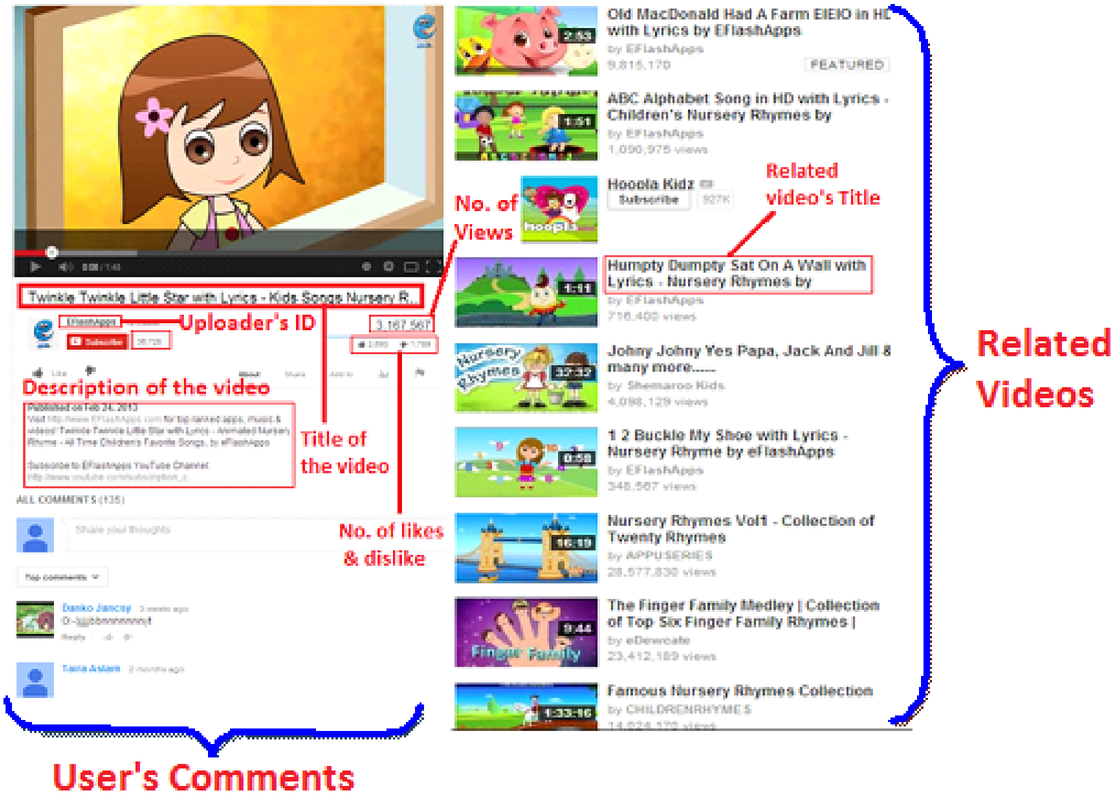
Flagged videos and users are reviewed by YouTube staff 24 hours a day, seven days a week to determine whether they violate Community Guideline. Accounts are penalised for Community Guidelines violations and serious or repeated violations can lead to account termination. If you would like to report a channel, please click here

Figure 1.1: The screenshot of the user's various activities on YouTube

In fact, YouTube not only allow users to upload and share videos, but also allows to post the comments in textual form, subscription for a particular channel, search any video using keywords & category and also provide features for users to interact with other users by comments and replying on the comments [8]. Users can also send the private messages to the other users in order to contact them. YouTube also allows its users to rate (like or dislike) the videos. Figure 1.1 shows the user's various activities on YouTube. A video contextual features includes title, a brief description about the video, textual comments, category of the video (entertainment, music, people & blogs etc). These user-generated content (UGC) may have some explicit information about the uploaded video content. Figure 1.2 shows the screenshot of the various contextual

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

features available for a video. Malicious content degrades the reputation of users who are involved in the videos and wastage the bandwidth for the user who are not willing to watch these videos and user-generated data help to identify such video on YouTube. There is no systematic framework for automatic identification to detect objectionable videos that a user is uploading.



Uploader's ID
Description of the video
Title of the video
No. of Views
No. of likes & dislike
Related video's Title
Related Videos
User's Comments
Twinkle Twinkle Little Star with Lyrics - Kids Songs Nursery R...
3,167,567
ALL COMMENTS (135)
Old MacDonald Had A Farm EIEIO in HD with Lyrics by EFlashApps
ABC Alphabet Song in HD with Lyrics - Children's Nursery Rhymes by
Hooola Kidz
Humpty Dumpty Sat On A Wall with Lyrics - Nursery Rhymes by
Johny Johny Yes Papa, Jack And Jill & many more......
1 2 Buckle My Shoe with Lyrics - Nursery Rhyme by eFlashApps
Nursery Rhymes Vol1 - Collection of Twenty Rhymes
The Finger Family Medley | Collection of Top Six Finger Family Rhymes |
Famous Nursery Rhymes Collection

Figure 1.2: The screenshot of the various contextual features available for a YouTube video

1.0.2 Privacy Invading and Harassment on YouTube

Privacy Invading is the wrongful or unauthorized taking and use of facts in order to disclosure of embarrassing others private information intentionally in public. YouTube allows its users to upload video without checking the content of the video. Therefore some malignant users take the advantage of this and make the small clips of their friends or other people in order to insult, make fun or with the intention of taking revenge. And those user post such clips and videos on YouTube which is publicly available to others. Therefore these kind of activities make the target person being harassed. Harassment can be as violence, in the form of fights, abuse etc. Generally prohibited material includes sexually explicit content, videos of animal abuse, shock videos, content uploaded without the copyright holder's consent, hate speech, spam, and predatory behaviour (refer to Figure 1.5). YouTube has a set of community guideline aimed to reduce abuse of the site's features. Despite the guidelines, YouTube has faced criticism from news sources for content in violation of these guidelines. According to an article in Chicago Tribune News[12], a California educator has resigned after a woman accused her in a YouTube video of abusing her when she was a 12-year-old student.

1.1 Research Motivation

YouTube is the largest video sharing website and contains a large amount videos being posted on it in every second. Low publication barriers (self-publishing model) and anonymity allows users to upload such content which is malicious and objectinable [18]. Figure 1.3 shows some offensive videos posted on YouTube. There is YouTube services of posting complaints by the users to remove the offensive content but that is not much good idea as YouTube administrators (refer to Figure 1.4 that shows screenshot of Safety and Abuse Center[13]) take time to review and remove them from YouTube. Other problem is that some users use YouTube as electronics aggression for the intention of harming others in terms of harassment, reputation. Cyber bullying and cyber-harassment which cause depression, low self-esteem and suicide [4]. Therefore a systematic framework for automatic identification of privacy invading harassment videos on YouTube needs to be developed.

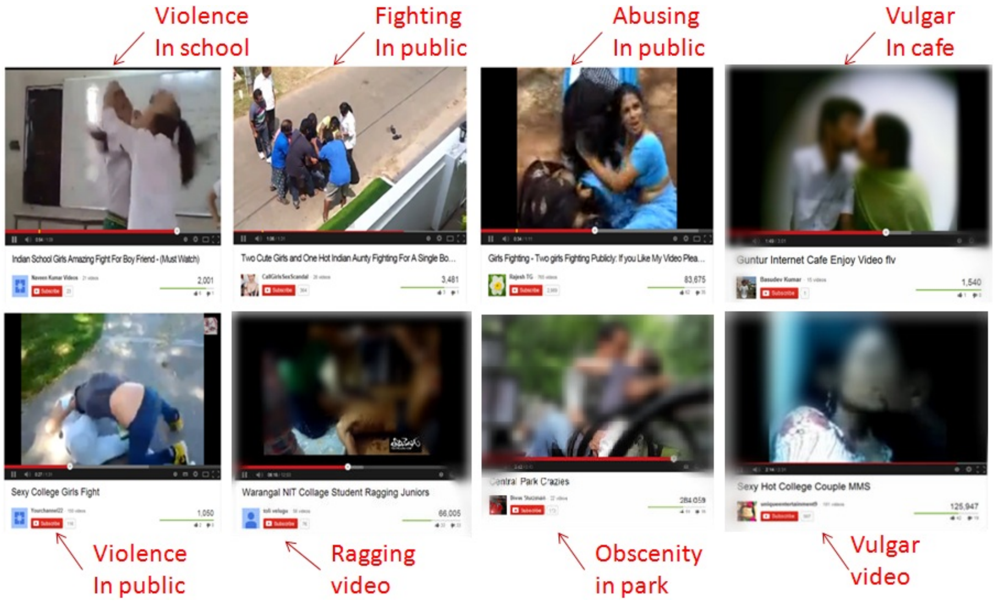

Violence
In school
Fighting
In public
Abusing
In public
Vulgar
In cafe
Indian School Girls Amazing Fight For Boy Friend - (Must Watch)
2,001
Two Cute Girls and One Hot Indian Aunty Fighting For A Single Bo...
3,481
Girls Fighting - Two girls Fighting Publicly: If you Like My Video Plea...
83,675
Guntur Internet Cafe Enjoy Video flv
1,540
Sexy College Girls Fight
1,050
Warangal NIT Collage Student Ragging Juniors
66,005
Central Park Crazies
284,059
Sexy Hot College Couple MMS
125,947
Violence
In public
Ragging
video
Obscenity
in park
Vulgar
video

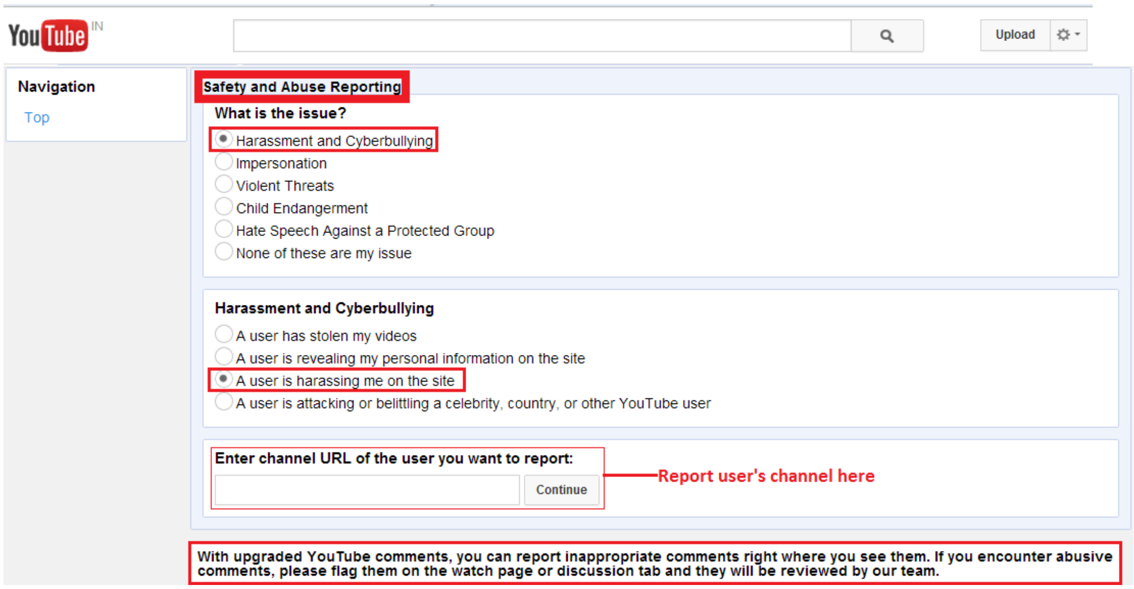
Figure 1.3: The screenshot of the harassment type videos posted on YouTube

The work presented in this paper is motivated by the fact that

1. A large amount of videos are posted on YouTube every minute and sometimes due to lack

[12] http://articles.chicagotribune.com/2014-01-21/news/sns-rt-usa-californiaabuse-20140121_1_youtube-video-abuse-accusations-educator

[13] https://support.google.com/youtube/answer/142443



Safety and Abuse Reporting

What is the issue?

● Harassment and Cyberbullying
○ Impersonation
○ Violent Threats
○ Child Endangerment
○ Hate Speech Against a Protected Group
○ None of these are my issue

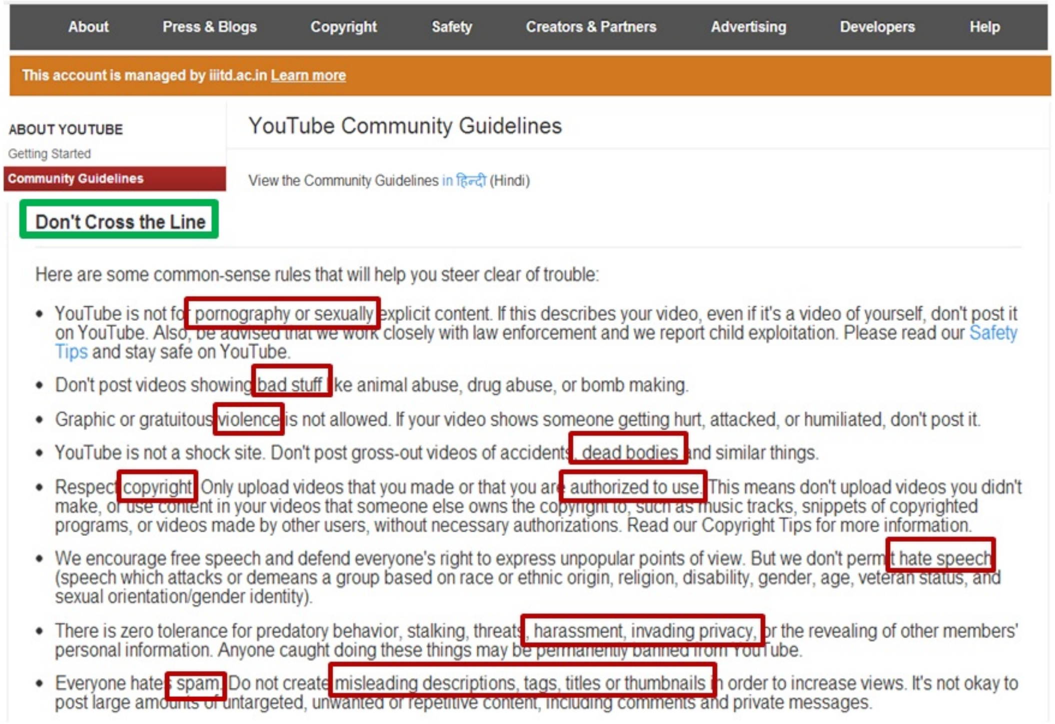
Harassment and Cyberbullying

○ A user has stolen my videos
○ A user is revealing my personal information on the site
● A user is harassing me on the site
○ A user is attacking or belittling a celebrity, country, or other YouTube user

Enter channel URL of the user you want to report:

With upgraded YouTube comments, you can report inappropriate comments right where you see them. If you encounter abusive comments, please flag them on the watch page or discussion tab and they will be reviewed by our team.

Figure 1.4: The screenshot of Safety and Abuse Reporting Center on YouTube



YouTube Community Guidelines

View the Community Guidelines in हिन्दी (Hindi)

Don't Cross the Line

Here are some common-sense rules that will help you steer clear of trouble:

- YouTube is not for pornography or sexually explicit content. If this describes your video, even if it's a video of yourself, don't post it on YouTube. Also, be advised that we work closely with law enforcement and we report child exploitation. Please read our Safety Tips and stay safe on YouTube.
- Don't post videos showing bad stuff like animal abuse, drug abuse, or bomb making.
- Graphic or gratuitous violence is not allowed. If your video shows someone getting hurt, attacked, or humiliated, don't post it.
- YouTube is not a shock site. Don't post gross-out videos of accidents, dead bodies and similar things.
- Respect copyright. Only upload videos that you made or that you are authorized to use. This means don't upload videos you didn't make, or use content in your videos that someone else owns the copyright to, such as music tracks, snippets of copyrighted programs, or videos made by other users, without necessary authorizations. Read our Copyright Tips for more information.
- We encourage free speech and defend everyone's right to express unpopular points of view. But we don't permit hate speech (speech which attacks or demeans a group based on race or ethnic origin, religion, disability, gender, age, veteran status, and sexual orientation/gender identity).
- There is zero tolerance for predatory behavior, stalking, threats, harassment, invading privacy, or the revealing of other members' personal information. Anyone caught doing these things may be permanently banned from YouTube.
- Everyone hates spam. Do not create misleading descriptions, tags, titles or thumbnails in order to increase views. It's not okay to post large amounts of untargeted, unwanted or repetitive content, including comments and private messages.

Figure 1.5: The screenshot of the YouTube Community Guidelines

of information people watch these offensive videos which cause bandwidth wastage for the users who are not willing to watch such videos and it violates the reputation of website.

2. Despite the several community guidelines[14] (refer to Figure 1.5) we found many offensive and malicious videos on YouTube. For example, violence in school, violence in public, vulgar in cafe, ragging in school etc (refer to Figure 1.3). Such videos can have negative

[14] https://support.google.com/youtube/answer/2802268?hl=en&ref_topic=2803240

impact on society and children.

3. We found that some previously watched malicious videos had been deleted from YouTube. The reasons behind for posting such kind of videos are: some users want to get more popularity and the channel subscription on YouTube, to make fun or to embarrass someone for the purpose of taking revenge etc.

1.2 Research Aim

As we have seen despite the YouTube community guidelines, some of the harassment videos are present on YouTube. Therefore the aim of the work presented in this research is to make the attention of the researcher's to solve the privacy invading harassment problem on YouTube.

The research objective of the work is presented in this paper is following:

1. **Broad Objective:** To increase our understanding of Harassment and Abused videos on YouTube. To investigate effective solution to combat the cyber harassment problem by mining the video metadata and identifying discriminatory features which can be used in harassment recognition framework.

2. **Specific Objective:** To examine one class classification approach for the task of recognizing Privacy Invading Harassment detection and misdemeanour videos on YouTube based upon contextual data such as title and description of the video, temporal data (duration of video) and demographic data (number of likes and dislikes, number of comments etc). To conduct a characterization study and empirical analysis on a real-world dataset to measure the effectiveness of the proposed hypothesis.

Chapter 2

Related Work and Research Contributions

2.1 Related Work

The work presented in this paper belongs to the area of privacy invading harassment detection on YouTube. We conduct a literature survey (refer to Table 2.1) in the area of cyber bullying and harassing content detection, personal insult detection on online social media. Table 2.1 reveals that most of the researches and techniques for violence and objectionable video detection are related to mining the comments and messages posted by the users and content (images and frames) analysis of the videos.

1. Theodoros et. al presented a multi-model approach for detecting violent content in video sharing sites. They proposed a 9-D feature vector based on audio, visual and textual features using binary classification to detect video is violent or not and more emphasis has been given to audio (like gunshots voice) feature. [9]

2. Deniz O. et. al proposed a method of extreme acceleration pattern (used acceleration method vectors) and action recognition techniques as a main discriminatory feature to detect fighting in videos. They discussed that efficiency of the extreme acceleration can be achieved efficiently by analyzing consecutive frames. [5]

3. Dadvar et. al investigate the cyber bullying detection in social network (e.g. MySpace) users' gender-specific (gender and age) information. They analysed some foul words (posted by each gender) present in the post and compared foul words that were used most frequently by each gender. They determined that male and female authors used significantly based on Wilcoxon signed rank test. [4]

4. Chaudhary et. al proposed a contextual based one class classifier approach which detect video response spam (botnet, promotional and offensive videos) on YouTube. The

Table 2.1: Summary of literature survey of 12 papers, arranged in reverse chronological order, identifying harassment and offensive content on social media.

Research Study	**Objective & Analysis**
Singhal et. al; 2013 [16]	proposed an approach to detect Cyber harassment and bully users based on the text written by the user and their real identity.
Chaudhary et. al; 2013 [1]	A one class classifier approach for detecting video response spam (Promotional and offensive videos) on YouTube.
Deniz O. et. al; 2012 [5]	Proposed a method to detect fighting in videos using by analysing acceleration patterns.
Dadvar et. al; 2012 [4]	Investigate the user information (gender and age) to detect cyber-bullying in MySpace.
Chen et. al; 2012 [2]	Distinguishing abusive and obscene words from text in order to identify a harassing and cyber bullying content.
Sood et. al; 2012 [17]	Identifying users' contribution in off topic, negative and personal insulting comments on social news sites.
Dinakar et. al; 2011 [6]	An approach to build a topic sensitive classifier by analysing identifying cyberbullying in textual comments posted on a topic.
Theodoros et. al; 2010 [9]	Describes a multi-model approach to identify violence in videos by using audio, visual and textual features.
Lee et. al; 2009 [13]	Describes a multimedia and contextual mining based approach to detect offensive content in videos.
Yin et. al; 2009 [19]	Detecting harassing posts in chat rooms & discussion forums by using local, sentiment, & contextual features.
Kim et. al; 2008 [11]	Proposed a multi-media approach to detect obscene videos by analysing shape, size and colour of video frames.
Mahmud et. al; 2008 [14]	A automated system to analyze contextual information to identify insulting or abusive text.

proposed method retrieved the metadata of the response videos and based on the discriminatory features (linguistic, temporal and popularity based features) detect spam videos. [1]

5. Lee et. al proposed a multilevel hierarchical system having 3 phases in different temporal domains based on multimedia (frame, colour) and contextual mining approach to detect offensive content in videos. Phases includes early detection based on textual features (file size, frame rate) of the videos header encrypted by hash signatures (and compared with predefined signatures), real time detection based on shape features that uses SVM to detect harmfulness of each frame and the last phase posterior detection based on GoF(group of frame) that uses skin color feature. [13]

6. Kim et. al proposed a multimedia approach based on segmentation (splitting the frames into the region based on colors) for detecting obscene videos. They analysed each frame of the input videos based on some features such as shape, size and colour of video frames to detect objectionable or benign frame. [11]

2.2 Research Contributions

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

1. In comparison to previous work, the work presented in this paper is the first step in the direction of applying a one-class classifier based approach for detecting privacy invading harassment and misdemeanour content on YouTube using video metadata. The work is presented into four category i.e. vulgar video detection, abuse and violence detection in the educational area as well in public places and abuse detection in terms of ragging in school and colleges based on video's 13 contextual features which is novel task in this area.

2. We conduct a set of experiments and perform an empirical analysis on real world dataset to evaluate the effectiveness of the proposed system based upon discriminatory features.

Chapter 3

Proposed Solution Approach

Figure 3.1 presents a general research framework for the proposed solution approach. We divide the privacy invading harassment detection problem into four sub-problems: vulgar video detection (VVD), violence and abuse video detection in school and colleges (VAVDS), violence and abuse video detection in public places (VAVDP), ragging video detection in school and colleges (RVDC). VVD, VAVDS, VAVDP and RVDC are employed as one class classification approach that performed recognition task for detection of harassment. As shown in Figure 3.1, the proposed solution approach consists of three phase: Dataset Extraction, Features Identification & Features Selection and Classification.

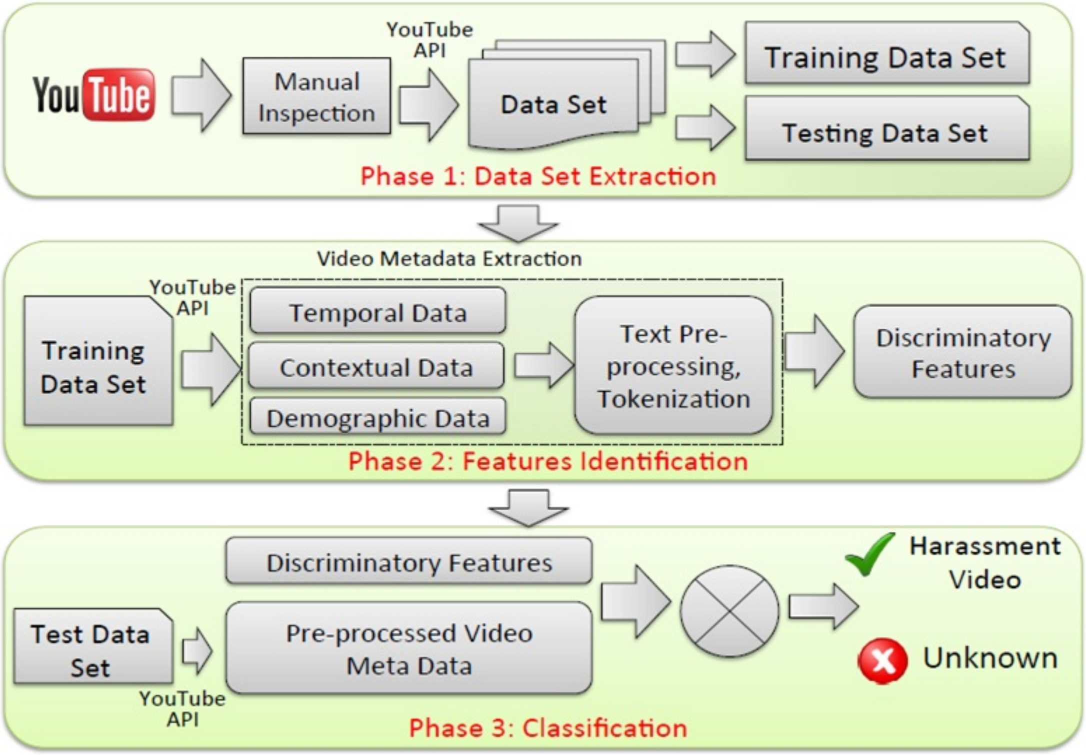


Figure 3.1: Research Framework

3.1 Dataset Extraction

The first step is used to acquire the training dataset using manual inspection of the videos and further to divide the training dataset into the four given sub-problems. We also acquire the experimental testing dataset (mostly viewed YouTube videos using YouTube API) of the given problem for the purpose of testing the proposed solution approach. The process of extracting dataset is a time consuming and monotonous.

3.2 Feature Identification and Feature Selection

Feature identification is process of identifying all the available features of YouTube video. During manual analysis we visually inspect the features of the video.

3.2.1 Video's Metadata Extraction

In this step, we fetch all the metadata of available features of a video on YouTube using YouTube API and we have divided the feature set into four categories: linguistic based feature, popularity based feature, YouTube category feature, temporal based feature. Linguistic based features include percentage of terms (dirty, violence, ragging related terms and also people type related terms) present in title and description of the video. Popularity based features include ratio of likes by view count, ratio of comments by view count etc. YouTube category feature tell us about the category of the video such as entertainment, music, sports etc and temporal based feature include duration of the video. Linguistic based feature analysis reveals that higher number of terms present, higher the chances that video is harassment type. Therefore percentage of terms present in the title and description is computed using text preprocessing and textual similarity.

3.2.1.1 Preprocessing

In this subphase we pre-process the fetched (linguistic based) training dataset. Text preprocessing involves tokenization and stop-words removal. Tokenization is a process (word level) of breaking up the complete sentence or line into smallest meaningful elements called as tokens. Stop words are very common words used in a sentence which are filtered out prior to, or after processing of text. Since the presence of stop words affect the performance of the algorithm. So, after tokenization we use the standard english stop-words list[1] given on the web to remove such words (like unigram, bigram, trigram) from the token list. Figure 3.2 shows the cloud of few stop words lexicons used in the algorithm.

[1] http://norm.al/2009/04/14/list-of-english-stop-words/

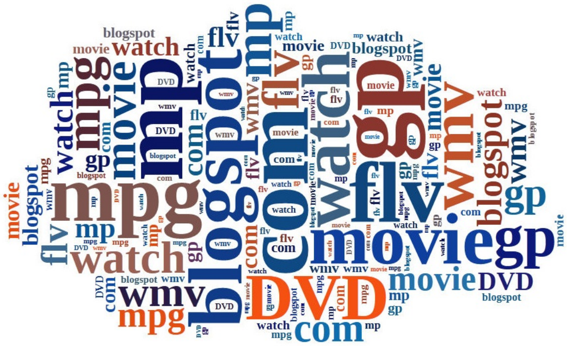


Figure 3.2: A Cloud of a few Stop Words Lexicon

3.2.1.2 Textual Similarity

In this step we find the similarity between the token of the title and description of sub-problems with their respective lexicon list. To compute the similarity of dirty terms of VVD we used a Standard bad-words lexicon given over the web[2] and for the remaining categories (VAVDS, VAVDP, RVDC) we made our own lexicon list (violence terms in educational area and in public places, ragging related terms, people type) based on manual analysis and visual inspection. Figure 3.3 shows the cloud of few dirty and violence lexicons that we used for computation. We also use lexicon of people type (refer to Figure 3.4) to compute the percentage of people involved in the action of video.

Figure 3.3: A Cloud of a few Dirty and Violence Lexicon

3.2.2 Feature Selection

Feature selection is the process of selecting a subset of relevant features use in proposed model and using only that subset for classification task. We conduct an in-depth manual analysis and visual inspection of the extracted metadata of the videos to identify the relevant features (refer Table 3.1) for the purpose of classification.

[2] http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/

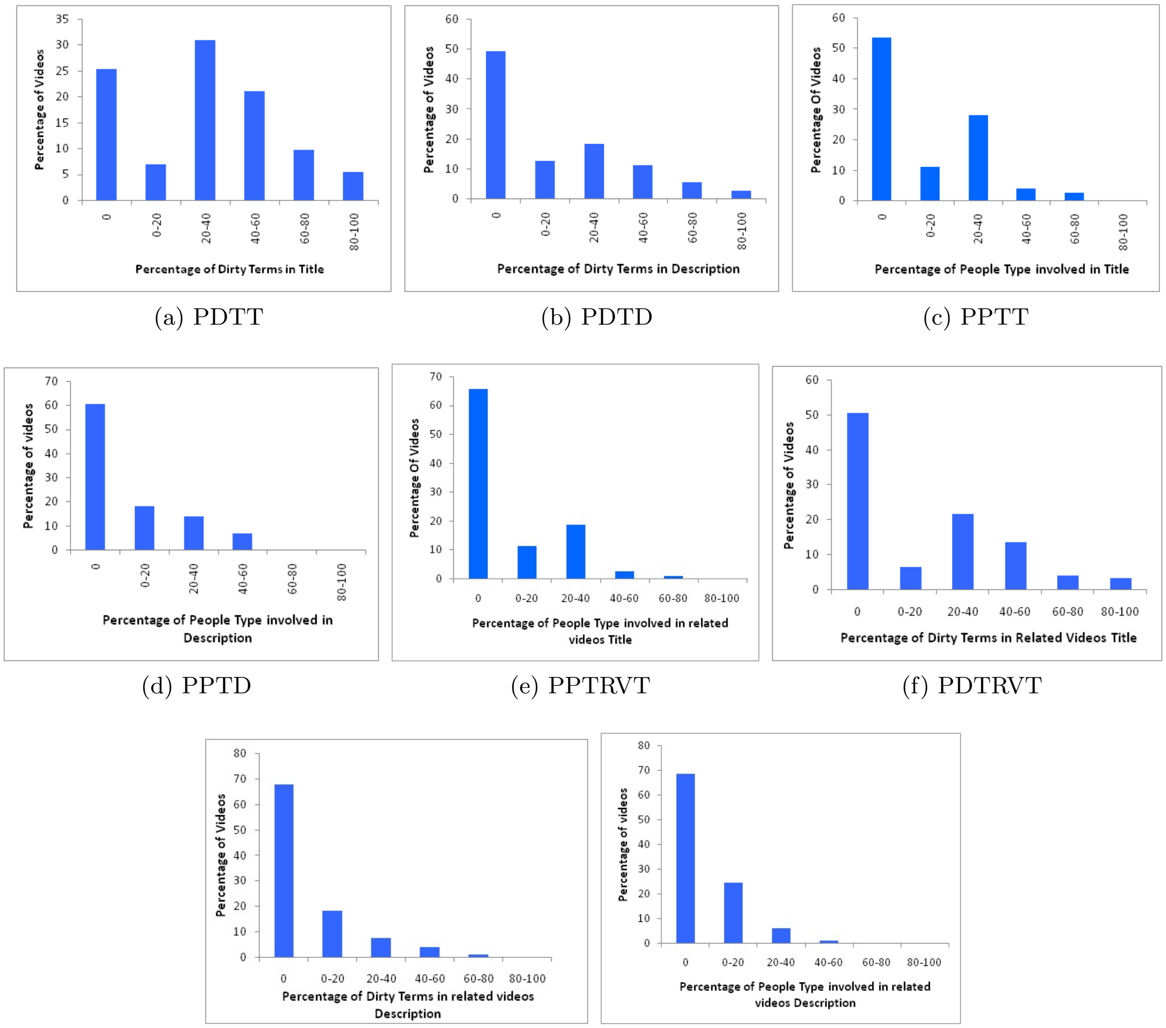
Figure 3.4: A Cloud of few People Type Lexicon

Table 3.1: List of 13 Contextual Features in Priority Based Order for all *Four* Category of Video's Harassment i.e. VVD, VAVDS, VAVDP and RVDC

Abbr.	**Feature Title**	**Type**	**Remarks**
DYTV	Duration of the YouTube Video	Temporal	Duration shows the length of the Content of the Video. By manual analysis duration of such videos is between 30 sec to 250 sec.
PXTT	% of X-Terms present in the title	Linguistic	Terms like MMS, kiss, sex, hot, violence, fight, fighting, ragging present in title and description shows the video category corresponding to term present in the video.
PXTD	% of X-Terms present in the Description	Linguistic	
PXTRVT	% of X-Terms in Related-Videos' title	Linguistic	
PXTRVD	% of X-Terms in Related-Videos' Description	Linguistic	
CatV	YouTube Category of the video	YT Category	YouTube category of such kind of videos usually belongs to entertainment, people & blogs, comedy.
PPTT	% of People Type present in the title	Linguistic	People Type like couple, lover, girls, boys, lady, students, uncle, aunty, junior, senior, freshers, kids present in title shows that people are doing some kind of activity in the video.
PPTD	% of People Type present in the Description	Linguistic	
PPTRVT	% of People type in Related-Videos' title	Linguistic	
PPTRVD	% of People Type in Related-Videos' Description	Linguistic	
RLBV	Ratio of #likes by #views	Popularity	Like and View feature shows the popularity of the video. As #views of such kind of videos are more as compare to like feature. So ratio of like by view count is usually less.
RCBV	Ratio of #Comments by #views	Popularity	Comments can be as an indication about the content of video and also is a trace about the user's mentality.So #comments on such videos is considered as less. Therefore ratio of comments by view count is usually less.
RRBV	Ratio of #Raters by #views	Popularity	#raters (Like & dislike) and View feature shows the popularity of the video. So ratio of raters by view count is usually less than the ratio of comment by view count.

3.2.2.1 Linguistic Based Features

1. **Percentage of X-Terms present in Title and Description:** Based on manual analysis we hypothesize that presence of dirty, violence, abuse, ragging related terms in title and description is an indicator to recognize harassment videos. Our observation shows that in vulgar videos 75% (refer Figure 3.5a) and 50% of videos (refer Figure 3.5b) having some dirty terms in their title and description respectively. Similarly 90% and more than 45% of VAVDS videos contain some violence terms (refer Figure 3.6a, 3.6b) in their title and description respectively. Figure 3.7a and Figure 3.7b shows that more than 90% of VAVDP videos having violence terms in their title and 60% of the VAVDP videos contain some violence terms in their description. In case of RVDC videos more than 90% and 35% of videos have some ragging terms (refer Figure 3.8a and Figure 3.8b) in their title and description respectively. This analysis shows that linguistic features are very important and discriminatory features for classification task.



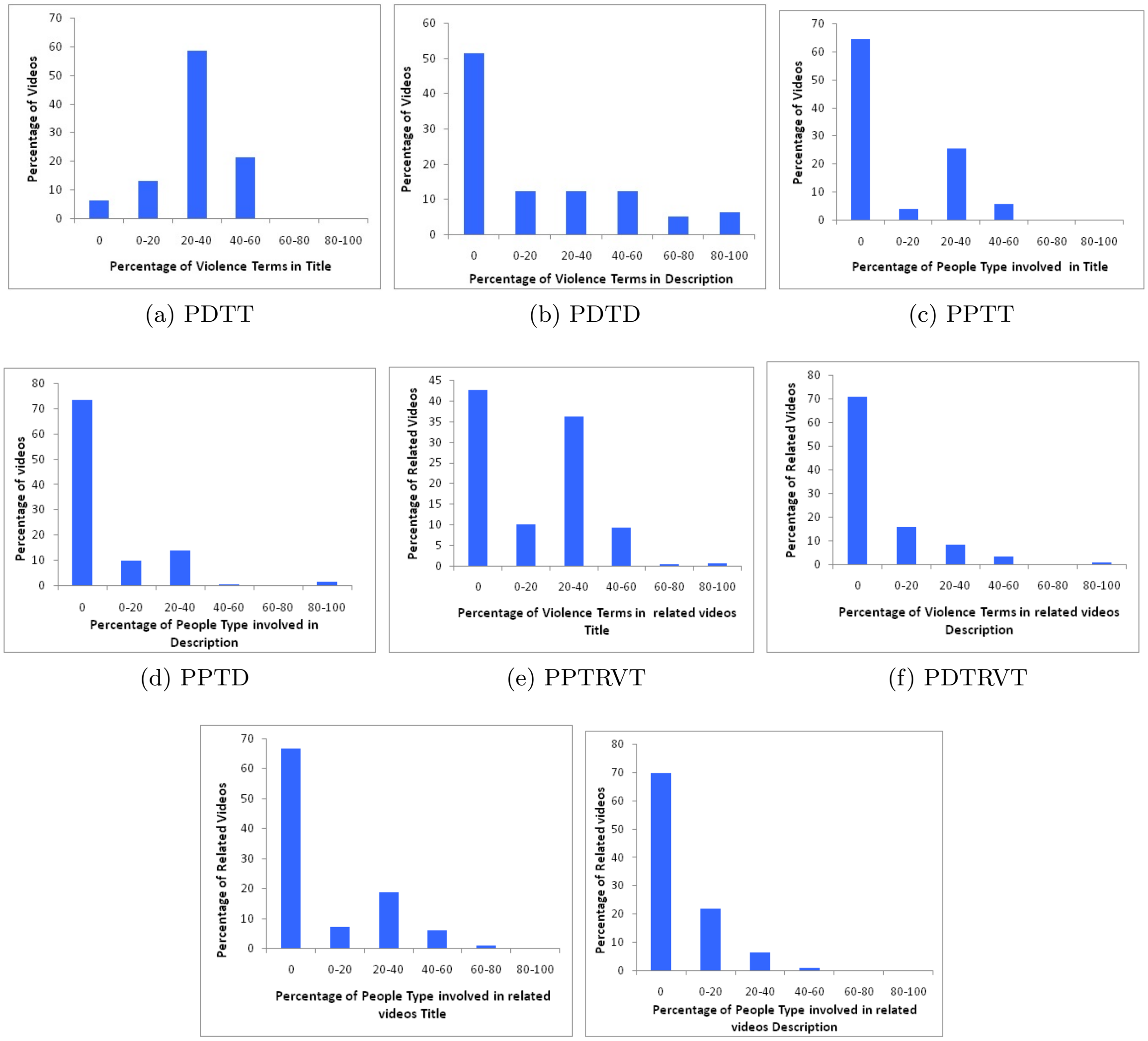
(a) PDTT (b) PDTD (c) PPTT

(d) PPTD (e) PPTRVT (f) PDTRVT

(g) PDTRVD (h) PPTRVD

Figure 3.5: Linguistic Based Features for VVD

2. **Percentage of People Type present in Title and Description:** We have analysed that presence of people type like girls, boys, lovers, students, junior, senior, lady, people etc. in title and description shows that some kind of actions are going on in the video content. We hypothesized that if some kind of dirty and violence terms present in the videos title and description as well as people type terms then the video is misdemeanour video. We observe that 45% and 40% of the vulgar videos (VVD) contain some people type related terms (refer Figure 3.5c & 3.5d) in their title and description. Similarly 35% and more than 25% of VAVDS (refer Figure 3.6c, 3.6d), more than 40% and 35% of the VAVDP (refer Figure 3.7c, 3.7d), more than 15% and 20% of RVDC have some people type related terms (refer Figure 3.8c, 3.8d) in their title and description respectively. Our analysis shows that presence of people type related terms in title and description is also an important and discriminatory feature.



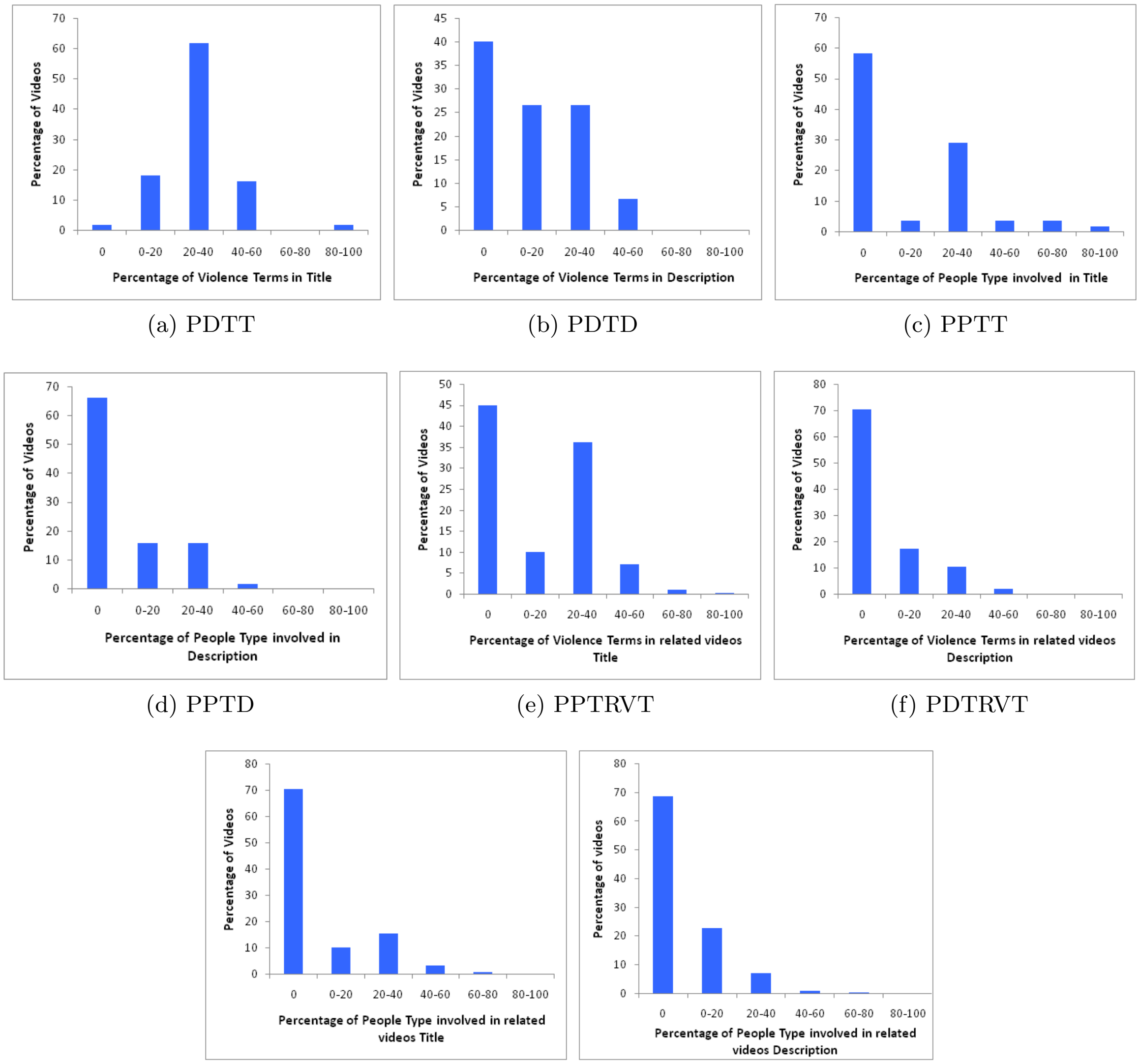
(a) PDTT (b) PDTD (c) PPTT

(d) PPTD (e) PPTRVT (f) PDTRVT

(g) PDTRVD (h) PPTRVD

Figure 3.6: Linguistic Based Features for VAVDS

3. **Percentage of X-Terms present in Related Video's Title and Description:** Here assumption is made based on the manual analysis of the related videos of the given video i.e. if related videos have some X-terms in their title and description then given video should also belong to same category as related videos. In related videos, priority is given to relevance and matching words present in the given videos title, description and tag. Sometimes usernames when relevance arent great that much. We observe that 50% and 30% of VVDs related videos contain some dirty terms (refer Figure 3.5f, 3.5g) in their title and description respectively. Similarly 55% and more than 30% of VAVDSs related videos contain some violence terms (refer Figure 3.6e, 3.6f), more than 55% and 30% of the VAVDP's related videos contain some violence terms (refer Figure 3.7e, 3.7f) in their title and description respectively. In case of ragging's related videos, more than 50% and 25% of videos have some ragging terms (refer Figure 3.8e, 3.8f) in their title and description respectively.



(a) PDTT

(b) PDTD

(c) PPTT

(d) PPTD

(e) PPTRVT

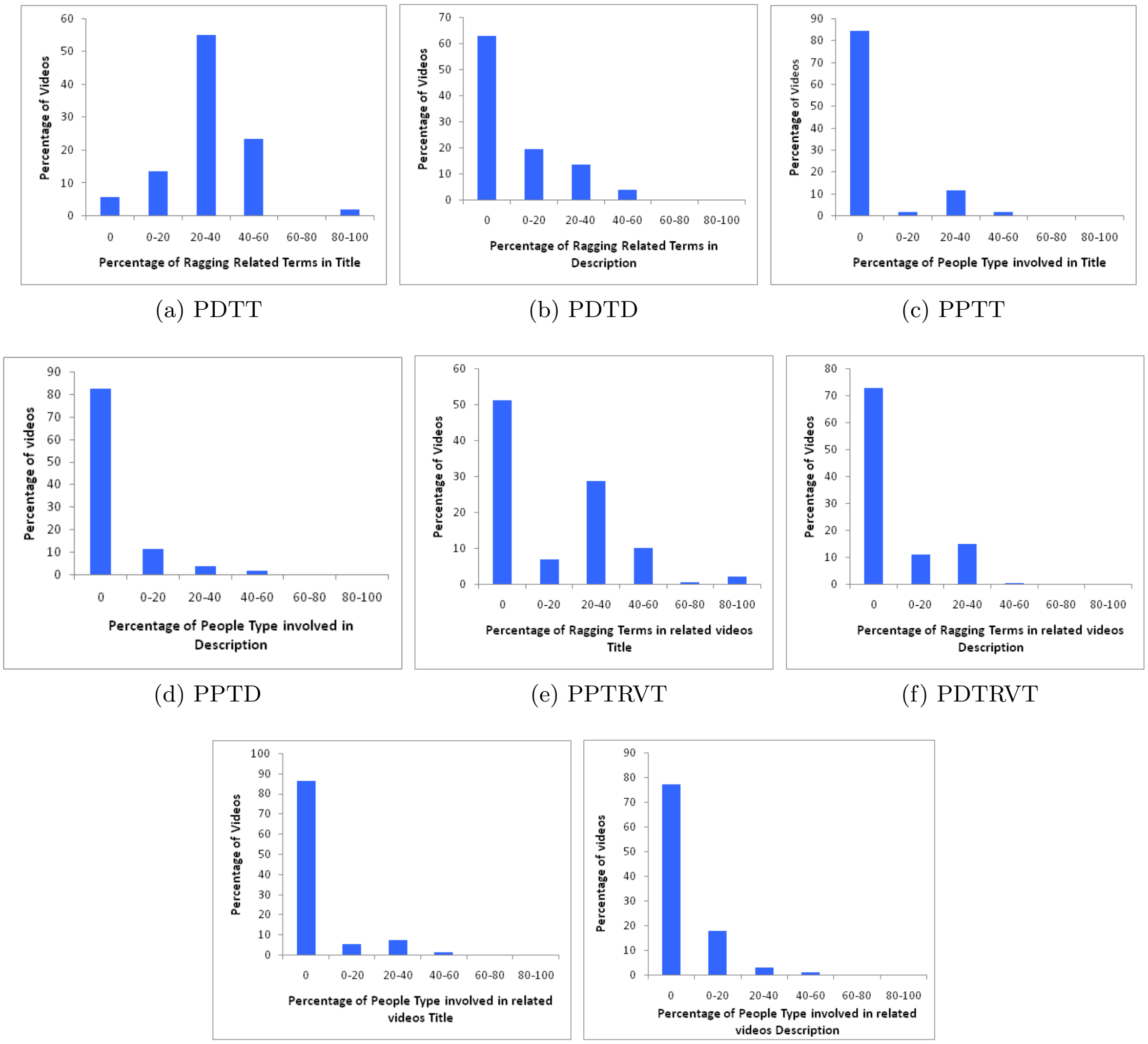
(f) PDTRVT

(g) PDTRVD

(h) PPTRVD

Figure 3.7: Linguistic Based Features for VAVDP

4. **Percentage of People Type present in Related Video's Title and Description:** We observe presence of people type like girls, boys, lovers, students, junior, senior, lady, people etc. in title and description of the related videos shows that some kind of actions would be there in the video content. Our assumption is based on the fact that if related videos of the given video have some X-terms in their title and description as well as people type related terms then given video should also belongs to same category as related videos. Our analyze that 35% and 30% of the vulgar videos (VVD) contain some people type related terms (refer Figure 3.5e & Figure 3.5h). Similarly 32% and more than 30% of VAVDS (refer Figure 3.6g and Figure 3.6h), more than 30% and 31% of the VAVDP (refer Figure 3.7g and Figure 3.7h), more than 12% and 22% of RVDC have some people type related terms (refer Figure 3.8g and Figure 3.8h) in their title and description respectively.



(a) PDTT

(b) PDTD

(c) PPTT

(d) PPTD

(e) PPTRVT

(f) PDTRVT

(g) PDTRVD

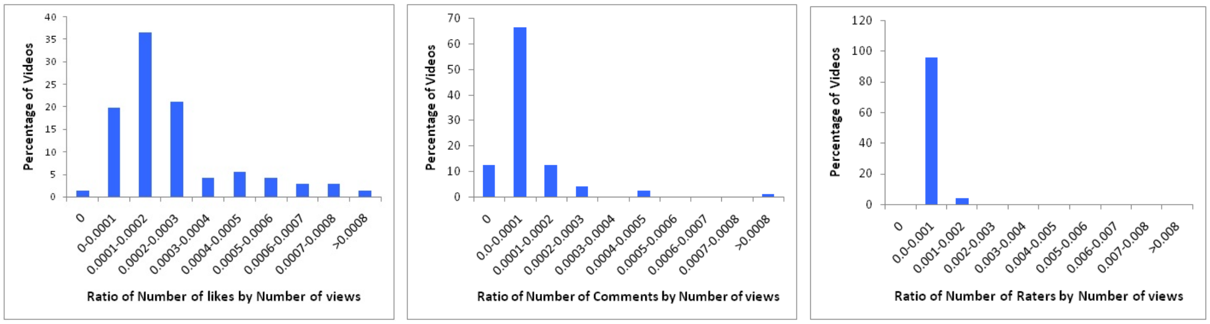
(h) PPTRVD

Figure 3.8: Linguistic Based Features for RVDC

3.2.2.2 Popularity Based Features

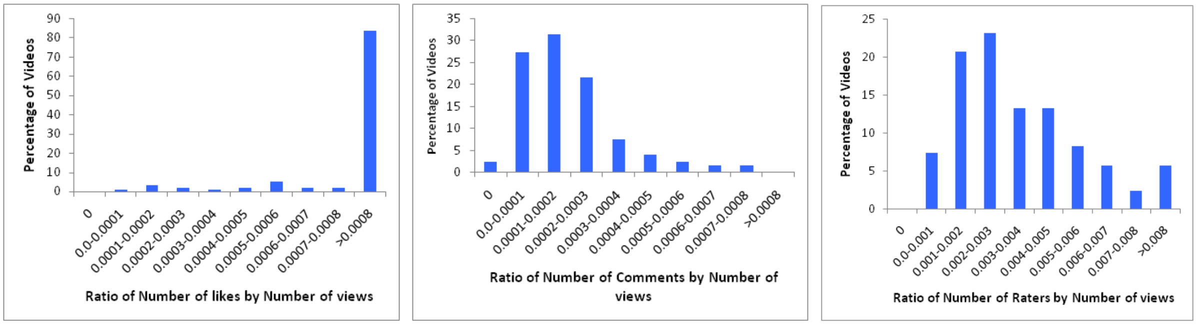
1. **Ratio of number of likes by numbers of views:** We first fetch the number of likes and number of views for our training dataset and then we compute the RLBV value. Based on manual analysis of the training dataset we found that 77% of the vulgar videos have RLBV (refer Figure 3.9a) value between $0.00001 - 0.0003$ which is very less because people don't take interest to like such kind of video. Violence in school and college i.e. fighting in such places, generally people likes such kind of video. Therefore 83% of RLBV (refer Figure 3.10a) value is greater than 0.0008. For VAVDP videos, Figure 3.11a shows that 44% and 26% of the video have RLBV value between $0.0001 - 0.0004$ and more than 0.0008 respectively. In case of RVDC, 47% and 45% (refer Figure 3.12a) of the video have RLBV

value between 0.0001 − 0.0004 and more than 0.0004 respectively.



(a) RLBV (b) RCBV (c) RRBV

Figure 3.9: Populrity Based Features for VVD

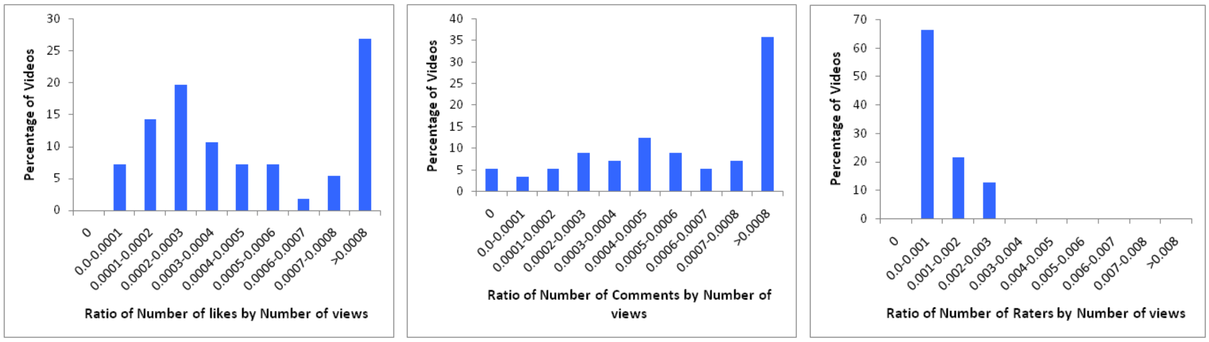


(a) RLBV (b) RCBV (c) RRBV

Figure 3.10: Populrity Based Features for VAVDS

2. **Ratio of number of comments by numbers of views:** We observed that number of comments are less on such kind of videos as compared to non-harassment videos. We first fetch the number of comments for our training dataset and then we compute the RCBV value. We found that 78% of the vulgar videos have RCBV (refer Figure 3.9b) value between 0.00001 − 0.0002 which is very less.The reason behind this is people dont want to comments on such kind of video (as it shows the bad mentality of the user). Violence in school and college i.e. fighting in such places generally people comment on such kind of video for fun. Therefore 82% of RCBV (refer Figure 3.10b) value is greater than 0.0003 in VAVDS videos. Figure 3.11b shows that 37% and 35% of the violence video in public place have RCBV value between 0.0002−0.0006 and more than 0.0008 respectively. Figure 3.12b shows that 50% and 45% of the ragging video have RCBV value between 0.0001 − 0.0005 and more than 0.0008 respectively.

3. **Ratio of number of raters by numbers of views:** Number of raters (#likes + #dislikes) is also a good indication of for finding the popularity of the videos. We first fetch



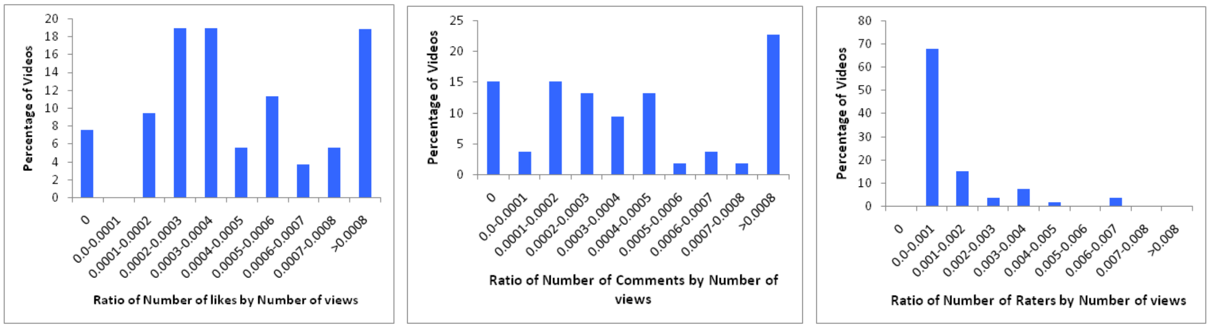
(a) RLBV

(b) RCBV

(c) RRBV

Figure 3.11: Populrity Based Features for VAVDP

the number of raters for our training dataset and then we compute the RRBV value. We observe that RRBV value is more as compared to RLBV and RCBV. Figure 3.9c shows that 95% of the vulgar videos have RRBV value between $0.0001 - 0.001$ which is more as compared to RVBV because people dislike such kind of videos. in case of violence in school and college i.e. fighting in educational places, generally people either like or dislike such kind of videos. Therefore 78% of RRBV (refer Figure 3.10c) value between $0.001 - 0.005$. Figure 3.11c shows that 87% of the VAVDP videos have RRBVP value between $0.0001 - 0.002$. In RVDC videos (refer to Figure 3.12c) 82% of the video have RRBV value between $0.0001 - 0.002$.



(a) RLBV

(b) RCBV

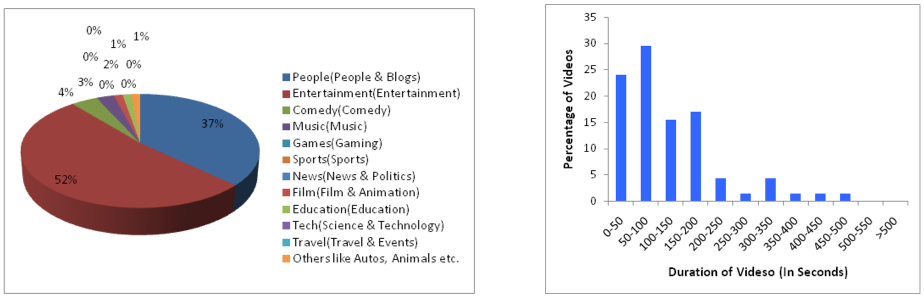
(c) RRBV

Figure 3.12: Populrity Based Features for RVDC

3.2.2.3 Temporal Based Feature

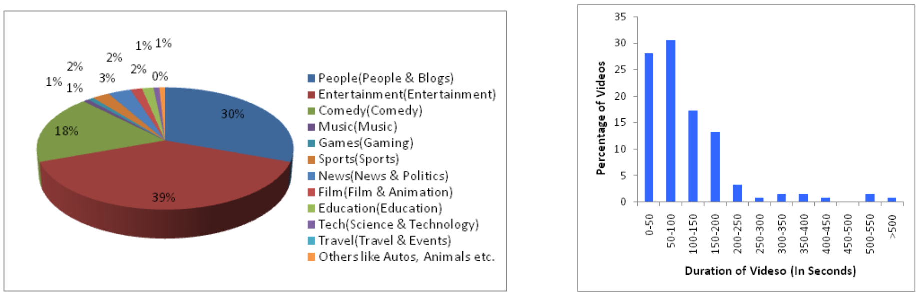
Temporal based feature is related to time i.e. *duration of the video* that shows the content length of the video. By visual inspection we found that most the videos fall under the category of harassment having duration between 30 second to 250 second. Because such kind of the videos are capture as a small clips and then uploaded on YouTube. We observed that 85% and 62% (refer Figure 3.13b), 89% and 61% (refer Figure 3.14b), 80%

and 60% (refer Figure 3.15b) of the vulgar videos, VAVDS and VAVDP have duration less than 200 second and between 50 − 200 second respectively. Similarly we observed that 54% and 74% (refer Figure 3.16b) of RDVC have duration less than 200 seconds and less than 350 seconds respectively. Therefore duration of the video is a very good indicator for detecting harassment videos.



(a) CatV (b) DYTV

Figure 3.13: YouTube Category Feature and Duration of Video for VVD

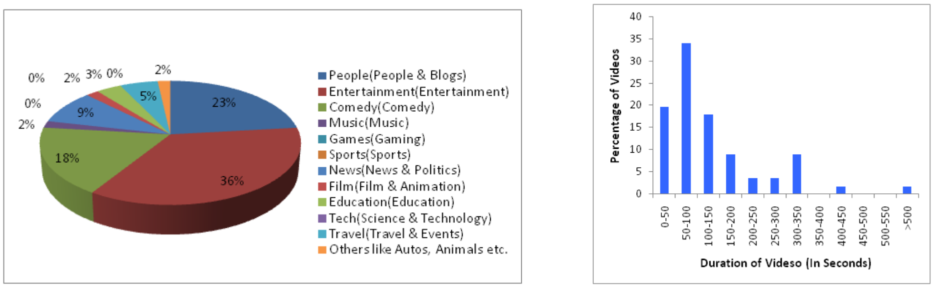


(a) CatV (b) DYTV

Figure 3.14: YouTube Category Feature and Duration of Video for VAVDS

3.2.2.4 YouTube Category Feature

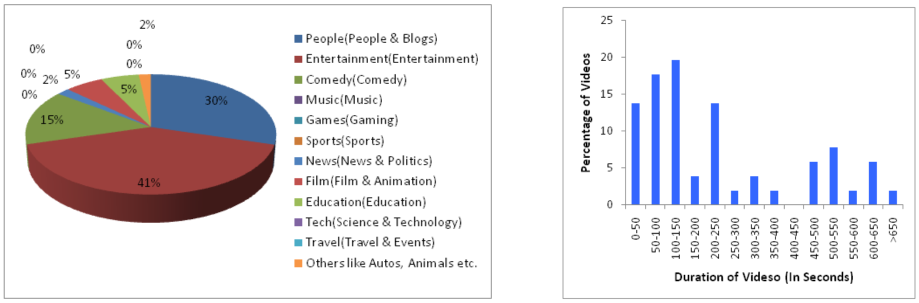
YouTube Category is the feature which shows the category of the video i.e. entertainment, sports, music, news, education etc. Based on manual analysis of the videos shows that out of total 16 YouTube categories, 52%, 37% and 4% of vulgar videos (refer Figure 3.13a) fall under the category of entertainment, people & blogs and comedy respectively. Similarly 39%, 30% and 18% of VAVDS videos (refer Figure 3.14a), 36%, 23% and 18% of VAVDP videos (refer Figure 3.15a) belong to the category of entertainment, people & blogs and comedy respectively. Figure 3.16a shows that 41%, 30% and 15% of the ragging videos are from the category of entertainment, people & blogs and comedy respectively. Therefore, the videos fall under the category of News (News & Politics), music, film are not consider as harassment videos.



(a) CatV

(b) DYTV

Figure 3.15: YouTube Category Feature and Duration of Video for VAVDP



(a) CatV

(b) DYTV

Figure 3.16: YouTube Category Feature and Duration of Video for VAVDR

3.3 Classification

In this subphase we use one-class classification approach to classify the video as harassment or unknown for each sub-problem. This classifier classifies the video as harassment class if it satisfy all the discriminatory features' value with their respective threshold values otherwise video belongs to unknown class. We perform a manual search on YouTube and created a few lexicons (refer to Figure 3.17) of news-channel due to fact that videos from the news channel cannot be consider as privacy invading harassment videos. If video uploader's channel name is listed in the news channel lexicons, we skip that video otherwise using one classification approach we classify that video either belongs to harassment class or unknown class [10].

3.3.1 Solution Implementation to Detect Privacy Invading Harassment Videos

In one class classification problem, either the video belongs to harassment or unknown class. The algorithm does the similarity computation of video's 13 discriminatory features with the threshold value obtained from training dataset to perform the recognition task of videos as harassment or unknown. In this section, we describe the classifier that we have developed to detect the videos class.

Algorithm 1: Identification of Harassment Videos

```
Data: News Channels $NC \in LexiconL1$, X terms $Key_{xxx} \in L2$, People type
      $Key_{people} \in L3$, English Stop words Dictionary $D_{st}$, Video IDs $V_{id} \in V$, Threshold
      $th \in \{t_1, t_2 ... t_{13}\}$
Result: Class of Video
for all $V_{id} \in V$ do
    $U \leftarrow V_{id}.getUploader$;
    if $(U \in NC)$ then
        $Display$ News Channel Video;
    else
        $V_t \leftarrow V_{id}.getTitle$; $V_d \leftarrow V_{id}.getDescription$; $V_l \leftarrow V_{id}.getDuration$;
        $V_c \leftarrow V_{id}.getCategory$; $V_{nl} \leftarrow V_{id}.getLikes$; $V_{nv} \leftarrow V_{id}.getNumViews$;
        $V_{nr} \leftarrow V_{id}.getNumRaters$; $RelV \leftarrow V_{id}.getTopKRelatedVideo$;
        $V_{nc} \leftarrow V_{id}.getNumComments$;
        for all $R_{id} \in RelV$ do
            $R_t \leftarrow R_{id}.getTitle$; $Rd \leftarrow R_{id}.getDescription$;
        $f1 \leftarrow V_l$; $f6 \leftarrow V_c$; $f11 \leftarrow V_{nl} : V_{nv}$; $f12 \leftarrow V_{nc} : V_{nv}$; $f13 \leftarrow V_{nr} : V_{nv}$;
        for all $V_t, V_d, R_t, Rd$ do
            $V_t \leftarrow V_t.remove(D_{st})$; $V_d \leftarrow V_d.remove(D_{st})$;
            $R_t \leftarrow R_t.remove(D_{st})$; $R_d \leftarrow R_d.remove(D_{st})$;
            $f2 \leftarrow$%Terms in $V_t \in L2$; $f3 \leftarrow$%Terms in $V_d \in L2$;
            $R_{TT} \leftarrow$%Terms in $R_t \in L2$; $R_{TD} \leftarrow$%Terms in $R_d \in L2$;
            $f7 \leftarrow$%People in $V_t \in L3$; $f8 \leftarrow$%People in $V_d \in L3$;
            $R_{PT} \leftarrow$%People in $R_t \in L3$; $R_{PD} \leftarrow$%People in $R_d \in L3$;
        for all $R_{TT}, R_{TD}, R_{PT}, R_{PD}$ do
            $f4 \leftarrow$ Harmonic Mean$[R_{TT}]$; $f5 \leftarrow$ Harmonic Mean$[R_{TD}]$;
            $f9 \leftarrow$ Harmonic Mean$[R_{PT}]$; $f10 \leftarrow$ Harmonic Mean$[R_{PD}]$;
    if
    $(\{f1 \wedge (f2 \vee f3)\} \vee [\{(f4 \vee f5) \wedge (f6)\} \wedge \{(f7 \vee f8) \wedge (f9 \vee f10)\} \wedge \{f11 \vee f12 \vee f13\}]) ::$
    $\{t_1, t_2 ... t_{13}\}$ then
        return $Class \leftarrow$Harassment;
    else
        return $Class \leftarrow$Unknown;
```

Figure 3.17: A Cloud of a Few Official News-Channels on YouTube (News & Politics Category)

Algorithm 1 shows our proposed solution implementation of classifier for classifying the harassment videos and unknown class videos. The result of the algorithm shows the class of video i.e. either harassment or unknown class.

In Step 2 we retrieve the UploaderID U corresponding to each VideoID V_{id}.

In Step 3 we look for whether the retrieved UploaderID is from news channel or not. If it belongs to news channel then return "video is from news channel".

Steps 4 to 6 extract all required metadata for all V_{id} of the video i.e. title and description of the video, duration of the video, video category, number of likes, number of views and number of raters. We also retrieve Top K related videoIDs corresponding to each V_{id} and store them in $RelV$.

Steps $7-8$ retrieves the title R_t and description R_d of each related videoID R_{id}.

Step 9 performs the likes by views ratio, comments by views ratio and raters by views ratio.

Steps $10-16$ represent the preprocessing and textual similarity of linguistic features. In Steps $11-12$ we perform the title and description pre-processing i.e. removing the stop words from the title and description of the video and their related videos. In Steps $13-16$ we perform the textual similarity of X-Terms present in title and description with the lexicon $L2$ and also for people type present in title and description with the lexicon $L3$ and we calculate the percentage of X-terms and people type present in the title and description respectively.

In case of related videos, we calculate the percentage of the X-terms and people type for each Top K videos corresponding to V_{id}.

In the Steps $17-19$ we find the harmonic mean of the calculated percentage of R_{TT}, R_{PT}, R_{TD} and R_{PD} of all related video's using formula as given below.

$$\boxed{f_4 = \frac{K}{\sum_{i=1}^{K} \frac{1}{R_{TT_i}}}} \tag{3.1}$$

$$\boxed{f_5 = \frac{K}{\sum_{i=1}^{K} \frac{1}{R_{PT_i}}}} \tag{3.2}$$

$$\boxed{f_9 = \frac{K}{\sum_{i=1}^{K} \frac{1}{R_{TD_i}}}} \tag{3.3}$$

$$\boxed{f_{10} = \frac{K}{\sum_{i=1}^{K} \frac{1}{R_{PD_i}}}} \tag{3.4}$$

Steps $19 - 22$ perform the classification procedure. In step 19, we compare the all the discriminatory features' value $f1, f2, f3, ..., f13$ with their respective threshold values (obtained from training dataset) i.e. $t_1, t_2, t_3, ..., t_{13}$. If all discriminatory features' value satisfy corresponding threshold values then it classifies the video as harassment class otherwise unknown class.

Chapter 4

Empirical Analysis and Performance Evaluation

4.1 Experimental Dataset

We collect experimental dataset from the YouTube using YouTube API. First we collect the videoIDs (unique value for each video) of videos manually that are related to our problem statement i.e. vulgar videos, violence and abuse in educational area as well in public places and the last category is related to ragging in school and colleges. Then we download all the metadata of the collected videoIDs and top 10 related videos of their corresponding videoIDs. We are able to fetch a total of 318 videos and a total of 6401 related videos for training dataset. Among 6401 videos, we have 1752 related videos of vulgar category. Similarly $2025, 1344$ and 1280 are the related videos of violence and abuse in school & colleges, violence & abuse in public places and ragging in school & colleges respectively. We analyze the collected main videos with the help of annotator and label the videos as harassment and categorized the dataset (harassment videos) into four categories of harassment (vulgar, violence in school and colleges, violence in public places, ragging in school & college). We classify $71, 121, 55$ and 71 videos as vulgar, violence and abuse in school & colleges, violence and abuse in public places and ragging in school & colleges respectively. For the testing dataset we acquire data from YouTube API similar to training dataset. First we collect some harassment videos' videoIDs and then download the metadata of the collected videoIDs and top 25 related videos of their corresponding videoIDs as well as most viewed videos (YouTube Chart[1]). We are able to fetch a total of $960, 1256, 1561$ and 1398 videos (total 5175 videos for all categories) from vulgar, violence in school and colleges, violence in public places and ragging videos respectively.

Our training dataset contain videos of only positive class i.e harassment videos. And testing dataset contains both videos of harassment class as well as unknown class dataset to find false positives and true negatives. Therefore normal distribution technique of $40 : 60$ ratio to build

[1] http://www.youtube.com/charts

training and testing dataset cannot be applied. Due to the fact we have very less number of training dataset as compared to testing dataset. The size of the training and testing dataset for the four class of videos are shown in Table 4.1.

Table 4.1: Experimental Dataset

	Training Dataset	**Related Videos(Training Dataset)**	**Testing Dataset**
VVD	71	1752	960
VAVDS	121	2025	1256
VAVDP	55	1344	1561
RVDC	71	1208	1398

4.2 Evaluation Metric

To measure the effectiveness of our solution approach we used Standard performance measures i.e. confusion matrix where each rows of the matrix represents the instances of the actual class and each column represents the instances of the predicted class. In case of one class classifier each instance can only assigned to either target class (harassment) or unknown class. Precision of a class X is the ratio of number of videos correctly classified to the total predicted as videos of class X. Recall of class X is the ratio of the number of videos correctly classified to the number of videos present in class X. Table 4.2 shows the standard confusion matrix. Accuracy is the proportion of true results with both true and false results.

Let ’a’ represents the number of harassment videos correctly classified as harassment type, ’b’ represents the number of harassment videos incorrectly classified as unknown type, ’c’ represents the number of unknown videos incorrectly classified as harassment type, and ’d’ represents the number of unknown videos correctly classified as unknown.

Table 4.2: Standard Confusion Matrix

		Predicted		
		Harassment	Unknown	Total
Actual	Harassment	a	b	a+b
	Unknown	c	d	c+d
	Total	a+c	b+d	

- True Positive(TP) = a/a+b
- True Negative(TN) = d/c+d
- False Positive(FP) = b/a+b
- False Negative(FN) = c/c+d
- Accuracy = a+d/a+b+c+d

4.3 Classifier Accuracy Results

4.3.1 Vulgar Video Detection (VVD) Classifier

Table 4.3 shows the confusion matrix for VVD classifier. We run VVD classifier for 960 (videos) testing dataset. Among them classifier classifies 278 videos as vulgar harassment videos and 682 videos as unknown class based on all discriminatory features' value. We annotate all these videos and based on annotation we found that 71 (10.6%) and 88 (29.8%) videos are wrongly classified as harassment and unknown respectively using classifier. The reason of this misclassification is presence of noisy data such as misspell mistakes or misleading information because of commercial video that cannot be classify as vulgar harassment.

Table 4.3: Confusion Matrix for VVD Classifier

		Predicted	
		Vulgar Harassment	**Unknown**
Actual	**Vulgar Harassment**	207	88
	Unknown	71	594

4.3.2 Violence and abuse Video Detection in School & Colleges (VAVDS) Classifier

Table 4.4 illustrates the confusion matrix for VAVDS classifier. We run VAVDS classifier for 1256 testing dataset. Based on discriminatory feature, VAVDS classifier classifies 388 as violence harassment in school & colleges and 867 or unknown class. We annotate these videos and we conclude that 93 (10.9%) and 107 (26.6%) videos are wrongly classified as violence harassment and unknown respectively using classifier. The reason of this misclassification is lack of information such as violence at some other places are also considered as violence in school and college. Similarly because of misleading information such as the words *fight* is used in place of *match*, for example "wrestling fight between X and Y school".

Table 4.4: Confusion Matrix for VAVDS Classifier

		Predicted	
		Violence Harassment	**Unknown**
Actual	**Violence Harassment**	295	107
	Unknown	93	760

4.3.3 Violence and abuse Video Detection in Public Places (VAVDP) Classifier

Table 4.5 shows the confusion matrix for VAVDP classifier. We run our classifier for 1561 videos. Among them our classifier classifies 254 videos as violence harassment videos in public places and 1315 videos belong to unknown class. We annotate all these videos and based on annotation we result that 71 (5.4%) and 91 (33.2%) videos are wrongly classified as violence harassment and unknown respectively using classifier. The reason of this misclassification is lack of information such as violence in school and college is also considered as violence at public place.

Table 4.5: Confusion Matrix for VAVDP Classifier

		Predicted	
Actual		**Violence Harassment**	**Unknown**
	Violence Harassment	183	91
	Unknown	71	1224

4.3.4 Ragging Video Detection in School & Colleges (RVDC) Classifier

Table 4.6 illustrates the confusion matrix for RVDC classifier. We run our RVDC classifier for 1398 videos taken from the testing dataset. The classifier classifies 78 videos as ragging harassment videos and 1320 videos as unknown class. With the help of annotators, we found that 22 (1.65%) and 15 (21.1%) videos are wrongly classified as ragging harassment and unknown class respectively using classifier.

The performance of all the classifier is computed in terms of Specificity and Sensitivity. Sensitivity measures the proportion of actual positives which are correctly identified as positive and specificity measures the proportion of negatives which are correctly identified as negative are computed in terms of TPR (true positive rate), TNR (true negative rate), PPV (positive predicted value), NPV (negative predicted value) and F1-Score. TPR (i.e. recall) and TNR are also termed as sensitivity and specificity respectively. PPV (Precision) is proportion of true positives to combined true and false positives. NPV is proportion of negative prediction, that are true negative. F1-Score is the weighted harmonic mean between precision and sensitivity.

Table 4.6: Confusion Matrix for RVDC Classifier

		Predicted	
Actual		**Ragging Harassment**	**Unknown**
	Ragging Harassment	56	15
	Unknown	22	1305

Table 4.7: Accuracy Results.

Classifier	**TPR**	**TNR**	**PPV**	**NPV**	**F1-Score**	**Accuracy**
VVD	0.70	0.89	0.74	0.87	0.72	0.83
VAVDS	0.73	0.89	0.76	0.88	0.75	0.84
VAVDP	0.66	0.95	0.72	0.93	0.69	0.90
RVDC	0.79	0.98	0.72	0.99	0.75	0.97

Table 4.7 shows the performance and accuracy results of each classifier based on the given

parameter i.e. TPR, TNR, PPV, NPV, F1-score and Accuracy. Accuracy for VVD classifier is approximately 83%. Similarly accuracy for VAVDS claasifier is 84% and accuracy for VAVDP, RVDC are 90% ,93% respectively.

Chapter 5

Conclusions

In this thesis, we present an approach based on one-class classifier to detect the privacy invading harassment and misdemeanour videos having objectionable content on YouTube. Our empirical analysis results reveal that accuracy of proposed approach is more than 80%. It indicates that the presence of discriminatory features can be used to exploit the harassment detection on YouTube. We propose 13 discriminatory features based on our manual analysis and visual inspection for each category. Our empirical analysis on real world test dataset using YouTube APIs and performance of classifier reveal that certain features like linguistic features of the video and temporal based features are more influential for the accuracy of the proposed solution approach.

Chapter 6

Future Work

The proposed solution approach for the detection of privacy invading harassment on YouTube results accuracy upto 80%. The future work of this approach is to improve the accuracy of the proposed classifier. To achieve the accuracy upto 100% there is a need to analyse more contextual features like analysis of threaded comments, more contextual features of related videos and trust based feature for VVD claasifier etc.

The work presented in this report is limited to detect either video belongs to harassment or unknown class. The future work will be to find more category of the harassment and to make single classifier for the detection of harassment for all given sub-problems i.e VVD, VAVDS, VAVDP and RVDC.

Bibliography

[1] Chaudhary, V., and Sureka, A. Contextual feature based one-class classifier approach for detecting video response spam on youtube. In *Privacy, Security and Trust (PST), 2013 Eleventh Annual International Conference on* (2013), IEEE, pp. 195–204.

[2] Chen, Y., Zhou, Y., Zhu, S., and Xu, H. Detecting offensive language in social media to protect adolescent online safety. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)* (2012), IEEE, pp. 71–80.

[3] Chisholm, J. F. Cyberspace violence against girls and adolescent females. *Annals of the New York Academy of Sciences 1087*, 1 (2006), 74–89.

[4] Dadvar, M., and De Jong, F. Cyberbullying detection: a step toward a safer internet yard. In *Proceedings of the 21st international conference companion on World Wide Web* (2012), ACM, pp. 121–126.

[5] Deniz, O., Serrano, I., Bueno, G., and Kim, T. Fast violence detection in video.

[6] Dinakar, K., Reichart, R., and Lieberman, H. Modeling the detection of textual cyberbullying. In *The Social Mobile Web* (2011).

[7] Eickhoff, C., and de Vries, A. P. Identifying suitable youtube videos for children. *3rd Networked and electronic media summit (NEM)* (2010).

[8] Fu, T., Huang, C.-N., and Chen, H. Identification of extremist videos in online video sharing sites. In *Intelligence and Security Informatics, 2009. ISI'09. IEEE International Conference on* (2009), IEEE, pp. 179–181.

[9] Giannakopoulos, T., Pikrakis, A., and Theodoridis, S. A multimodal approach to violence detection in video sharing sites. In *Pattern Recognition (ICPR), 2010 20th International Conference on* (2010), IEEE, pp. 3244–3247.

[10] Khan, S. S., and Madden, M. G. A survey of recent trends in one class classification. In *Artificial Intelligence and Cognitive Science*. Springer, 2010, pp. 188–197.

[11] Kim, C.-Y., Kwon, O.-J., Kim, W.-G., and Choi, S.-R. Automatic system for filtering obscene video. In *Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference on* (2008), vol. 2, IEEE, pp. 1435–1438.

[12] Kontostathis, A., Edwards, L., and Leatherman, A. Text mining and cybercrime. *Text Mining: Applications and Theory. John Wiley & Sons, Ltd, Chichester, UK* (2010).

[13] Lee, S., Shim, W., and Kim, S. Hierarchical system for objectionable video detection. *Consumer Electronics, IEEE Transactions on 55*, 2 (2009), 677–684.

[14] Mahmud, A., Ahmed, K. Z., and Khan, M. Detecting flames and insults in text.

[15] Roher, E. M. Confronting facebook, youtube and myspace: Cyberbullying in schools. *Principal Connections* (2007), 19–21.

[16] Singhal, P., and Bansal, A. Improved textual cyberbullying detection using data mining.

[17] Sood, S. O., Churchill, E. F., and Antin, J. Automatic identification of personal insults on social news sites. *Journal of the American Society for Information Science and Technology 63*, 2 (2012), 270–285.

[18] Sureka, A., Kumaraguru, P., Goyal, A., and Chhabra, S. Mining youtube to discover extremist videos, users and hidden communities. In *Information Retrieval Technology*, P.-J. Cheng, M.-Y. Kan, W. Lam, and P. Nakov, Eds., vol. 6458 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2010, pp. 13–24.

[19] Yin, D., Xue, Z., Hong, L., Davison, B. D., Kontostathis, A., and Edwards, L. Detection of harassment on web 2.0. *Proceedings of the Content Analysis in the WEB 2* (2009).