2016

Identifying Offensive Videos on YouTube

Rajeshwari Kandakatla
Wright State University

Identifying Offensive Videos on YouTube

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

By

Rajeshwari Kandakatla

B.Tech., Kakatiya University, 2014

2016
Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

December 15, 2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Rajeshwari Kandakatla ENTITLED Identifying Offensive Videos on YouTube BE ACCEPT -ED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science.

Krishnaprasad Thirunarayan, Ph.D.
Thesis Director

Mateen Rizki
Chair, Department of Computer Science and Engineering

Committee on
Final Examination

Krishnaprasad Thirunarayan, Ph.D.

Amit Sheth, Ph.D.

Valerie L. Shalin, Ph.D.

Robert E.W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

Abstract

Kandakatla, Rajeshwari. M.S. Department of Computer Science and Engineering, Wright State University, 2016. Identifying Offensive Videos on YouTube.

Harassment on social media has become a critical problem and social media content depicting harassment is becoming common place. Video-sharing websites such as YouTube contain content that may be offensive to certain community, insulting to certain religion, race etc., or make fun of disabilities. These videos can also provoke and promote altercations leading to online harassment of individuals and groups.

In this thesis, we present a system that identifies offensive videos on YouTube. Our goal is to determine features that can be used to detect offensive videos efficiently and reliably. We conducted experiments using content and metadata available for each YouTube video such as comments, title, description and number of views to develop Naïve Bayes and Support Vector Machine classifiers. We used training dataset of 300 videos and test dataset of 86 videos and obtained a classification F-Score of 0.86. It was surprising to note that sentiment and content of the comments were less effective in detecting offensive videos than the unigrams and bigrams in the video title and any other feature combinations does not improve the performance appreciably.Thus, the simplicity of these features contributes to the efficiency of computation and implies that the up-loaders provide good titles.

Contents

List of Figures

List of Tables

Acknowledgements

I like to express my sincere gratitude to my thesis advisor Dr. Krishnaprasad Thirunarayan for this opportunity. I am always thankful for his incredible support and patience throughout my research. He always took time out from a busy schedule to keep me on the right track by his valuable guidance and resources.

I would also like to thank Dr. Amit Sheth and Dr. Valerie Shalin, for their valuable feedback and encouragement.

I owe my thanks to Dr. Lu Chen, Dr. Sumanth Kulkarni and Dr. Wenbo Wang for their valuable help throughout my project. I would like to thank the whole Kno.e.sis family for answering my questions patiently.

I would also like to thank my brother K. Kaushik for all his motivation and encouragement.

Chapter 1

Introduction

Social media tools allow people, companies and other organizations to create, share or exchange information, such as career interests, ideas, pictures/videos in virtual communities and networks[1]. There are many social-networking websites, video sharing sites and blogs that provide interactive platform for virtual communities that have gained significant popularity in the recent past. According to a recent survey[2], on an average, an American spends more than 3 hours in a day exclusively on social networking websites.

Websites such as, Facebook, Twitter, YouTube, LinkedIn and Pinterest provide effective communication platform among people all around the world. The rapid growth in the usage and adaption of these social media websites can be attributed to the ease of networking through the devices such as smart phones and tablets. On the other hand, misuse of these websites to insult, harass or harm others is not uncommon since these websites are used by public at large and topics of conversation can be polarizing or controversial.

[1] https://en.wikipedia.org/wiki/Social_media

[2] http://www.marketingcharts.com/online/social-networking

"EVERYTHING you post on social media impacts your PERSONAL BRAND. How do you want to be known?"

–Lisa Horn

Recent studies have shown that, YouTube, Vimeo, Dailymotion and Twitch are popular video sharing websites today. Among them, YouTube is the largest and global video-sharing website where 400 hours of videos are uploaded every minute and one-third of the population on Internet use YouTube[3].

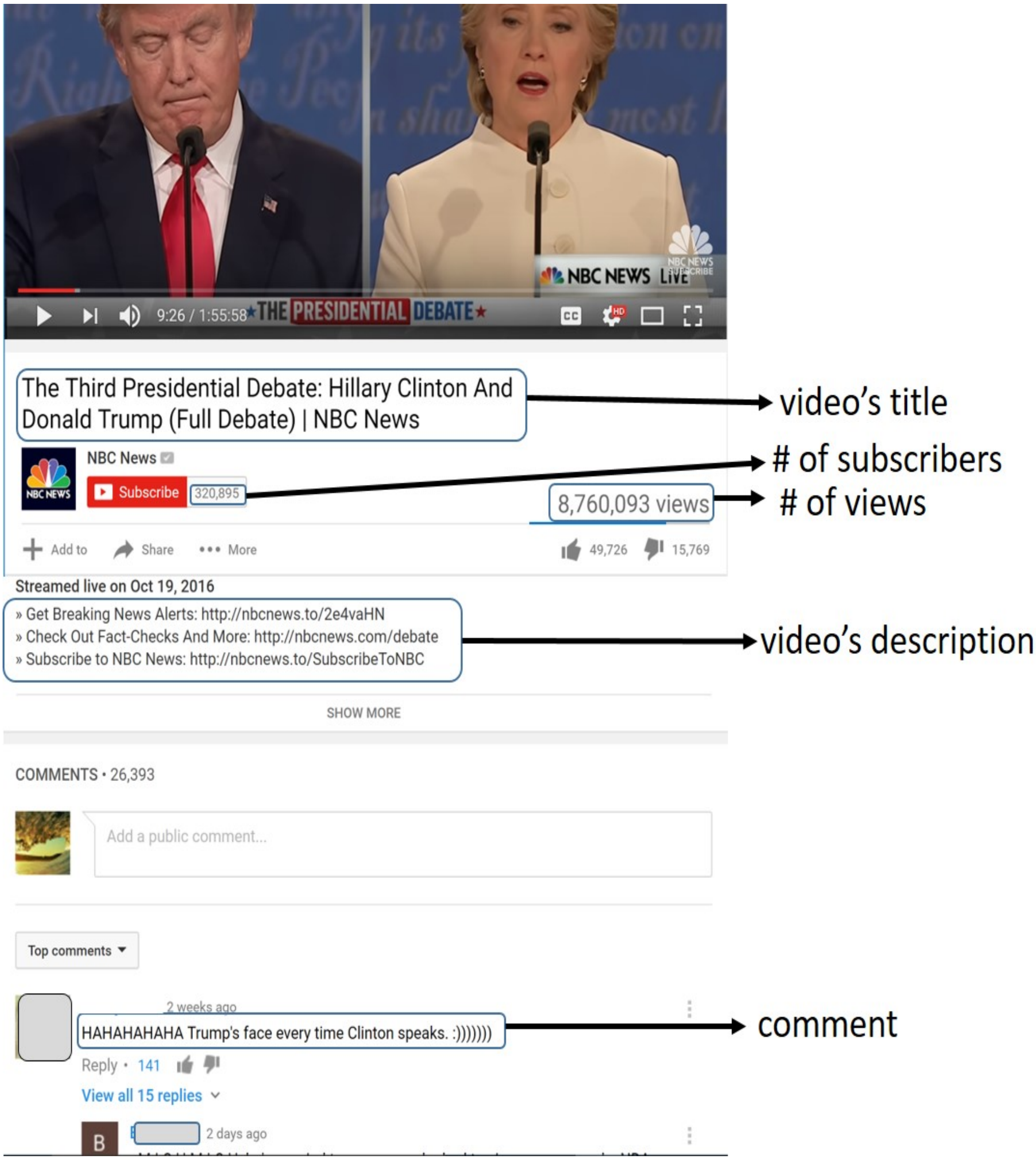
1.1 YouTube

YouTube is one of largest video-sharing websites today. In the United States, the number of adults watching television has declined compared to the number of people watching YouTube videos on daily basis. This is because of wide availability of YouTube on different smart devices like mobile phones, tablets and smart TVs that allow us to watch whatever we want, wherever we want and whenever we want. Many companies also take advantage of this by posting their ads in the middle of YouTube videos, in fact, one of significant ways of monetization. YouTube allows its users to upload, rate, share, view, subscribe and comment on the videos. It has no restrictions on the number of views, shares or uploads of videos. Here videos can be of a variety of type like movies, recording clips and animations. According to YouTube statistics[4], YouTube has over billion users and has been released locally in more than 88 countries in 76 different

[3] http://21-amazingly-interesting-youtube-facts-2016/

[4] http://21-amazingly-interesting-youtube-facts-2016/

languages which covers 95% of the Internet population. These statistics clearly show the outstanding popularity of YouTube. YouTube users can perform different activities on the website like post comments, interact with other users by replying to their comments, subscribe to a variety of channels and rate videos. Information about the video include title of the video, description of the video,category



The Third Presidential Debate: Hillary Clinton And Donald Trump (Full Debate) | NBC News

NBC News

Streamed live on Oct 19, 2016

» Get Breaking News Alerts: http://nbcnews.to/2e4vaHN
» Check Out Fact-Checks And More: http://nbcnews.com/debate
» Subscribe to NBC News: http://nbcnews.to/SubscribeToNBC

COMMENTS • 26,393

2 weeks ago

HAHAHAHAHA Trump's face every time Clinton speaks. :))))))

Figure 1.1: Features of YouTube

of the video, comments and replies to them. A glimpse of all these activities is shown in Figure 1.1.

There is no apriori filter to check the contents of the video that are being uploaded. This unfiltered access is the root cause for existence of videos that insult some religion, criticize race, make sexiest comments and make fun of disability. This can hurt the emotions of a large group of people and sometimes even have serious impact on individuals. There is a need for a framework to automatically remove such videos from being uploaded on YouTube or atleast have the ability to filter from view on a voluntary basis.

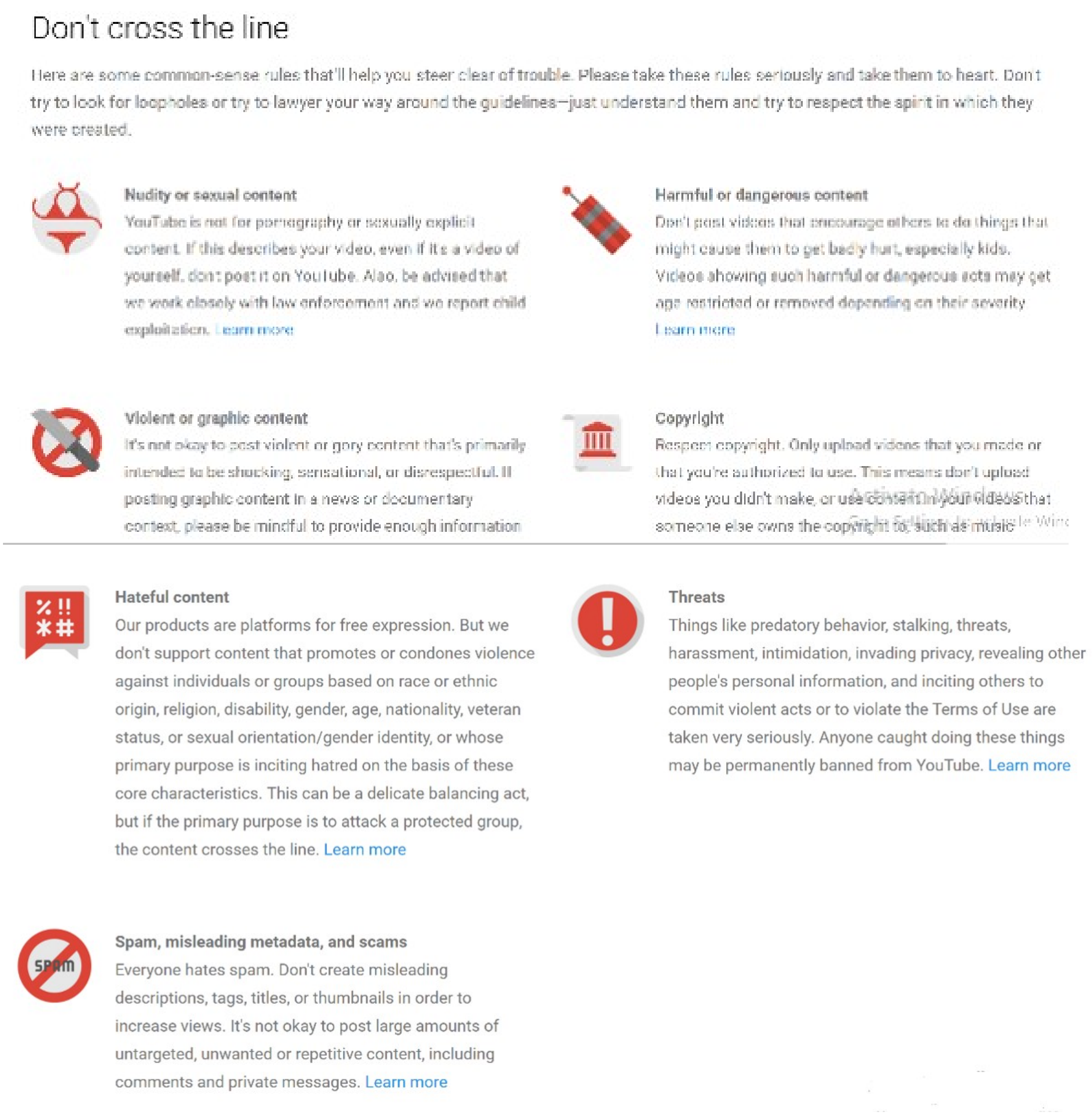
1.2 Offensive videos on YouTube

Offensive videos are those that create a negative impact on a large group of people after watching the video. YouTube grants permission to users to upload videos without checking the content of the videos. Users take advantage of this freedom to insult a minority community by uploading recordings that were captured without their consent. Such activities can offend and target a community.

YouTube has a set of community guidelines aimed to reduce the aforementioned videos on their website. Despite these guidelines, there is significant proportion of videos that violate these guidelines. Figure 1.2 describes some of the community guidelines on YouTube. For example, two high school students were suspended because of the racist videos they uploaded to YouTube[5].

[5] http://www.huffingtonpost.com/post

[6] https://www.youtube.com/yt/policyandsafety/communityguidelines.html



Don't cross the line

Here are some common-sense rules that'll help you steer clear of trouble. Please take these rules seriously and take them to heart. Don't try to look for loopholes or try to lawyer your way around the guidelines—just understand them and try to respect the spirit in which they were created.

Nudity or sexual content
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. Learn more

Harmful or dangerous content
Don't post videos that encourage others to do things that might cause them to get badly hurt, especially kids. Videos showing such harmful or dangerous acts may get age restricted or removed depending on their severity. Learn more

Violent or graphic content
It's not okay to post violent or gory content that's primarily intended to be shocking, sensational, or disrespectful. If posting graphic content in a news or documentary context, please be mindful to provide enough information

Copyright
Respect copyright. Only upload videos that you made or that you're 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

Hateful content
Our products are platforms for free expression. But we don't support content that promotes or condones violence against individuals or groups based on race or ethnic origin, religion, disability, gender, age, nationality, veteran status, or sexual orientation/gender identity, or whose primary purpose is inciting hatred on the basis of these core characteristics. This can be a delicate balancing act, but if the primary purpose is to attack a protected group, the content crosses the line. Learn more

Threats
Things like predatory behavior, stalking, threats, harassment, intimidation, invading privacy, revealing other people's personal information, and inciting others to commit violent acts or to violate the Terms of Use are taken very seriously. Anyone caught doing these things may be permanently banned from YouTube. Learn more

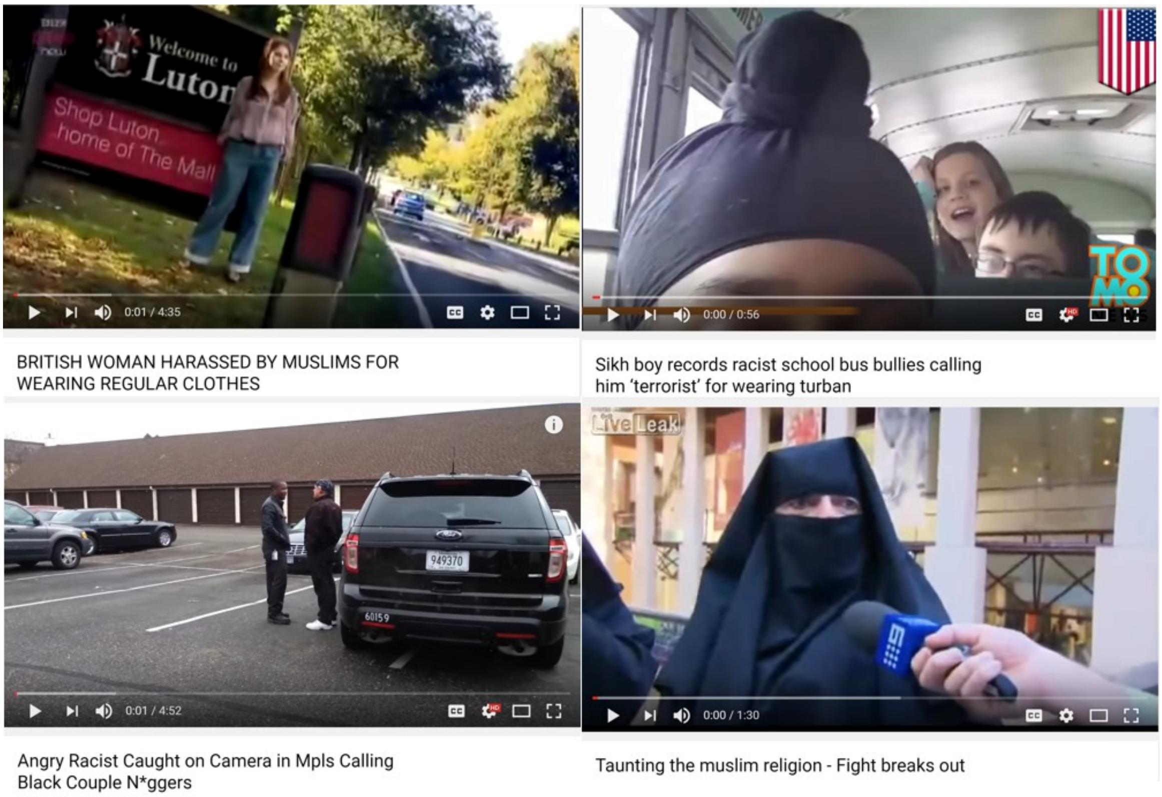
Spam, misleading metadata, and scams
Everyone hates spam. Don't create misleading descriptions, tags, titles, or thumbnails in order to increase views. It's not okay to post large amounts of untargeted, unwanted or repetitive content, including comments and private messages. Learn more

Figure 1.2: YouTube community guidelines [6]

1.3 Motivation

YouTube is the second most used popular website after Facebook. Due to low publication barriers, anonymous users can potentially misuse this website by uploading harmful content which has serious effect on a large group of people. Figure 1.3 shows some instances of offensive videos.

In fact, there are some YouTube services where one can report offensive video (Figure 1.4). However, this is not very effective to



BRITISH WOMAN HARASSED BY MUSLIMS FOR WEARING REGULAR CLOTHES

Sikh boy records racist school bus bullies calling him 'terrorist' for wearing turban

Angry Racist Caught on Camera in Mpls Calling Black Couple N*ggers

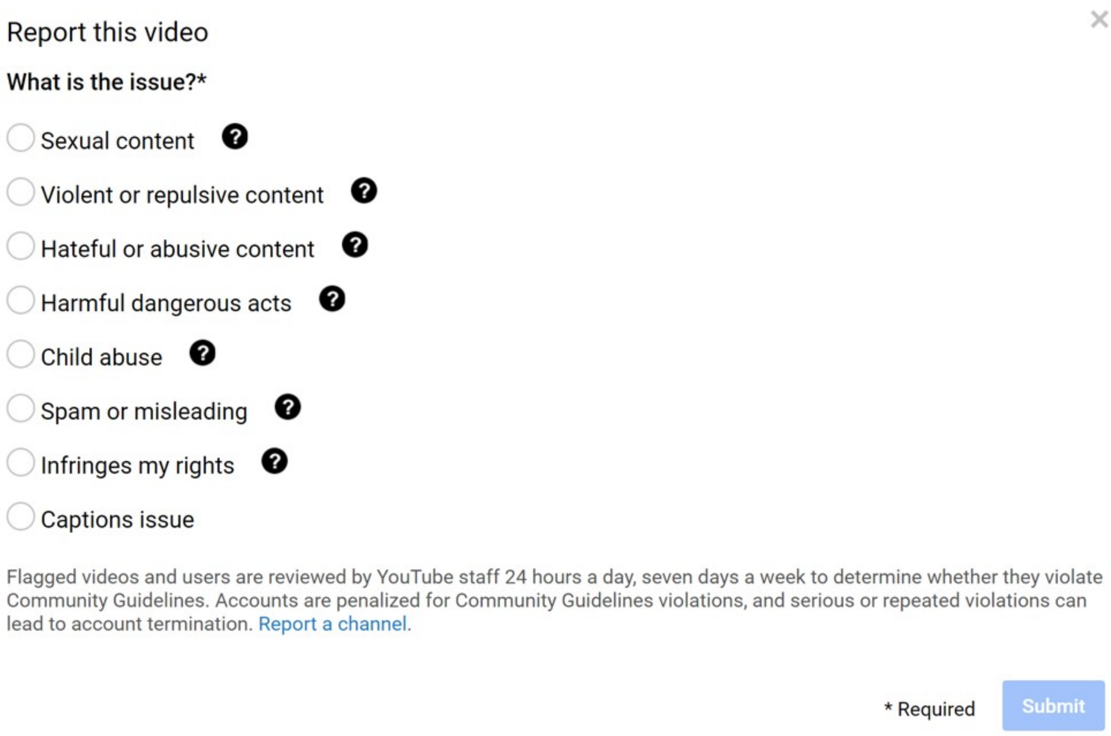
Taunting the muslim religion - Fight breaks out

Figure 1.3: Examples of offensive videos

flag offensive videos, given the delays in processing the requests for removing videos from the website.

There is a need for techniques and tools to automatically identify offensive videos on YouTube. Our work has been motivated by the following facts:

1. Despite many community guidelines, we see many videos that are offensive to a large group of people readily available.
2. We find instances where these videos have had serious negative effect on people's lives such as suspension from school and suspension from work.(See Figure 1.5)
3. These videos can lead to subsequent discussions /conversations /altercation (in comments section) which can further fuel and



Report this video

What is the issue?*

○ Sexual content

○ Violent or repulsive content

○ Hateful or abusive content

○ Harmful dangerous acts

○ Child abuse

○ Spam or misleading

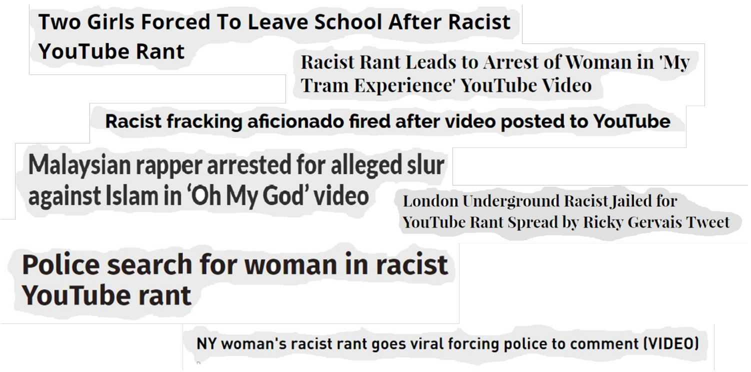
○ Infringes my rights

○ Captions issue

Flagged videos and users are reviewed by YouTube staff 24 hours a day, seven days a week to determine whether they violate Community Guidelines. Accounts are penalized for Community Guidelines violations, and serious or repeated violations can lead to account termination. Report a channel.

* Required

Figure 1.4: Snapshot of reporting video to YouTube community



Two Girls Forced To Leave School After Racist YouTube Rant

Racist Rant Leads to Arrest of Woman in 'My Tram Experience' YouTube Video

Racist fracking aficionado fired after video posted to YouTube

Malaysian rapper arrested for alleged slur against Islam in 'Oh My God' video

London Underground Racist Jailed for YouTube Rant Spread by Ricky Gervais Tweet

Police search for woman in racist YouTube rant

NY woman's racist rant goes viral forcing police to comment (VIDEO)

Figure 1.5: Consequences

harassment of groups and individuals, both online and in the physical world.

4. Factors such as racist and demeaning videos are responsible for the rise in the terrorism.

1.4 Outline

In this thesis, we propose a system that identifies offensive videos on YouTube. We determine a set of useful features and then design a classifier that identifies offensive videos. Further, we evaluate the results of our classifier.

Here is the organization for the rest of this document. In Chapter 2, we present related works that focus on YouTube and video-sharing websites. In Chapter 3, we mention different types of data available publicly on YouTube. We also discuss various techniques to filter noisy comments and our approach to identify offensive videos using machine learning algorithms. In Chapter 4, we evaluate the performance of the approach discussed in Chapter 3. In Chapter 5, we present our conclusions and insights of and discuss possible future research directions.

Chapter 2

State of the art

We summarize the literature in the area of cyber bullying and harassing content detection in social media. Most of the techniques on identifying textual cyberbullying are based on analyzing the comments and content(images).

2.1 Offensive textual content detection on YouTube

Chen et al. [1] introduced the idea of identifying offensive language in social media. They proposed a lexical syntactic architecture which incorporates both message-level and user-level features to identify offensive messages and offensive users. It is based on the idea that in a sentence if profanities are associated with user or contain other offensive word, then the sentence is offensive. For this purpose, this approach uses two dictionaries, i.e., strong and weak offensive word dictionaries. For instance, *f**k* and *s**k* come under strong profanities, *stupid* and *liar* are weak profanities. Table 2.1 shows some of the examples of offensive texts.

Table 2.1: Examples of offensive texts

Comment	Comment structure	Result
You stupid!	user identifier + weak offensive word	offensive message
Game is stupid	weak offensive word	not an offensive message
Fucking stupid	strong offensive word + weak offensive	offensive message

The architecture assigns different weights for strong and weak offensive words. It uses Stanford dependency parser to identify association of offensive word/s or user identifier to offensive word and defined offensive levels (defined as intensity) for:

- user identifier associated with offensive word/s
- other offensive word/s associated with offensive word

Message-level offensiveness (O_s) is given by:

$$O_s = \sum O_w I_w, \tag{2.1}$$

where, O_w is the offensive level of the word and I_w is the user-identifier/offensive word.

This approach aggregates sentence offensiveness values to find user offensiveness. It also extracts style features (ratio of short sentences, punctuations, uppercase letters w.r.t all sentences), structural features (ratio of imperative sentences, nouns, verbs), content-specific features (race, religion, violence, clothes, accent) to determine overall user-level offensiveness.

Dinakar et al. [2] proposed an approach to detect textual cyberbullying. This approach first uses binary classifier to identify sensitive text. After spotting sensitive text, this framework builds topic sensitive classifiers by analyzing textual comments posted on a topic. This approach makes use of a list of profane words, TF-IDF, lexicon of negative words and label-specific unigrams, bigrams as features. This model couldn't distinguish sarcasm posts since they do not contain profane or negative words. For example, given below is a comment of video about a famous politician.

Example:*He is an expert in tossing coins.*

Dadvar et al. [3] improved detection of cyberbullying with user information. They collected top three videos from different video categories such as entertainment, politics and sports. This framework comprises of three sets of features, namely, content-based features, cyber-bullying features and user-based features. Content-based features are composed of number of profane words, first and second person pronouns, profanity window of different sizes and the number of emotions. Cyber-bullying features include number of cyberbullying words and phrases. User-based features involve history of user's comments along with age of user.

Reynolds et al. [4] proposed an approach to detect cyberbullying using machine learning. It discovers language patterns used by bullies and victims based on hard coded rules. The idea of this work is to focus on number of bad words along with their intensity.

The aforementioned works on cyberbullying use YouTube dataset to detect cyberbullying. We noticed that these works consider every post as an individual post rather than a group of posts associated with a video. This leads to misclassification due to missing context of posts.

Example: *She is heavily built... OMG!*

Positive comment in video titled “World’s biggest container ship CSCL globe maiden call”.

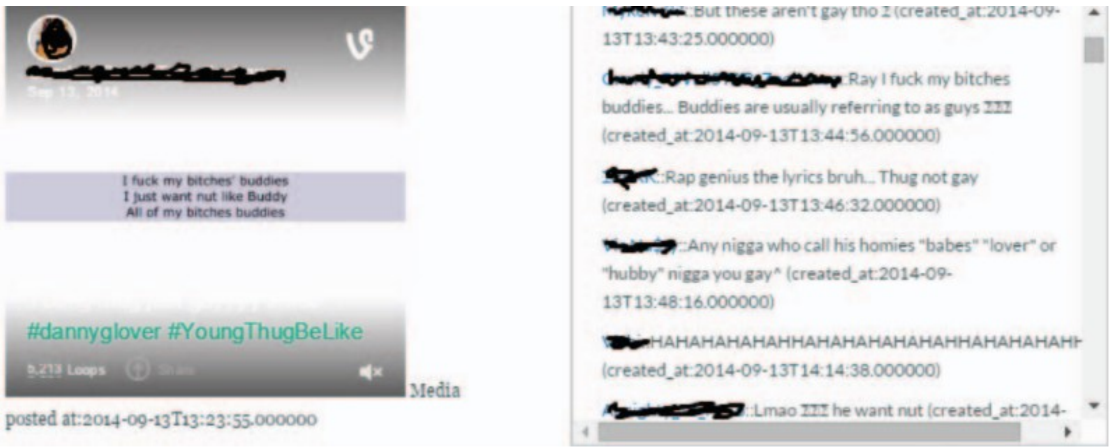
Negative comment in video titled “Old man insults fat woman”.

2.2 Video-sharing websites

Agarwal et al. [5] focused on detecting privacy invading harassment and misdemeanor videos. This work decomposes problem of identifying objectionable content as a problem of identifying vulgar videos, abuse and violence in public places and ragging videos in schools and colleges. Feature set of their approach includes linguistic (percentage of keywords in title and description), popularity (ratio of likes and views, ratio of comments and views), duration (duration of video) and category (YouTube category). It uses one class classifier to detect objectionable videos.

Rafiq et al. [6] designed a technique to detect cyberbullying instances in Vine[1], a mobile based video-sharing social network that allows user to record and edit six-second videos. A screenshot of Vine media session is shown in Figure 2.1. This work distinguishes cyberbullying and cyber aggression. They define cyber aggression as a type of behavior in an electronic context that is meant to intentionally harm another person [7]. Cyberbullying is defined in a stronger and more specific way as aggressive behavior that is carried

[1] https://vine.co/



[redacted]

Sep 13, 2014

I fuck my bitches' buddies
I just want nut like Buddy
All of my bitches buddies

#dannyglover #YoungThugBeLike

5,213 Loops

Media posted at:2014-09-13T13:23:55.000000

[redacted]::But these aren't gay tho Z (created_at:2014-09-13T13:43:25.000000)

[redacted]::Ray I fuck my bitches buddies... Buddies are usually referring to as guys ZZZ (created_at:2014-09-13T13:44:56.000000)

[redacted]::Rap genius the lyrics bruh... Thug not gay (created_at:2014-09-13T13:46:32.000000)

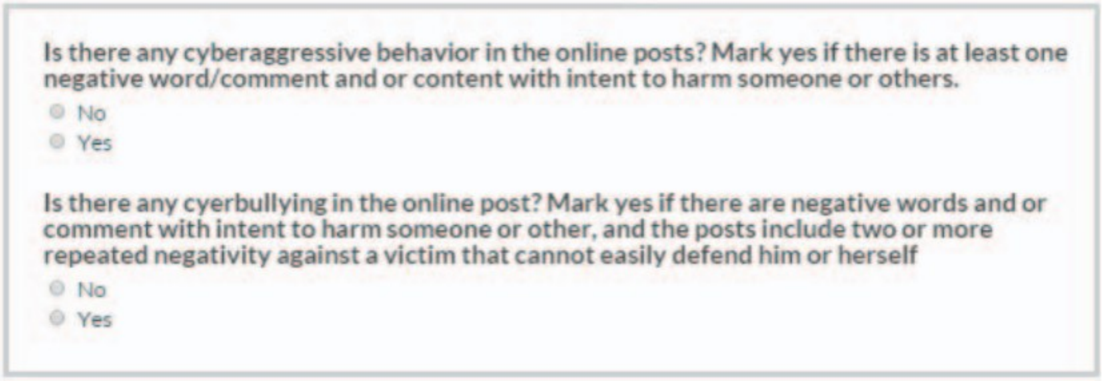
[redacted]::Any nigga who call his homies "babes" "lover" or "hubby" nigga you gay^ (created_at:2014-09-13T13:48:16.000000)

[redacted]HAHAHAHAHAHHAHAHAHAHAHAHHAHAHAHAHH (created_at:2014-09-13T14:14:38.000000)

[redacted]::Lmao ZZZ he want nut (created_at:2014-

Figure 2.1: Instance of Vine media session

out repeatedly against a person who cannot easily defend himself or herself, creating a power imbalance [7] [8]. It uses CrowdFlower[2], a crowd-sourced website to annotate media sessions for cyberbullying and cyber aggression. Labelling of media sessions is shown in Figure 2.2. Feature set of this framework includes information about media-session (e.g., number of likes and comments), profile-owner (e.g., number of followers, followings and media-posted by profile owner), comments (e.g., sentiment of each comment) and n-grams features to train the classifier.



Is there any cyberaggressive behavior in the online posts? Mark yes if there is at least one negative word/comment and or content with intent to harm someone or others.

○ No
○ Yes

Is there any cyerbullying in the online post? Mark yes if there are negative words and or comment with intent to harm someone or other, and the posts include two or more repeated negativity against a victim that cannot easily defend him or herself

○ No
○ Yes

Figure 2.2: Labelling media sessions

Fu et. al [9] proposed an approach to identify extremist videos in online video sharing sites. This work uses user-generated content

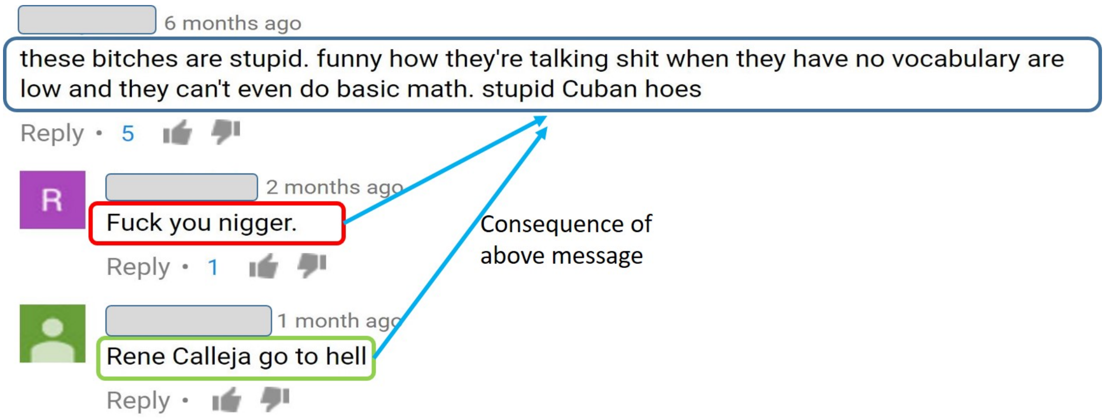
[2] https://www.crowdflower.com/

such as comments to identify extremist groups. Feature set of this work includes lexical, syntactic and content-specific features to identify extremist groups. Lexical features contain number of characters, numbers and average word length. The patterns used to form sentences, frequency of punctuations are the syntactic features of this framework. Content-specific features include important keywords and phrases. This work selects features for classification using information gain. SVM with 10-fold cross validation was adopted to classify extremist videos.

2.3 Research gap

- Most of the research studies consider message/comment as an individual entity. However, the message/comment might be part of conversation or a reply to another harassing message.(see Figure 2.3)

  We consider conversations instead of individual messages/comments. So, we can detect the original source for harassment.



[redacted] 6 months ago

these bitches are stupid. funny how they're talking shit when they have no vocabulary are low and they can't even do basic math. stupid Cuban hoes

[redacted] 2 months ago

Fuck you nigger.

[redacted] 1 month ago

Rene Calleja go to hell

Consequence of
above message

Figure 2.3: Limitation of existing approaches

- We also found that these conversations took place only because of video uploaded on YouTube. In this scenario, source of harassment is video but not messages.

  None of research studies focused on this aspect.

In this thesis, we solve the second limitation, the source of harassment, i.e., videos that has the potential of above mentioned conversations.

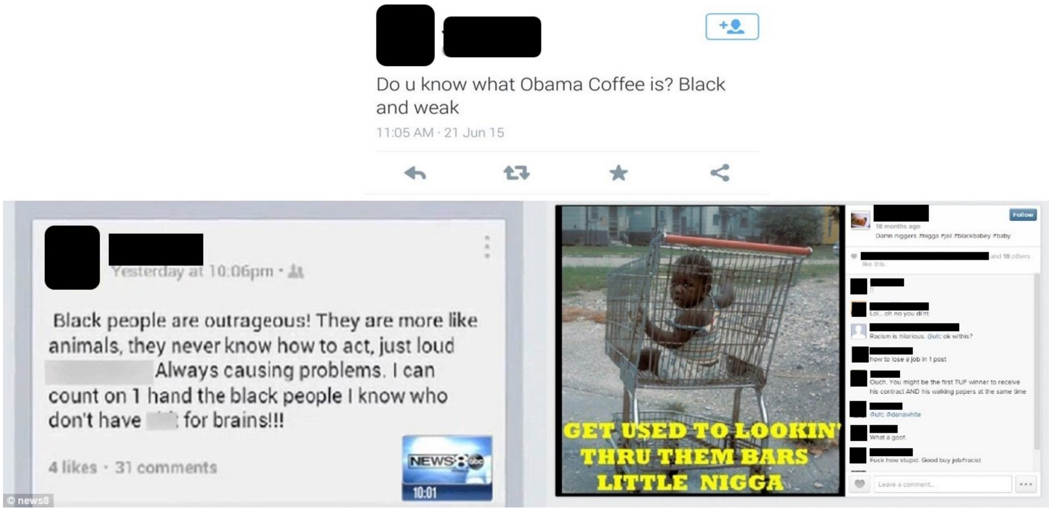
Chapter 3

Proposed Approach

We discuss an approach to detect offensive videos on YouTube. Before discussing the general framework, we define offensiveness. According to English literature, offending is an act to cause someone to feel resentful, upset or annoyed[1]. In the context of social media, offensive posts are those which cause someone to feel upset or annoyed and that create negative impact on a group of people. There are numerous instances of such posts in different social-networking websites such as Facebook, Twitter, YouTube and Instagram. Few instances of such posts are shown in Figure 3.1.

Our goal is to design a framework that detects offensive videos. Our approach is general in nature and can be applied to any video sharing website with typical features such as likes and comments. As YouTube is the world's top viewing and sharing website, we consider it as a representative sample and perform our research based on YouTube. Instances of offensive videos on YouTube are shown in Figure 3.2.

[1] https://en.oxforddictionaries.com/definition/offensive



[redacted]

Do u know what Obama Coffee is? Black and weak

11:05 AM · 21 Jun 15

[redacted]

Yesterday at 10:06pm ·

Black people are outrageous! They are more like animals, they never know how to act, just loud [redacted] Always causing problems. I can count on 1 hand the black people I know who don't have [redacted] for brains!!!

4 likes · 31 comments

[redacted]

18 months ago

Damn niggers #nigga #jail #blackbabey #baby

[redacted] and 18 others like this.

[redacted]

[redacted] Lol... oh no you dint

[redacted] Racism is hilarious. @ufc ok w/this?

[redacted] how to lose a job in 1 post

[redacted] Ouch. You might be the first TUF winner to receive his contract AND his walking papers at the same time

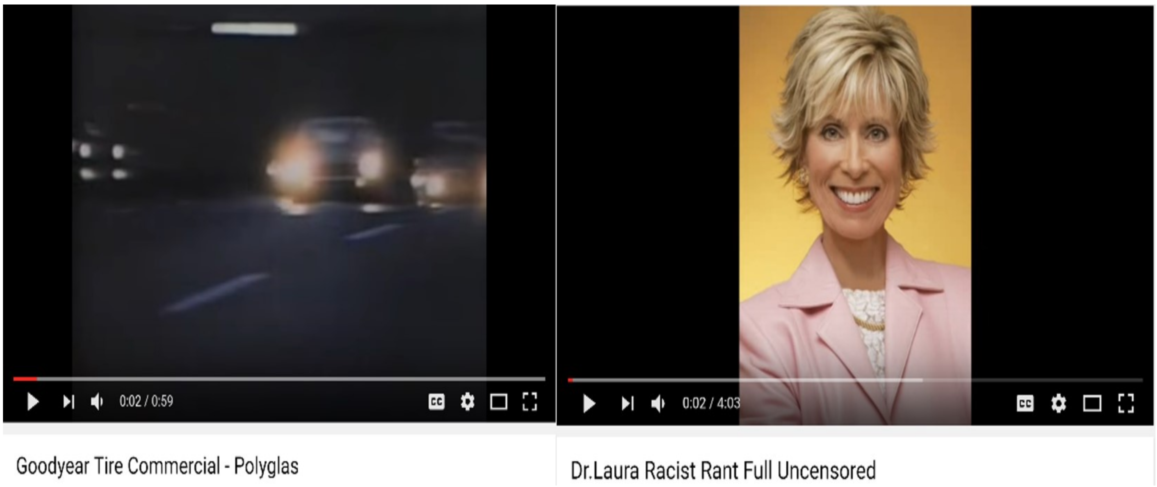
[redacted] @ufc @danawhite

[redacted] What a goof.

[redacted] Fuck how stupid. Good buy job#racist

Leave a comment...

Figure 3.1: Instances of offensive post on social media



Goodyear Tire Commercial - Polyglas

Dr.Laura Racist Rant Full Uncensored

Figure 3.2: Instances of offensive videos

We divide the problem of identifying offensive videos into three phases, namely, creation of data set, identifying features to detect offensives has reliably and efficiently and then classifying videos with the selected set of features. These phases represent a broad approach to detect offensive content on YouTube.

3.1 Dataset Creation

In order to create a dataset for our analysis, we initially started with random video ids. To get a random video id, we generate a random string and check if that is a valid video id or not using a web service. If yes, then we consider that video id for our analysis. In this way, we created the dataset for our analysis. But from manual verification, we observed that this random video id collection doesn't have sufficient number of offensive video ids for our analysis. So, we enhanced our dataset by obtaining video ids related to offensiveness using keywords shown in Figure 3.3.

Figure 3.3: Cloud of sensitive keywords[2]

We focused on controversial and sensitive-topic keywords to enhance our dataset because these kind of videos gain attention from a large number of people. We used keywords such as racism, racist slurs, feminism and black keywords to collect offensive video ids. Now we have acquired a qualitative dataset with good number of offensive video ids. We obtained a dataset of 300 video ids and gave it to domain experts for annotation. Given a video id, they were asked to annotate whether video is offensive to any community or not along with type of offence. From this dataset, we used 90 percent as training and 10 percent as validation.

This annotated dataset is used for training the classifier. But

[2] http://www.shutterstock.com/sensitive-key-words

Table 3.1: Examples of offensive texts

Video URL	**Offensive to community/not**	**Type of Offence**
https://www.youtube.com/watch?v=Y9k2CUZJDp0	yes	gender
https://www.youtube.com/watch?v=NXSKX3XX52M	yes	race
https://www.youtube.com/watch?v=t2NX9OVeAEU	yes	appearance
https://www.youtube.com/watch?v=god_1Pa8XYo	yes	religion
https://www.youtube.com/watch?v=Ren0LZHUBEY	yes	none

this process of acquiring dataset is time consuming. We then acquired test dataset in two steps.

1. We collected random video ids by generating random string and validating the string using web service.
2. We enhanced test set by combining related videos of offensive video ids that were annotated before.

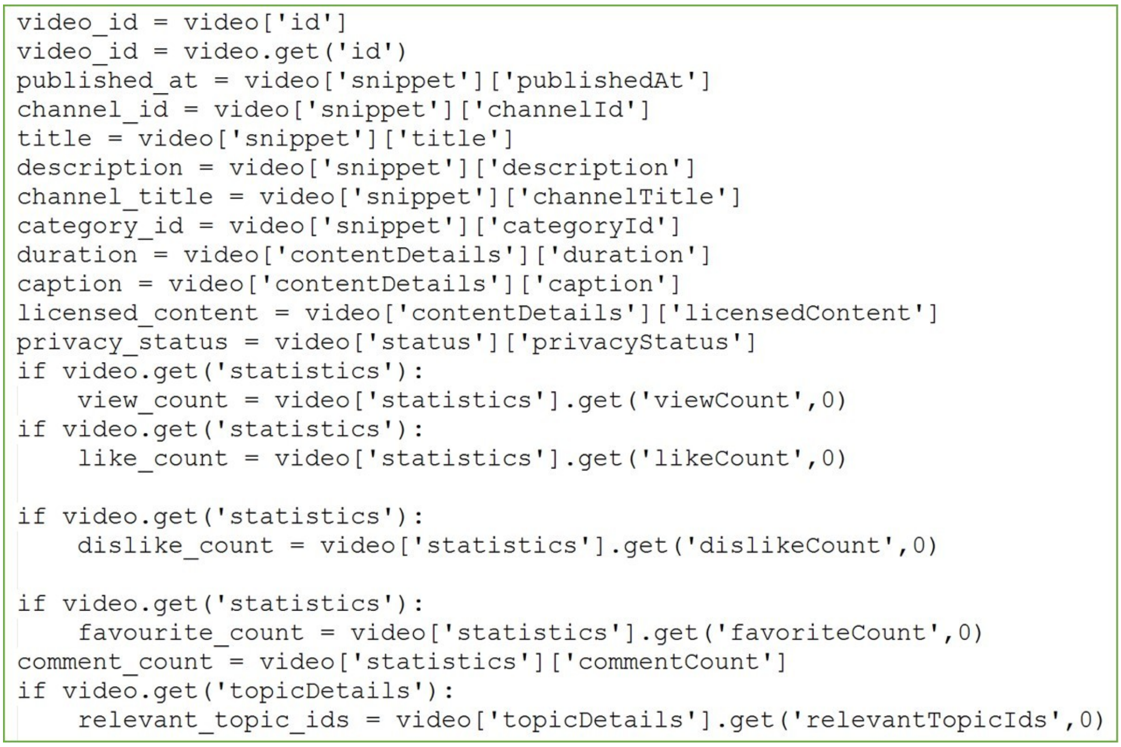
This way, we acquired training and testing datasets. Figure 3.4 shows code snippet for crawling data using YouTube API.

3.2 Identifying and Selecting Features

In this phase, we investigate all available features of YouTube videos and further select the features that are useful for our analysis.

3.2.1 Data Extraction

As YouTube is public, we can extract any information regarding the video. We use YouTube API for extracting all the data available for a YouTube video. We decomposed the feature set into two categories: comment based features and metadata based features. Comment based features involve comments and replies to the comments and metadata based features involve features such as title, description, upload date and time, channel id, number of likes, number of dislikes and number of views.



```
video_id = video['id']
video_id = video.get('id')
published_at = video['snippet']['publishedAt']
channel_id = video['snippet']['channelId']
title = video['snippet']['title']
description = video['snippet']['description']
channel_title = video['snippet']['channelTitle']
category_id = video['snippet']['categoryId']
duration = video['contentDetails']['duration']
caption = video['contentDetails']['caption']
licensed_content = video['contentDetails']['licensedContent']
privacy_status = video['status']['privacyStatus']
if video.get('statistics'):
    view_count = video['statistics'].get('viewCount',0)
if video.get('statistics'):
    like_count = video['statistics'].get('likeCount',0)

if video.get('statistics'):
    dislike_count = video['statistics'].get('dislikeCount',0)

if video.get('statistics'):
    favourite_count = video['statistics'].get('favoriteCount',0)
comment_count = video['statistics']['commentCount']
if video.get('topicDetails'):
    relevant_topic_ids = video['topicDetails'].get('relevantTopicIds',0)
```

Figure 3.4: Code snippet for crawling data using YouTube API

Preprosessing

In this sub phase, we pre-process the crawled dataset. This pre-processing involves removal of stop words, tokenization and removal of special symbols from text-based features such as comments. Tokenization is the process of breaking a sentence into meaningful elements called tokens. Tokens that are very common but that do not contribute to significant relevant content are called stop words which need to be eliminated/filtered. After tokenization, we provide a standard stop-word list to remove such words from token list. Similarly, we remove unnecessary special symbols. We used regular expression pattern to capture special symbols to be eliminated.We used stop words list from Onix Text Retrieval Toolkit[3]. Figure 3.5 shows some of frequently used stop words in English language.

Figure 3.5: Stop word list

3.2.2 Feature Selection

Feature Selection is the process of selecting a subset of relevant features that help in identifying offensive videos and that can be used in the classification task. We perform experiments to identify relevant features for our purpose and subsequently evaluate them.

We focused more on text based features compared to image/frame. We usually use image processing to extract meaningful information

[3] http://www.lextek.com/manuals/onix/stopwords2.html

from the image/frame. However, we did not perform such analysis because our goal in identifying offensive videos is more related to the topic being discussed/conversations. That is, our focus was more on what was discussed in the video. On the other hand, analyzing or processing this images/frames gives no idea about the discussion in the video. So, analyzing video frames is not a good discriminatory feature for our problem.

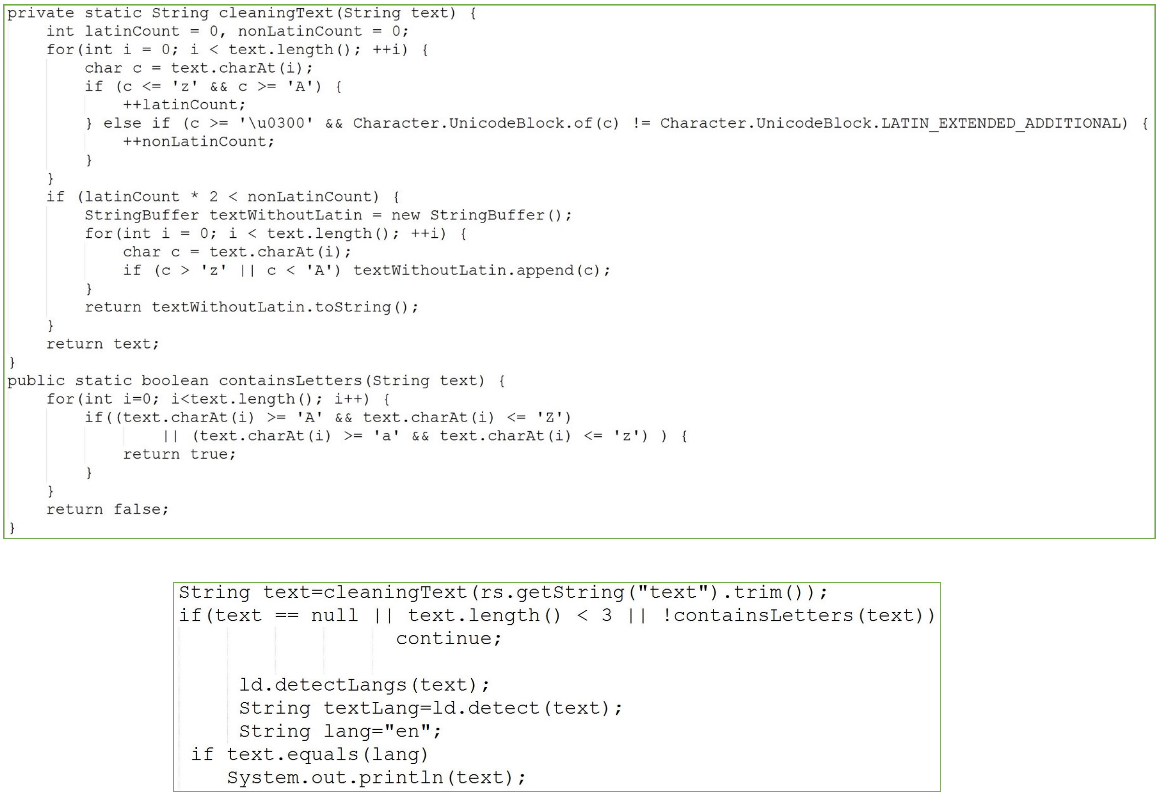
As discussed above, our focus was more on what was discussed in the video. For this reason, we can make use of audio transcripts of the video. But, we didn't use them because that requires processing the video which is computationally expensive.

3.2.2.1 Comment Based Features

Our next focus was on comments of each YouTube video. Comments are messages/reviews that are written after watching a YouTube video. These are the statements that describe or discuss the video. Along with comments, YouTube allows reply to each posted comment. YouTube users reply to the posted comments, supporting or criticizing them and thus give rise to conversations.

On closer scrutiny, we discovered that initial comments discuss the content of the video while subsequent replies end up discussing other comments or quarrelling with previous commenters. The top-level comments are highly relevant to the content of the video, while subsequent replies are less relevant or biased or contain attacks on previous comments. In order to improve relevance and consistency of discrimination of a video, we focused on top-level comments and ignored reply-comments.

As YouTube supports 76 different languages, comments can be in any of these languages. But, we restrict our analysis to only English comments. To filter English comments from all other comments, we used language detector web service. Figure 3.6 is the code snippet for retrieving only English comments.



```
private static String cleaningText(String text) {
    int latinCount = 0, nonLatinCount = 0;
    for(int i = 0; i < text.length(); ++i) {
        char c = text.charAt(i);
        if (c <= 'z' && c >= 'A') {
            ++latinCount;
        } else if (c >= '\u0300' && Character.UnicodeBlock.of(c) != Character.UnicodeBlock.LATIN_EXTENDED_ADDITIONAL) {
            ++nonLatinCount;
        }
    }
    if (latinCount * 2 < nonLatinCount) {
        StringBuffer textWithoutLatin = new StringBuffer();
        for(int i = 0; i < text.length(); ++i) {
            char c = text.charAt(i);
            if (c > 'z' || c < 'A') textWithoutLatin.append(c);
        }
        return textWithoutLatin.toString();
    }
    return text;
}
public static boolean containsLetters(String text) {
    for(int i=0; i<text.length(); i++) {
        if((text.charAt(i) >= 'A' && text.charAt(i) <= 'Z')
                || (text.charAt(i) >= 'a' && text.charAt(i) <= 'z') ) {
            return true;
        }
    }
    return false;
}
```

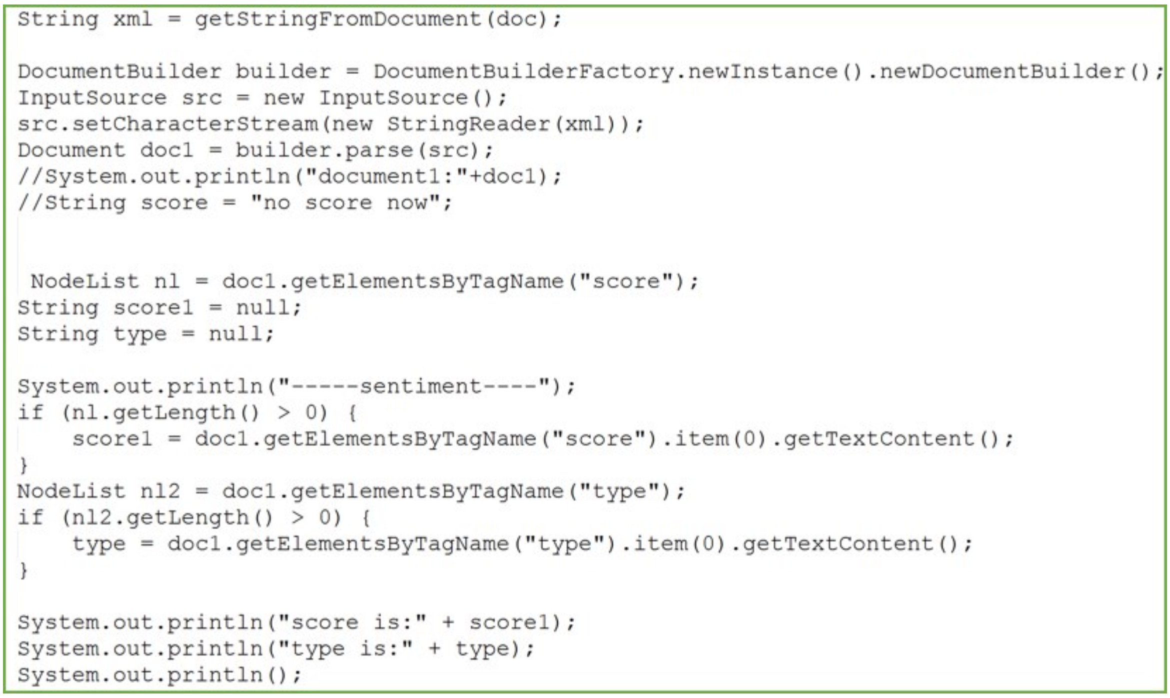
```
String text=cleaningText(rs.getString("text").trim());
if(text == null || text.length() < 3 || !containsLetters(text))
                  continue;

     ld.detectLangs(text);
     String textLang=ld.detect(text);
     String lang="en";
 if text.equals(lang)
    System.out.println(text);
```

Figure 3.6: Code snippet for retrieving english comments

Now our focus was to use these comments to identify offensive videos. We need a reliable metric that captures the intention of comments, i.e., opinion of the comment which indirectly states the nature of the video. Subsequently, we used sentiment analysis, the process of identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative or neutral[4]

[4] https://en.oxforddictionaries.com/definition/sentiment_analysis

We analyzed sentiment for each top-level comment and aggregated the number of all positive and negative comments. There are numerous API's to compute sentiment analysis but we used Alchemy API to compute sentiment of each comment and then average of sentiments was computed.



```
String xml = getStringFromDocument(doc);

DocumentBuilder builder = DocumentBuilderFactory.newInstance().newDocumentBuilder();
InputSource src = new InputSource();
src.setCharacterStream(new StringReader(xml));
Document doc1 = builder.parse(src);
//System.out.println("document1:"+doc1);
//String score = "no score now";


 NodeList nl = doc1.getElementsByTagName("score");
String score1 = null;
String type = null;

System.out.println("-----sentiment----");
if (nl.getLength() > 0) {
    score1 = doc1.getElementsByTagName("score").item(0).getTextContent();
}
NodeList nl2 = doc1.getElementsByTagName("type");
if (nl2.getLength() > 0) {
    type = doc1.getElementsByTagName("type").item(0).getTextContent();
}

System.out.println("score is:" + score1);
System.out.println("type is:" + type);
System.out.println();
```

Figure 3.7: Code snippet for extracting sentiment of comments

We also found another measure to extract meaningful information from comments. A measure which gleans information from the comment about video. We then came up with the idea of lexical features of comments.

These lexical based features involve unigrams and bigrams of text. Unigrams are single word tokens and bigrams are sequences of two words. We obtained unigrams and bigrams for comments after preprocessing them. As discussed above, we focus only on comments but not replies to the comments, and acquired unigrams and bigrams of comments for each video.

3.2.2.2 Metadata based features

We focused on features obtained from structured metadata that are about the uploaded video. We performed analysis on each individual feature to ascertain their relevance and discriminating ability.

We collected the exact upload date and time of each YouTube video in our training dataset. Each video's uploaded time and date vary significantly. As, our dataset is generated randomly, it contains arbitrary videos but not confined to a particular period of time.

We presumed that title of video is a very good indicator of offensive videos. The title usually summarizes video well. So, we decided to analyze this as a candidate feature and computed n-grams of title, i.e., unigrams and bigrams of title.

Description is another feature that provide detailed information about uploaded YouTube video. The same analysis is again performed for text in description, i.e., unigrams and bigrams of description were computed to analyze.

Channel Title is name of channel that video subscribed to. This feature captures the type of videos uploaded in that channel, which gives us some insight about the type of YouTube Video.

Duration is a temporal feature that gives the duration of video. From our manual inspection, offensive videos are short and crisp.

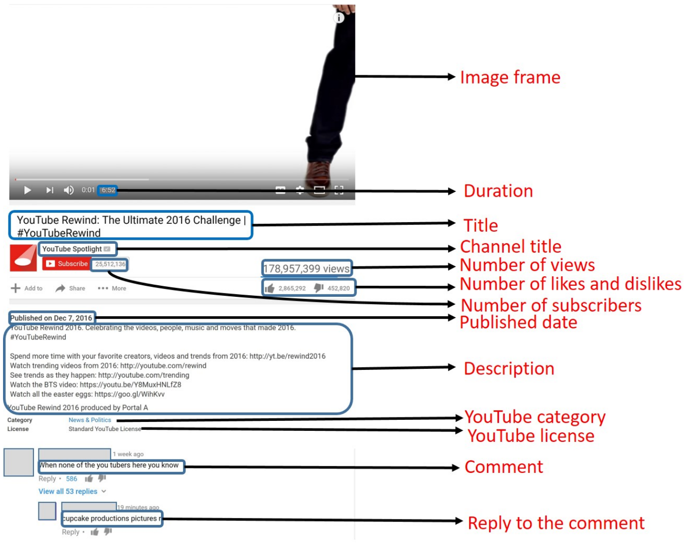
Counts of YouTube video include number of likes, number of dislikes, number of favorites and number of views. We first fetched all counts for each YouTube video in the training set.

Another feature that is associated with a YouTube video is the category of YouTube video such as sports, entertainment, music and politics.

Licensed content is another feature that is related to the uploaded YouTube video. In general, YouTube has either standard YouTube license or creative commons license.

Privacy status of YouTube video is the settings made by person who uploaded the YouTube video. A YouTube video can have any of three privacy status namely private, public and unlisted.

Figure 3.8 is an instance of YouTube video with comment and metadata features.



YouTube Rewind: The Ultimate 2016 Challenge | #YouTubeRewind

YouTube Spotlight

Published on Dec 7, 2016

ouTube Rewind 2016. Celebrating the videos, people, music and moves that made 2016.
#YouTubeRewind

Spend more time with your favorite creators, videos and trends from 2016: http://yt.be/rewind2016
Watch trending videos from 2016: http://youtube.com/rewind
See trends as they happen: http://youtube.com/trending
Watch the BTS video: https://youtu.be/Y8MuxHNLfZ8
Watch all the easter eggs: https://goo.gl/WihKvv

YouTube Rewind 2016 produced by Portal A

Category News & Politics
License Standard YouTube License

1 week ago
When none of the you tubers here you know

19 minutes ago
cupcake productions pictures r

Figure 3.8: Instance of YouTube video along with its features

Table 3.2 represents notations for each feature in the feature set.

Table 3.2: Notations for feature set

Comment Based Features	**Notation C**
Sentiment of comments	C_s
Unigrams of comments	C_u
Bigrams of comments	C_b
Metadata Based Features	**Notation M**
Unigrams of Title	M_u
Bigrams of title	M_b
Unigrams of channel title	M_{cu}
Bigrams of channel title	M_{cb}
Published time stamp	M_p
Unigrams of description	M_{du}
Bigrams of description	M_{db}
Category of video	M_c
Duration of video	M_d
Licensed of video	M_l
Privacy status of video	M_p
Number of views	M_v
Number of likes	M_{lk}
Number of dislikes	M_{dk}

3.3 Experiments and Results of Classifier

3.3.1 Machine Learning Classifiers

We experimented with our features using machine-learning algorithms such as Naïve Bayes and SVM that have been effective for text classification.

Naïve Bayes [10, 11] is one of popular text-based classifier which classifies text belonging to one category or another based on conditional probability and strong independence assumptions.

SVM [10, 11] is other supervised learning approach, i.e., a linear classifier but can deal with non-linearity by mapping input vectors to higher dimensional space with kernel function.

We present our experiments and analyze our results. We evaluate the performance of individual features for our training dataset and finally select only the features that are discriminating offensive videos. We test our test dataset with selected features. The following table gives the F-measures of the selected features with respect to the two supervised learning algorithms.

Table 3.3: F-Measures for different algorithms

	Feature set	**Naive Bayes**	**SVM**
Comment based features	C_s	0.70	0.61
	C_u	0.79	**0.83**
	C_b	0.79	0.80
	$C_s + C_u$	0.79	**0.83**
Metadata based features	M_u	0.80	0.77
	M_b	0.80	0.76
	$M_u + M_{cu}$	0.84	0.83
	$M_u + M_{cu}$	0.84	0.83
	$M_u + M_{cb}$	0.84	0.82
	$M_u + M_{cu} + M_p$	0.84	0.83
	$M_u + M_{cu} + M_{du}$	0.84	0.85
	$M_u + M_{cu} + M_{du} + M_d$	0.83	0.85
	$M_u + M_{cu} + M_{du} + M_l$	0.84	0.85
	$M_u + M_{cu} + M_{du} + M_p$	0.84	0.85
	$M_u + M_{cu} + M_{du} + M_c$	0.85	**0.86**
	$M_u + M_{cu} + M_{du} + M_c + M_{lk}$	0.83	0.83
	$M_u + M_{cu} + M_{du} + M_c + M_{dk}$	0.83	0.85
	$M_u + M_{cu} + M_{du} + M_c + M_v$	0.83	0.83
All features	$C_u + M_u + M_{cu} + M_{du} + M_c$	0.79	0.83

3.3.2 Experimental analysis of the effectiveness of the various features

Based on the experiments, we noticed that offensive videos have more negative comments compared to positive comments. We also observed that other videos such as sad songs and natural hazards also have more negative comments.

Examples:

1. "This shit is classic"

   Sentiment: negative sentiment but not offensive

2. "White people were stupid morons. And by the comment in this post, you people haven't changed"

   Sentiment: negative sentiment and offensive

3. "Sounds like the good doctor is correct, again."

   Sentiment: positive sentiment and not offensive

So, sentiment of comments broadly classifies a video as positive or negative video but it cannot distinguish negative and offensive videos.

Our experiments revealed that unigrams are more meaningful than bigrams and we combined both of these features to observe the improvement in performance. Sentiment feature doesn't have sizable impact in discriminating offensive videos compared to unigrams.

Our analysis on different metadata features suggests that text-based features are good indicators rather than numeric and temporal features. We also noticed that privacy status and license status of the content doesn't have much impact on identifying offensive videos.

From all these experiments and analysis, it was surprising to note that sentiment and content of comments were less effective in detecting offensive videos than the unigrams and bigrams in the video title and any other feature combinations does not improve the performance appreciably. We test these effective features with testing dataset.

3.3.3 Implementation details of offensive video classifier

Algorithm 1 describes our strategy to classify videos as offensive videos or not offensive videos.

In Step 1, we acquire the metadata (title, description, channel-title and category-id of each YouTube video id). In Step 2, we perform batch-filtering for training set and obtain unigrams of title, channel title and description. In Step 3, we remove stop words, numbers and special-symbols from obtained unigrams, i.e., unigrams of title, unigrams of channel title, unigrams of description and category id. The Step 4, predicts the output, i.e., class of given video id based on discriminatory features of training data set.

Algorithm 1 Identify offensive video

```
**Require:** List of YouTube video ids $[V_1, V_2 \cdots V_n]$
**Ensure:** Status of the video $\chi = \begin{cases} 1, & \text{Offensive} \\ 0, & \text{Not offensive} \end{cases}$
$\forall\ V_i \in [V_1, V_2 \cdots V_n]$
$V_t \leftarrow$ *title of* $V_i$
$V_d \leftarrow$ *description of* $V_i$
$V_{ct} \leftarrow$ *channel-title of* $V_i$
$V_c \leftarrow$ *category-id of* $V_i$                                    → feature 4
$\forall\ V_t, V_{tc}, V_d$
$V_{et}$ = *eliminate_stopwords_symbols ($V_t$)*;
$V_{etc}$ = *eliminate_stopwords_symbols ($V_{tc}$)*;
$V_{edc}$ = *eliminate_stopwords_symbols ($V_d$)*;
$\forall\ V_{et}, V_{etc}, V_{edc}$
$V_1$ = *batch_filtering ($V_{et}$)*                                    → feature 1
$V_2$ = *batch_filtering ($V_{etc}$)*                                   → feature 2
$V_3$ = *batch_filtering ($V_{edc}$)*                                   → feature 3
**if** feature 1, feature 2, feature 3, feature 4← trained classifier **then**
    $\chi = 1$
**else**
    $\chi = 0$
```

Chapter 4

Evaluation and Performance Analysis

In this phase, we evaluate our classifier performance and analyze our results with training dataset. We first discuss our training and testing datasets, followed by evaluation metric, followed by evaluation results.

4.1 Training and Testing Datasets

We extracted our dataset from YouTube using YouTube Data API. We first manually acquired video ids of videos that serve our purpose using keywords. Then, we input these video ids to YouTube crawler to extract metadata of videos using YouTube API. We were able to acquire 96 video ids that were offensive. In addition to these video ids, we also gathered 204 random video ids metadata from

YouTube. Altogether, these 300 video ids were annotated by four domain experts and labelled as offensive video to community or not. Inter-judge agreement of these annotations was 0.77.

This task of creating training and test dataset was labor intensive because obtaining balance dataset containing offensive videos for the purpose of classifier development was difficult.

We then collected our testing dataset similar to our training dataset. We collected offensive video ids and non-offensive video ids for testing our classifier. On the whole, testing dataset contained 86 video ids, i.e., 27 offensive and 59 non-offensive video ids. Next, we crawled metadata for 86 video ids to test them with our trained classifier. Experimental dataset statistics of training and testing datasets are summarized below.

Table 4.1: Training and Testing datasets

Training Dataset	**Testing Dataset**
96 (Offensive) + 204 (Not offensive)	27 (Offensive) + 59 (Not Offensive)

4.2 Evaluation Metric

We now evaluate our approach using F-measure. We also used confusion matrix, i.e., a matrix where rows represent actual classes and columns represent predicted classes to see the classifier effectiveness.

Precision [10, 11] is defined as the ratio of number of correctly classified videos (diagonal sum of entities) to the total predicted videos.

Recall [10, 11] is defined as the ratio of correctly classified instances (sum of off diagonal) to the number of videos in that class.

Accuracy [10, 11] is defined as the ratio of number of correctly classified instances to the actual number of videos.

In addition to these measures, we also calculate F-measure [10, 11] which is a combined measure of precision and recall. A sample confusion matrix is shown below in Table 4.2

Table 4.2: Sample confusion matrix

		Predicted	
		Offsensive	**Not offensive**
Actual	**Offensive**	p	q
	Not offensive	r	s

In above figure, ‘p’ represents number of videos that are correctly classified as offensive (true positives), ‘q’ represents number of videos that are incorrectly classified as non-offensive (false negatives), ‘r’ represents the number of non-offensive videos that are classified as offensive videos (false positives), ‘s’ represents the number of non-offensive videos that are correctly classified as non-offensive (true negatives). We can now define true positive rate, true negative rate, false positive rate and false negative rate as follows:

Table 4.3: Evaluation Metrics

Evaluation Metric	**Formula**
True Positive Rate	$\frac{p}{p+q}$
False Positive Rate	$\frac{q}{p+q}$
False Negative Rate	$\frac{r}{r+s}$
True Negative Rate	$\frac{s}{r+s}$
Precision	$\frac{p}{p+r}$
Recall	$\frac{p}{p+q}$
Accuracy	$\frac{p+s}{p+q+r+s}$
F-Measure	$2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

4.3 Classification Results

We also conducted experiments on our test dataset to observe the performance of various classification algorithms and found that SVM algorithm is best suited for our classification. The following table represents the precision, recall and F-measure for various machine-

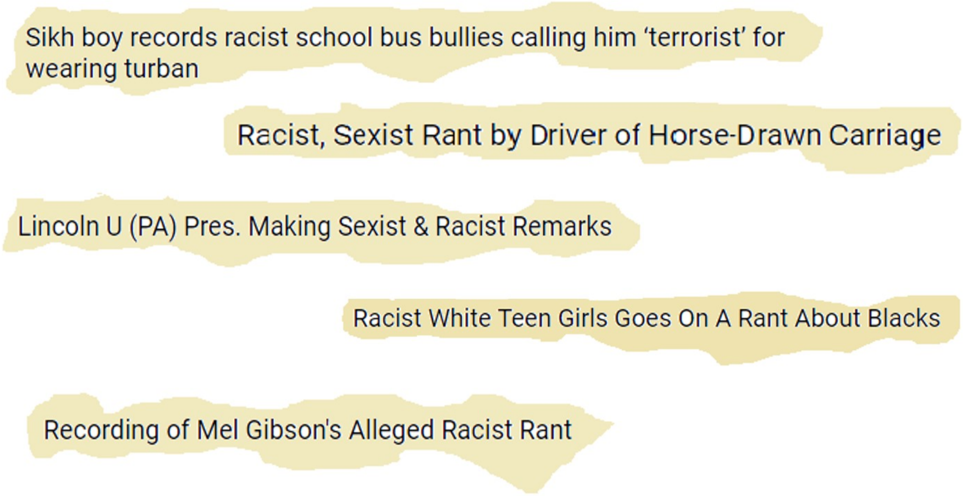
Table 4.4: Results of classifiers

	Precision	**Recall**	**F-Measure**
Naive Bayes	87.5	77.7	82.3
SVM	88	81.4	84.5

learning algorithms.

Table 4.5 represents the confusion matrix for the test dataset. From results of the classifier, we observed that our classifier classified 25 video ids as offensive and 61 video ids as non-offensive based on our discriminatory features.

Figure 4.1 are some of the YouTube videos that were identified as Offensive videos.



Sikh boy records racist school bus bullies calling him ‘terrorist’ for wearing turban

Racist, Sexist Rant by Driver of Horse-Drawn Carriage

Lincoln U (PA) Pres. Making Sexist & Racist Remarks

Racist White Teen Girls Goes On A Rant About Blacks

Recording of Mel Gibson's Alleged Racist Rant

Figure 4.1: YouTube video title instances

We found that 5 and 3 video ids are misclassified as offensive and non-offensive videos respectively.

Table 4.5: Confusion Matrix

	Offensive	**Not offensive**
Offensive	22	5
Not offensive	3	56

The evaluation metrics are computed as follows:

1. Precision $= 0.88$

2. Recall $= 0.81$

3. F-Measure$= 2 \times \frac{0.88 \times 0.81}{0.88 + 0.81} = 0.84$

Chapter 5

Conclusion and Future Work

In this thesis, we designed a framework to identify videos that offend a large group of people in a community. We collected 96 offensive videos based on keywords and later coupled them with 204 random videos (for instance, a video titled Trump vs Hillary) to enhance our dataset. But creating a balanced dataset was challenging. We experimented with various combinations of features such as comment based features and meta data based features and noticed that title, a meta data based feature, is more effective compared to comment based features. So this contributes to efficiency. We experimented using different machine learning algorithms and found that SVM, gave us best performance. This performance of classifier indicates that meta data features are effective in detecting offensive videos. Further, we also found that sentiment and content of the comments were less effective in detecting offensive videos than the unigrams and bigrams in the video title. Thus these simple features contribute to the efficiency of computation and this implies that uploaders provide good titles.

One of the limitations of our approach is the limited dataset. We can extend this dataset using lexicons of different categories of offensiveness such as sexual-orientation, religion, political-beliefs and appearance. The other limitation is the restriction of language. We restricted our framework to only English videos. We can extend this framework to support multilingual content metadata. Also, computation of other features for extracting audio transcript associated with the video as text and further analyzing it, topic-modeling of the comments may improve performance of the classifier.

Bibliography

[1] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, "Detecting offensive language in social media to protect adolescent online safety," in *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, Sept 2012, pp. 71–80.

[2] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 18:1–18:30, Sep. 2012.

[3] M. Dadvar, D. Trieschnigg, R. Ordelman, and F. de Jong, *Improving Cyberbullying Detection with User Context.* Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 693–696.

[4] K. Reynolds, A. Kontostathis, and L. Edwards, "Using machine learning to detect cyberbullying," in *2011 10th International Conference on Machine Learning and Applications and Workshops*, vol. 2, Dec 2011, pp. 241–244.

[5] N. Aggarwal, S. Agrawal, and A. Sureka, "Mining youtube metadata for detecting privacy invading harassment and misdemeanor videos," in *2014 Twelfth Annual International Conference on Privacy, Security and Trust*, July 2014, pp. 84–93.

[6] R. I. Rafiq, H. Hosseinmardi, R. Han, Q. Lv, S. Mishra, and S. A. Mattson, "Careful what you share in six seconds: Detecting cyberbullying instances in vine," in *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, Aug 2015, pp. 617–622.

[7] R. M. Kowalski, S. P. Limber, S. Limber, and P. W. Agatston, *Cyberbullying: Bullying in the digital age.* John Wiley & Sons, 2012.

[8] J. W. Patchin and S. Hinduja, *Cyberbullying prevention and response: Expert perspectives.* Routledge, 2012.

[9] T. Fu, C. N. Huang, and H. Chen, "Identification of extremist videos in online video sharing sites," in *2009 IEEE International Conference on Intelligence and Security Informatics*, June 2009, pp. 179–181.

[10] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval.* New York, NY, USA: Cambridge University Press, 2008.

[11] R. Baeza-Yates, B. Ribeiro-Neto *et al.*, *Modern information retrieval.* ACM press New York, 1999, vol. 463.