Contextual Feature Based One-Class Classifier Approach for Detecting Video Response Spam on YouTube

Student Name: Vidushi Chaudhary

IIIT-D-MTech-CS-IS-13-MT11016
March 07, 2013

Indraprastha Institute of Information Technology
New Delhi

Thesis Committee
Ashish Sureka (Chair)
Pushpendra Singh
Naveen Kumar

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

Keywords: Spam detection, YouTube, one-class classification, social-media analytic, video response spam detection, classifier feature evaluation and selection

Certificate

This is to certify that the thesis titled **"Contextual Feature Based One-Class Classifier Approach for Detecting Video Response Spam on YouTube"** submitted by **Vidushi Chaudhary** 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 largest video sharing websites (with social networking features) on the Internet. The immense popularity of YouTube, anonymity and low publication barrier has resulted in several forms of misuse and video pollution such as uploading of malicious, copyright violated and spam video or content. YouTube has a popular feature (commonly used) called as video response which allows users to post a video response to an uploaded or existing video. Some of the popular videos on YouTube receive thousands of video responses. We have observed the presence of opportunistic users posting unrelated, promotional, pornographic videos (spam videos posted manually or using automated scripts) as video responses to existing videos.

We present a method of mining YouTube to automatically detect video response spam. We formulate the problem of video response spam detection as a one-class classification problem (a recognition task) and divide the problem into three sub-problems: promotional video recognition, pornographic or dirty video recognition and automated script or botnet uploader recognition. We create a sample dataset of target class videos for each of the three sub-problems and identify contextual features (meta-data based or non-content based features) characterizing the target class. Our empirical analysis reveals that certain linguistic features (presence of certain terms in the title or description), temporal features, popularity based features, time based features can be used to predict the video type. We identify features with discriminatory powers and use it within a one-class classification framework to recognize video response spam. We conduct a series of experiments to validate the proposed approach and present evidences to demonstrate the effectiveness of the proposed solution with more than 80% accuracy.

Acknowledgments

Life as a researcher is a zig-zag way and has many ups and downs but i overcome from them because of the support and faith of many individuals.
First and foremost I would like to express my sincere gratitude to Dr. Ashish Sureka for providing me an opportunity to work with him. Your creativity encouraged me, your energy and dedication towards work motivates me. Your perfection always inspired me to do things perfectly. I feel motivated and encouraged everytime i attend your meeting. Without your encouragement, inspiration and guidance this project would not have materialized. Thank you so much for not only guiding me in this project but for every knowledge you give me. This thesis is dedicated to your creativity and enthusiasm of doing work.
I would like to thank Akshay Rajwanshi (B.Tech 3rd year student at IIIT- Delhi) for his tremendous help in initial phase of this thesis project when i did not know what would be the final outcome of this project. His help in manual analysis of the experimental dataset was the great support.
I wish to thank many people for their direct or indirect help in achieving this goal. Special thanks to my parents for their support and encouragement throughout my study, their faith in me ignites that spark to do anything for their happiness, my sister and jiju for being the best sis and jijs one can have.
Last but not least I would like to thank all my friends for being with me at each step when i need their support. This thesis would never be successful without your support and love.

Contents

List of Figures

List of Tables

Chapter 1

Research Motivation And Aim

The web 2.0 is now building up enormously which consist of search engines, social networking sites, video sharing sites, and photo sharing sites. Specially social networking sites such as Facebook [1] , Twitter, Flickr have increased a lot since the last decade which specializes in micro- blogging, video sharing, photo sharing and discussion forms. In particular, video is becoming a most important part of user's daily life, the reason being video is the most usable medium to share views with others and is a medium of many type of interactions among users such as political debates, educational tips etc. Out of many video sharing sites present on the Internet, YouTube is the largest and most popular video sharing site [2] .

1.0.1 YouTube

YouTube is one of the most popular video sharing websites (with social networking features) on the Internet. Figure 1.1 shows the popularity of YouTube with video contextual features. On YouTube, users can upload a video, view and share videos, subscribe to a particular channel, like or dislike any video, post textual comment on a video, post a video as a response to a particular video. Statistics shows the enormous popularity of YouTube[3] that over 800 million unique users visits YouTube each month and over 100 million users actively participate either by liking, disliking a video or by posting comments; about 72 hours of videos are uploaded to YouTube every minute. YouTube is the most popular medium to share videos on most popular social networking sites i.e. facebook and twitter. 3 hours of videos are uploaded per minute on YouTube from mobile devices shows the immense fame of YouTube on mobile devices. These statistics shows the huge popularity of YouTube on web and mobile devices.
Being such a popular video sharing site, it become a platform for spammers and promoters to post unrelated and irrelevant content [4] [1] either as video response or as related video to the most popular videos either to gain popularity or to promote their sites or products. The presence of spam then become a serious problem as there is huge amount of data that streams on YouTube every minute and presence of spam in such case cause bandwidth waste, time waste, degraded user experience etc which is undesirable.
In general, spam is some irrelevant, unsolicited message posted over the web, specially to large number of users with the intention of either getting publicity or to spread viruses, malwares. In our context, spam is some unrelated, unsolicited video posted as video response to a YouTube video.

[1] http://www.statista.com/topics/751/facebook/
[2] http://www.comscoredatamine.com/2010/09/youtube-viewing-hours-across-markets/
[3] http://www.youtube.com/yt/press/statistics.html

Figure 1.1: YouTube

Research shows Web 2.0 platforms and social media websites (such as online discussion forums, blogs, social networking websites, video sharing platforms, micro-blogging websites) are easy target for spammers and users with malicious intent [6] (because of low barriers to post content and anonymity). Previous studies shows that spam (video pollution, video response to uploaded videos , forum comments) is prevalent on YouTube and YouTube has taken several measures to counter the spam problem [1] [2] [5]. Figure 1.1 shows the video response feature of the YouTube video.

1.1 Research Motivation

While doing manual inspection of YouTube video responses, we found some spam videos (pornographic, promotional, and botnet) are posted as video response to non- spam videos. Presence of pornographic, commercial videos as video response to a non- pornographic, non- commercial videos respectively shows the spam behaviour. Figure1.2 shows that a pornographic video is posted as video response to a child game video. If a child will watch a pornographic video, that would have negative impact on child's mindset, which is not desirable and shows spam behaviour. Figure1.3 shows that a pornographic video is posted as video response to one of the most popular educational dance video, presence of porn video as reponse to an educational video shows the spam behavior. Figure1.4 is an image of all time most viewed music video, and a pornographic video is posted as video response to a most viewed video with the intention to promote pornography which is spam. By manual analysis, we observed that some commercial, botnet videos are posted as video response to most popular and viewed video with the intention to promote their site or product. Figure1.5 shows the presence of unrelated commercial videos as video response to most popular video waste bandwidth and time (at user part) and indicates spam behavior.

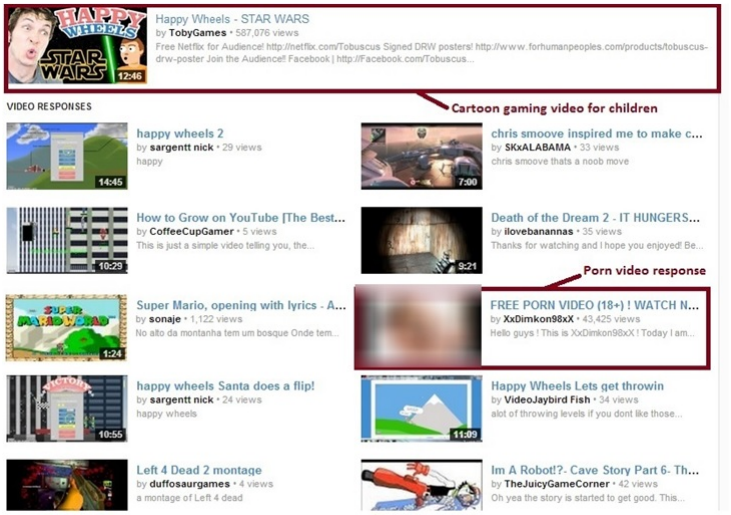

Figure 1.2: Screenshot of a pornographic video posted as a video response to child game video for children.

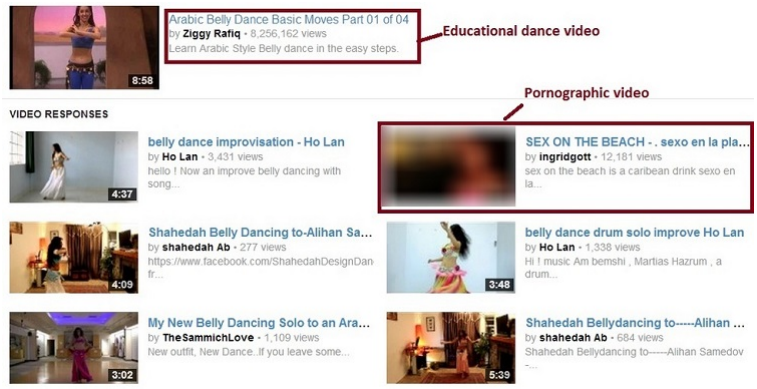

Figure 1.3: Screenshot of a pornographic video posted as a video response to an educational dance video.

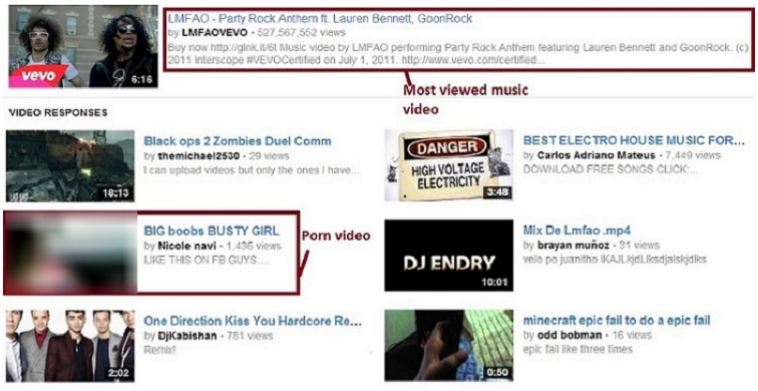

Figure 1.4: Screenshot of a pornographic video posted as a video response to a most viewed music video.

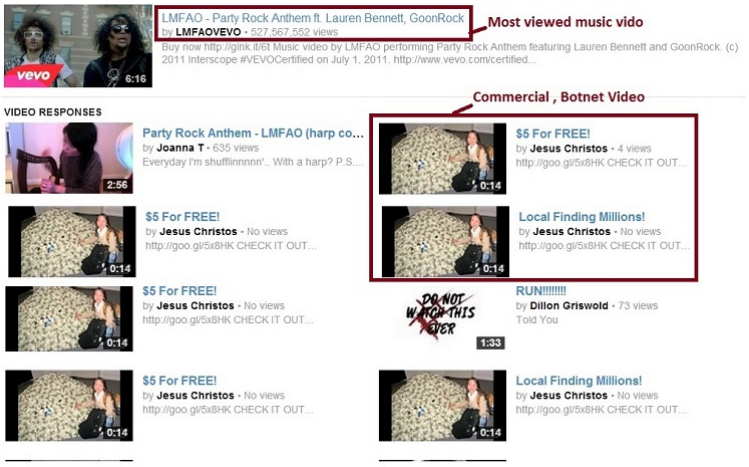

Figure 1.5: Screenshot of a commercial, botnet video posted as a video response to a most viewed music video.

These examples confirm the presence of spam videos as video response to legitimate videos. Presence of spam is completely undesirable and has many disadvantages so it require researchers attention to get solved. The work presented in this report is motivated by the facts that:

- Large amount of data streams on YouTube every minute, presence of spam in such case cause bandwidth waste.
- Posting spam on YouTube has several disadvantages like undesirable consumption of resources, decreased user experience, degraded quality of YouTube, lowered system reputation [1] [5] [6].
- Some pornographic videos are video response of a kids rhyme videos, cartoon video, gaming videos for children. Presence of pornographic video as video response to such videos is not legitimate and will have negative impact on kid's mental growth. Figure 1.2 shows a pornographic video posted as a video response to a cartoon gaming video and most popular music video.
- Botnet videos cannot be a video response of any video because only a human being can analyse the content of the video by watching it and can post a video as response to the main video which he think is related to the main video. An automatic script cannot analyse the content of a video [5].
- Presence of commercial videos as video response to most viewed and most discussed videos shows spam behaviour because the intention behind posting commercial video as video response to most viewed and most discussed video is to promote their sites or products and gain popularity. Promoters choose most viewed and most discussed videos as medium to promote their sites because if large number of people are watching the main video, higher the chances that their video will also get large number of views and they will gain popularity. These examples show the importance of the problem.

The focus of the work presented in this report is video response spam (responding to an uploaded video on YouTube using another YouTube video) detection. Previous research shows that video response spam on YouTube is prevalent and the problem of video response spam detection has attracted researcher's attention [1][2][3][4].

1.2 Research Aim

Previous section shows that presence of spam on the most popular video sharing site i.e. YouTube and has several disadvantages. It requires researcher's attention to solve the video response spam problem on YouTube. The research aim of the work presented in this report is the following:

- **Broad Objective:** To increase our understanding of video response spam problem on YouTube and investigate effective solutions to combat the video response spam problem (by mining video contextual data and identifying discriminatory features which can be used within a classification framework).
- **Specific Objective:** To examine the application of a one-class classification framework for recognizing pornographic video response spam, commercial video response spam and botnet video response spam based on several linguistic, temporal, trust and popularity-based features. To conduct a characterization study and empirical analysis on a real-world dataset to measure the effectiveness of the proposed hypothesis.

1.2.1 Advantages

The work presented in this report has several advantages:

- Better user experience.
- No compromise on system reputation.
- Less bandwidth wastage.
- No time wastage.

Chapter 2

Related Work And Research Contributions

2.1 Related Work

The work presented in this report belongs to the area of Spam detection on YouTube. In this section, we discuss some closely related work (to the experiment presented in this report) and present novel research contributions in context to existing work. We categorize the related work in 3 lines of research: Video Response Interactions and Video response spam, Social media spam detection, Classification of YouTube videos based on contextual features.

2.1.1 Video Response Interactions and Video Response Spam

Fabricio et al. analyzed the properties of the social network created by video response interactions on YouTube [3]. They characterize users interaction with each other on YouTube to understand how malicious users can behave. The main aim of their work is to find evidence of pollution (opportunistic behaviour of spammers and promoters). They also did some study on user behavioral patterns in video based environment [3].
Fabricio et al. present a binary classification strategy to detect spammers on YouTube. They contrive a number of YouTube users and their profile, social behaviour and finally propose a video spammer detection mechanism that classifies a user either as a spammer or a legitimate user [4]. Their results highlight the most important attributes for video response spam detection [4].
Fabricio et al. address the issue of detecting Spammers and Content Promoters and classify the real YouTube users as Spammers, Promoters or Legitimate users based on user behaviour attributes. They present experimental results which demonstrate that characterization of social and content attributes is helpful to distinguish each user class [1].

2.1.2 Social media spam detection

Sureka present a technique to automatically detect comment spammers in YouTube Forums by mining comment activity log of a user and extracting patterns which indicates the spam behaviour. Their empirical analysis on sample dataset demonstrate the effectiveness of proposed technique in identifying comment spammers [5].
Paul et al. survey potential solution for fighting spam detection on social wesites like Wikipedia,

Table 2.1: Literature survey of papers (chronological order) on YouTube video response spam detection using contextual features based one class classifier approach. VIVRS = YouTube video interaction and video response spam, SMS = Social media spam, CCF = classification of YouTube videos using contextual features.

Study	**Type**	**Purpose/ Goal**
Yiming et al., 1997 [7]	CCF	Presented a comparative study on Feature Selection in Text categorization
Paul et al., 2007 [6]	SMS	Survey potential solution for fighting spam on social web sites. Compare the results with prior solutions to email and web spam.
Yuan et al., 2007 [8]	CCF+SMS	Propose Contextual based analysis to automatically detect Forum spamming Define three perspective of Forum spamming and propose context- based detection technique to detect forum Spam
Fabrico et al. , 2008 [3]	VIVRS	Analyzed the properties of the social network created by video response interactions on YouTube They characterize users interaction with each other on YouTube to understand how malicious users can behave The main AIM of their work is to find evidence of pollution (opportunistic behaviour of spammers and promoters) They also did some study on user behavioral patterns in Video based environment
Fabrico et al. , 2008 [4]	VIVRS	Contrive a number of YouTube users and their social behavior to discriminate a spammer from a legitimate user. Their results highlight the most important attributes for video response spam detection
Yinglian et al., 2008 [11]	SMS	Developed a spam signature generation framework for botnet spam emails detection.
Fabrico et al. , 2009 [1]	VIVRS	Instead of classifying the content of the YouTube video, they are addressing the problem of detecting spammers and content promoters on YouTube.
Benjamin et al.,2009 [12]	SMS	Study of automatic detection of spammers in a social system. Analyze distinct features that address various properties of social spam.
Fabrico et al., 2010 [2]	VIVRS	Define existing pollution in video sharing systems, their negative impact to users and systems and possible solution to minimize the problem.
Ashish Sureka, 2011 [5]	SMS	Presented a method to automatically detect comment spammers on YouTube. Technique was based on mining comments feed and extracting patterns indicating spam behavior.

Flickr and finally presented a comparative study of their work with previous e-mail and web spamming. Their paper surveys three categories of potential countermeasures which have been proposed before email and web spamming and in this paper, the author find that their applicability to social websites differs [6].

2.1.3 Classification of YouTube videos based on contextual features

Yiming et al. present a comparative study on feature selection methods in reduction to a high dimensional feature space in text categorization problems. Their work is motivated by the fact that as more and more information is available online, effective retrieval is difficult without good indexing. They compare 5 methods of feature selection and find the effectiveness of these feature selection methods in text categorization [7].
Yuan et al. propose context-based analysis (redirection and cloaking analysis) to detect spam automatically and to overcome shortcomings of content-based analysis. They have conducted a comprehensive study of forum spamming from three perspectives: the search user, the spammer, and the forum hosting site and showed that redirection analysis and cloaking are very effective in identifying forum spammers [8].

2.2 Research Contributions

In context to closely related work, this report makes the following novel contributions:

1. The work presented in this report is the first step in the direction of applying a one-class classifier based approach using contextual features to detect video response spam on YouTube. While there has been work done in the area of detecting video response spammers and promoters on YouTube, the application of three one-class classifiers (pornographic video recognition, commercial video recognition and botnet uploader detection) based on 18 video contextual features (refer to Table) offers a fresh perspective and a novel research contributions of this work.

2. We conduct empirical analysis on real world dataset acquired from YouTube to train and test the effectiveness of the proposed features and classifier. We present the intuition behind each discriminatory feature and an empirical analysis demonstrating its influence or impact on the classification task.

Chapter 3

Proposed Solution Approach

Figure 3.1 presents the research method adopted in our study and proposed solution approach. We divide the spam video response detection problem into three sub- problems: pornographic video response detection (PVRD), botnet video response detection (BVRD), and promotional or commercial video response detection (CVRD). PVRD, BVRD and CVRD are framed as three independent one-class classification problems. We employ one-class classification approach due to the nature of the problem (a recognition task) in which resemblance (similarity between objects in the training dataset and the test object) is calculated to recognize if the test video is spam or not. As shown in Figure 3.1, the proposed solution approach is a four step process: positive class training dataset, characterization and training, verification or recognition, performance evaluation.

3.0.1 Positive Class Training Dataset

The first step consists of acquiring positive class training dataset (using YouTube APIs[1]) for the purpose of training a classifier. We download all the available meta-data of several popular YouTube videos and their video responses. We extract meta-data serving as basis for various types of features such as: linguistic features (title and description of the video), temporal and popularity based features (number of subscribers, likes, views and forum comments posted in response to the video) and times based features (duration, time-stamp of upload).

3.0.2 Characterization and Training

Characterization is the process by which features can reveal the behaviour of the object (YouTube video). The next step consists of characterization and identification of discriminatory features. We conduct an in-depth manual analysis and visual inspection of the meta-data of YouTube video responses to identify patterns which can be used as markers for the classification task. This step consists of characterizing the target class using various types of features.

3.0.3 Verification or Recognition

We propose a weighted similarity function (based on the type of the features and distribution of the variables representing the features) to compute the resemblance between the target class object and the objects in the training dataset.

[1] http://code.google.com/apis/youtube/overview.html

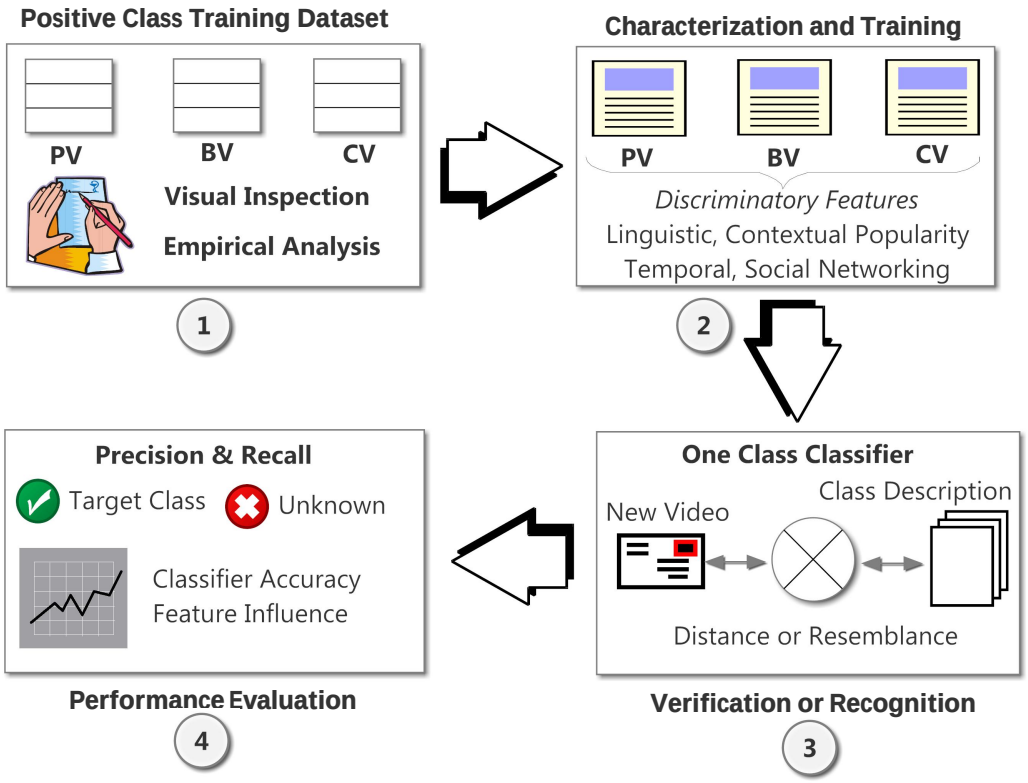

Figure 3.1: Research Framework

3.0.4 Performance Evaluation

The last step in the process is the performance evaluation of the three independent classifiers using standard information retrieval metrics such as precision and recall. This step evaluates the effectiveness of applying one- class classification approch to recognize spam video responses.

3.1 Solution approach to Detect Pornographic Video Response

A pornographic video is not considered as a spam video until it is posted as a video response to a non- pornographic video. Presence of pornographic video as video response to a kids rhyme video, political news video, and music video considered as spam. The aim of this section is to detect pornographic video responses which are posted as video response to a non- pornographic video. To detect pornographic video response, the first step is the characterization which involves fetching of meta data of pornographic video responses and find discriminatory features which can be helpful to recognize pornographic videos. We have divided our discriminatory feature set into 5 categories: linguistic features like percentage of pornographic terms in title and description, temporal or popularity based features like number of subscribers, likes and views, time based features like duration of the video, YouTube basic features like category of the YouTube video,

and trust features like safe search and web of trust.

3.1.1 Keyword based search technique

Keyword based search technique is applied on linguistic features to detect pornographic videos. Figure 3.2 shows our approach of keyword based search technique to detect pornographic video responses.

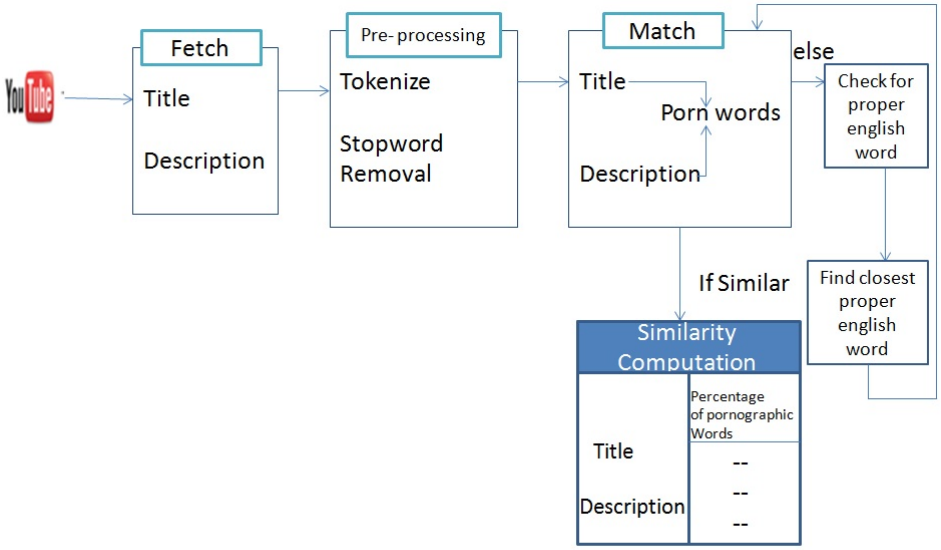

Figure 3.2: Keyword based search technique to detect pornographic video responses

Fetch

Fetch block in figure 3.2 shows the retrieval of linguistic features like title and description of the YouTube video. We fetch title and description of pornographic videos through YouTube api's and apply keyword based search technique to detect presence of pornographic terms in title and description as presence of pornographic terms in title and description indicates the pornographic behaviour of the video. Our hypothesis is based on the assumption that only pornographic videos contain pornographic terms in their title and description and a non- pornographic video will not contain pornographic terms.

Preprocess

Next step is to preprocess the fetched title and description. Preprocessing involve tokenization and stop word removal. Tokenization is the process of breaking a stream of text into words called tokens. After tokenization, stop words are removed from the token list. Standard english stop word list present over the web is used to remove stop words from title and description [2].

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

Similarity Computation

In the final step, we do the matching of the tokens present in title and description with the porn words list (standard porn word dictionary present over the web) and compute the percentage of pornographic terms present in title and description. Higher the percentage of pornographic terms present in title and description, higher the chances that the video is a pornographic video.

3.1.2 Solution Approach to Detect Botnet Video Response

Botnet video is the video posted by an automatic script and not by a human being. Botnet video response is considered as spam because an automatic script can not analyse the content of a video and post response which is related to the main video. Figure 3.3 is a screenshot of a botnet profile. We notice that as botnet videos are posted by an automatic script, time

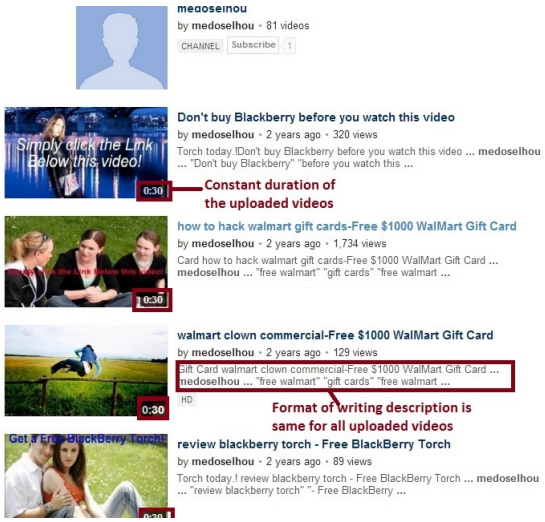

Figure 3.3: Screenshot of a botnet profile

difference between uploaded videos is very less say few seconds. It is infeasible for a human being to upload multiple videos in very few seconds. So we fetched uploaded time of all the videos uploaded by the uploader, sort by time and compute the time difference between each successive video. if the time difference is less than a threshold, it indicates the botnet behaviour of the video. Figure 3.4 shows our approach applied on uploaded time to detect botnet videos.

3.1.3 Solution Approach to Detect Commercial Video Response

Approach to detect commercial/promotional video response is same as pornographic video response detection. The difference is that feature set to detect commercial video is different from the pornographic video.

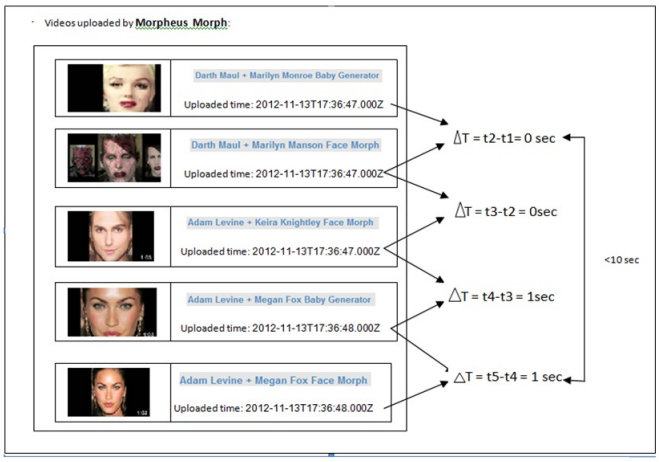

Figure 3.4: Approach applied to uploaded time feature to detect botnet video

3.2 Classifier

In one class classification problem, either negative class is not present or it is not properly sampled. The goal of one-class classifier is to recognize spam video response. In one class classification approach, each object is represented by a vector of values, say feature vector. The algorithm does the similarity computation of new video with the existing labeled (spam) dataset to recognize video as spam i.e. pornographic, promotional, or botnet.

In this section, we describe the classifier we have developed to detect spam video responses on YouTube. The first step of the classification algorithm is to define the feature vector(set of features) which defines the feature space.

Classifier feature vector X is:

$\mathbf{X= (x_1, x_2, x_3 \ldots ., x_n)}$,where n = Number of features from the feature space.

n = 8 in case of Pornographic Video Response Detection,
n = 6 in case of Commercial Video Response Detection,
n = 4 in case of Botnet Video Response Detection.

The training Dataset (TD) is the set of observation vectors along with corresponding class labels. The training dataset contains data only for spam videos so class label is same for all videos presented in training dataset.

$\mathbf{TD = ((x_1, x_2, x_4 \ldots ., x_n), y_j)}$, where j = Size of Training Dataset, y_j = Class Label.

n = 8, j = 250 in case of PVRD,
n = 6, j = 200 in case of CVRD,
n = 4, j = 61 in case of BVRD.

The testing Dataset (TS) is the vector of feature value without class label.

$\mathbf{TS = ((x_1, x_2, x_3 \ldots ., x_n)_k)}$, where k = Size of Testing Dataset.

n = 8, k = 1,000 in case of PVRD,
n = 6, k = 1,000 in case of CVRD,
n = 4, k = 3,389 in case of BVRD.

Each sub-problem has multiple features; weight to each feature is assigned which shows the contribution

of the corresponding feature in recognition of spam video. Let **W_i = Weight of the feature i s.t.**

$$\boxed{\sum_{i=1}^{n} W_i = 1} \tag{3.1}$$

Input: A list L of features.
Result: Weight of each feature.
initialization;
Assign equal weight to each feature s.t

$$\sum_{i=0}^{n} W_i = 1 \tag{3.2}$$

Run the classifier and calculate the accuracy of the system, say accuracy1.
for *each feature f in L* **do**
    Remove feature j from L;
    Adjust weights of rest of the features s.t.

$$\sum_{\forall i \neq j} W_i = 1 \tag{3.3}$$

    Run the classifier and check the accuracy of the system, let accuracy2;
    Let

$$\Delta = \textbf{percentage change in accuracy / 100} \tag{3.4}$$

    if *((Significant change in accuracy))* **then**
        Removed feature is an important feature and weight corresponding to this feature should be high;

$$feature_{weight_i} = feature_{weight_i} - \Delta \tag{3.5}$$

    else
        Removed feature is not an important feature and weight corresponding to this feature should be low;

$$feature_{weight_i} = feature_{weight_i} + \Delta \tag{3.6}$$

    end
end

Algorithm 1: Algorithm for Weight computation of each feature.

Algorithm 1 shows our approach of calculating the weight of each feature where the whole process is repeated until the accuracy is optimal. The result of the algorithm shows the contribution of each feature in the spam video response detection. Lower the weight, more important the feature is. Figure 3.5 shows the flow chart of computing weight of each individual feature based on their influence.

One class classification approach is based on similarity computation; we need to find the similarity of the new object with the existing dataset which is the score of that particular feature. Score of the feature is a unique value which represents that feature in comparison to the training dataset.

S_i = Score of the feature i s.t.

$$0 \leq Si \leq 1 \tag{3.7}$$

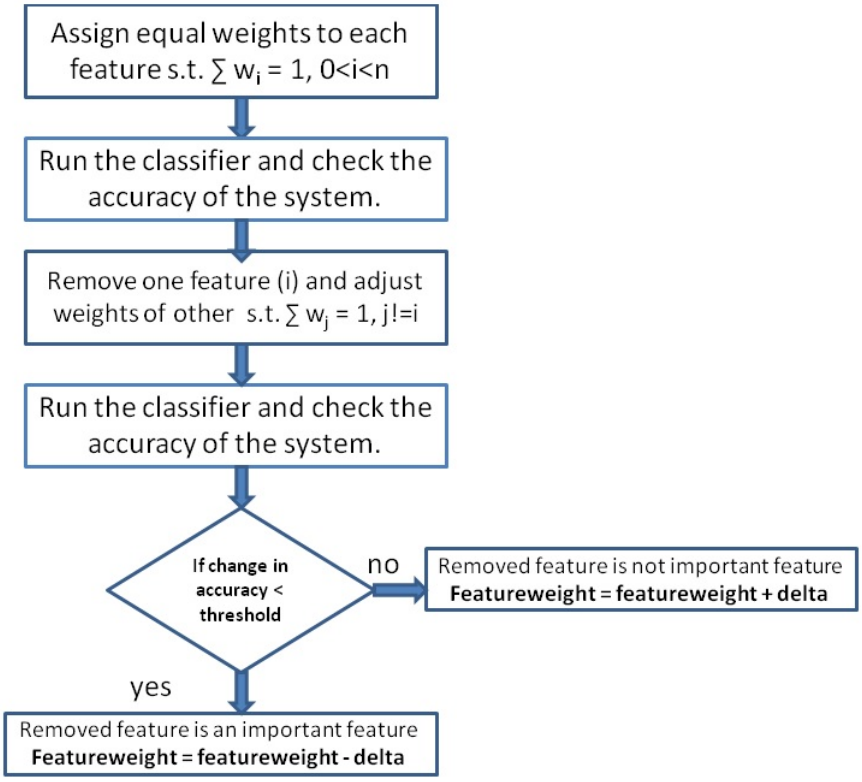

Figure 3.5: Flow chart of Weight Computation

Our experimental dataset consists of both numerical features (nf) like duration of the video, number of subscribers etc and categorical features (cf) like category of the YouTube video; there are different approaches to calculate score of these features. For numerical features, we are calculating the average difference of the new object with the existing training dataset. Lower the difference, higher the chance that new object is similar to the existing dataset. For categorical features, percentage of videos fall into the specific category contributes in finding the score of the feature.
If j is the size of training dataset, then equation to score of numerical feature is:

$$\boxed{Score_{nf} = \sum_{i=0}^{j}(|(new_{value} - TS[i])|/n)} \tag{3.8}$$

This equation of calculating score of numerical feature is not applicable to all the numerical features because for certain features like percentage of pornographic terms in title and description, higher the number of dirty terms present, higher the chances that video is a pornographic video. For such features, let x = Percentage of pornographic or commercial terms present in title or description.

$$\boxed{Score_{nf} = 1 - (x/100)} \tag{3.9}$$

Let y = Percentage of videos fall in the particular category

$$\boxed{Score_{cf} = 1 - (y/100)} \tag{3.10}$$

Because we consider the average difference, lower the value of the score, higher the chance that new object is similar to the training dataset objects. Based on weight and score of each feature, we compute the final value of the feature, C_{value} which is the resemblance of the feature with the target class and recognizes the Spam behaviour of the video.

$$\boxed{c_{value} = \sum_{i=0}^{n} W_i * S_i} \tag{3.11}$$

Chapter 4

Empirical Analysis and Performance Evaluation

The aim of this section is to present the empirical analysis done for characterization of the contextual features for each sub- problem i.e. pornographic video response detection, botnet video response detection, and commercial video response detection. At the end of this section, accuracy matrix is shown which shows the effectiveness of the proposed solution approach.

4.1 Experimental Dataset

We acquire experimental data using YouTube API. We download all the meta-data (available using YouTube API) of 50 most popular and 50 most discussed videos (from YouTube Charts [1] [all categories] during the month of November 2012) and their video responses. We selected most popular and most discussed videos as several video responses are posted to such videos and because of their popularity become targets for spammers. YouTube has an API limit which allows a maximum of 1000 videos responses to be fetched for a given video. We were able to fetch a total of 5900 videos. We manually analyzed each video and tagged videos which were uploaded by a botnet (classified based on our perception) and videos which seemed to be clear cases of pornographic videos and commercial promotion. We classified 532, 427 and 100 videos as pornographic, commercial promotion and botnet respectively. We notice that several commercial promotion videos are uploaded by botnet. Our findings shows that 9.01% of the video responses were pornographic and 7.23% are commercial promotion which is an evidence of the extent of video response spam on YouTube. We divide the botnet, pornographic and commercial promotion into training and testing dataset for one-class classifier model building and evaluation. The size of the training and test dataset for the three class of videos are: 250 training and 250 test dataset for pornographic videos, 200 training and 200 test dataset for commercial promotion, 61 training and 39 test dataset for botnet videos. Algorithm 2 shows our apporach of collecting experimental dataset. A video is said to be a responsive video if it is posted as a video response to any other video and a video is said to be a responded video if it has atleast one video response [1]. Algorithm 2 applies the concept of responsive video and responded video to collect experimental dataset of spam (pornographic, botnet, and commercial) video responses.

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

Input: A list L of most popular and discussed videos
Result: Experimental dataset of video response spam
initialization;
for *each video V in L* **do**
  if *V is a responded video* **then**
    fetch all video response of V;
    forall the *video response Vr* **do**
      Add Vr in the experimental dataset;
      manually assign label to each video;
    end
  end
  if *V is a responsive video* **then**
    Add V in the experimental dataset;
    Manually assign label to the video;
  end
end

Algorithm 2: Algorithm for Experimental Data Collection

	Training Dataset	**Testing Dataset**
PVRD	250	1000
CVRD	200	1000
BVRD	61	3389

Table 4.1: Experimental Dataset

Table 4.1 shows the final set of experimental data used to detect spam videos.

4.2 Evaluation Metric

To evaluate the effectiveness of the proposed solution approach, standard information retrieval techniques (precision, recall) are used. Precision (p) of a class X is the ratio of number of videos correctly classified to the total predicted as videos of class X. Recall (r) of class X is the ratio of the number of videos correctly classified to the number of videos present in class X. In order to evaluate these matrics, standard confusion matrix is used with each column of the matrix represents the instances of the predicted class while each row of the matrix represents the instances of the actual class. Each position of the confusion matrix represents the number of elements in that particular class. Table 4.2 shows the standard confusion matrix.

Table 4.2: Confusion Matrix

		Predicted		
		Spam	Unknown	Total
Actual	Spam	a	b	$a+b$
	Unknown	c	d	$c+d$
	Total	$a+c$	$b+d$	N

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

- True Positive (TP) = a/a+b.
- True Negative (TN) = d/c+d.
- false Positive (FP) = b/a+b.
- False Negative (FN) = c/c+d.
  Final accuracy of the system depends on true positives and false negatives.
- Accuracy = a+d/ a+b+c+d.

4.3 Empirical Analysis

This section present the characterization of the contextual features for each sub- problem. Characterization is the process by which features can reveal the behaviour of the video. As we are using contextual features to detect spam videos, in this section we are analysing various contextual features of the YouTube video to find out discriminatory features.

4.3.1 Pornographic Video Response Detection

Pornographic video response is not a spam video response until it is posted as a video response to a normal video (like kids rhyme video, music video, educational video). Presence of pornographic videos as video response to a normal video indicates the spam behaviour. Each video has a specific set of attributes that indicates the type of the video. The aim of this section is to present all discriminatory contextual attributes to detect pornographic video response. We characterize each video by its meta data(contextual features). These set of contextual features is divided into 5 categories: linguistic, temporal, trust, popularity and rating based, YouTube basic, and time based features.

Linguistic features

Percentage of pornographic terms in title and description (PPTT and PPTD): We hypothesize that presence of PPTT and PPTD is an indicator of the video being a pornographic video because it is highly unlikely that non- pornographic videos contain pornographic terms in their title and description. Our hypothesis is based on the observation that more than 50 % pornographic videos contain pornographic terms in title and description (refer to figure 4.1 4.2). A standard dictionary of pornographic words taken from the web [2] [3] is used to match predefined terms in title and description.

[2] http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/
[3] http://home.teleport.com/ stevena/scrabble/expurg.html

Table 4.3: Features

	Abbr.	Feature Title	Feature Type	Remarks
Pornographic Video Response Detection				
P1	PPTT	Percentage of pornographic terms in video title	Linguistic	Terms like sex, kiss, xxx present in title
P2	PPTD	Percentage of pornographic terms in video description	Linguistic	Terms like sex, kiss, xxx present in description
P3	CatV	YouTube category of the video	YouTube Basic	YouTube video category (selected by uploader) such as music, sports, gaming and movies
P4	WTRL	Web of Turst rating of the links	Trust	WoT calculates Reputation and Confidence of the links according to child safety. Reputation is estimate of Turst and Confidence is estimated reliability of the reputation. Ref: http://www.mywot.com/wiki/API.
P5	RSBV	Ratio of number of subscribers by number of views	Temporal + Popularity	Subscriber is an authentic user who has subscribed for a particular video to get regular updates. Analysis reveals that people watch pornographic videos but do not want to be an authentic user.
P6	RLBV	Ratio of number of likes by number of views	Temporal and Popularity	Like feature of the YouTube video shows the popularity of the video. Usually pornographic Videos are not popular videos so ration of Likes by Views is very less.
P7	DYTV	Duration of the YouTube video	Time based	Duration of the YouTube Video shows the length of the Content of the Video. As the Pornographic Videos does not contain much content, their duration is comparatively less.
P8	PVARF	Percentage of videos with age-restricted flag	Trust	Safe Search is the searching feature of YouTube to find out Age- restricted Videos. Pornographic Videos are usually marked as Age- Restricted videos by YouTube itself.
Botnet or Automatic Script Video Response Detection				
B1	TDUV	Time difference between uploaded videos	Time Based	As Videos are posted by an automatic script, time difference between uploaded videos is very less as a human being can not post multiple videos in less than 5 seconds.
B2	NSUB	Number of subscribers of the user	Temporal + Popularity	Botnet videos usually do not contain any useful content so Number of Subscribers (authentic and permanent users) of a Botnet video is very less.
B3	NCYV	Number of comments of the YouTube video	Temporal + Popularity	Number of Comments is the textual response posted by the viewers to the YouTube video. Botnet Videos contains very less (negligible) Number of Comments.
B4	CDUV	Constant duration of the uploaded YouTube videos	Time Based	Most of the Videos are of same duration which shows the Botnet behaviour as usually an automatic script can post all the videos of same duration.
Commercial and Promotional Video Response Detection				
C1	PCTT	Percentage of commercial terms in title	Linguistic	Terms like free, win, subscribe, click present in title which shows the commercial purpose of the Video.
C2	PCTD	Percentage of commercial terms in description	Linguistic	Terms like free, win, subscribe, click present in Description.
C3	NWLD	Number of web links present in description	Trust	Presence of large number of links in description of the Video shows the promotional behaviour of the Video as large number of links is posted just to promote the site of product.
C4	RSBV	Ratio of number of subscribers by number of views	Temporal + Popularity	Promotional videos does not contain any significance, neither they contain any legitimate content. Aim of the promotional videos is to promote their sites and products to gain popularity.
C5	DYTV	Duration of the YouTube video	Time Based	Duration of the Video shows the length of the content present in the Video. As the main aim of promotional videos is just to promote their website or product, duration is either less or constant as no legitimate content is present
C6	NCYV	Number of comments of the YouTube video	Temporal + Popularity	Number of Comments is the textual response posted by the viewers to the YouTube video. As promotional videos do not contain any legitimate content, viewers usually do not post any textual comment.

YouTube Basic Features

Category of the video (CatV) is an indicator of pornographic video response detection. A visual inspection of multiple pornographic videos across their category shows the discriminatory behaviour of the feature as out of total 14 YouTube video categories, more than 75% pornographic videos fall under the category entertainment and people & blogs (refer fig 4.5).

Temporal and Popularity Based Features

Popularity based features such as number of subscribers ,likes and views shows the popularity of the video among users and so can be a good indicator to detect pornographic videos. We fetch the number of subscribers, likes and views of each video response present in our training dataset and compute the ratio of number of subscribers by number of views(RSBV) and number of likes by number of views(RLBV). We hypothesize that low value of RSBV and RLBV signals pornographic behaviour. We confirm the effectiveness of this phenomenon in the evaluation dataset wherein the RSBV and RLBV exhibit a low value (refer to fig 4.7 4.8).

Time Based Feature

We observe the pattern of duration of multiple YouTube videos (DYTV). A visual inspection of this phenomenon clearly shows that duration of the video is a good indicator for pornographic video response detection as around 32% pornographic videos have duration less than 50 seconds and more than 55% videos have duration less than 100 seconds (refer to figure 4.11).

Trust Feature

Some pornographic videos contain web links in their description. Manual inspection of the links shows pornographic behaviour of the links. We have used Web of Trust(WoT) service [4] to detect trustworthiness of the links according to child safety. The WoT reputation system computes website reputations using information received from users and other sources and finally computes the reputation of the link. Lower value of the reputation indicates the less trustworthiness of the link.
Percentage of Videos with Age-Restricted Flag(PVARF): We compute the percentage of videos marked as age-restricted videos. We hypothesize that a significant percentage of videos marked as age-restricted video can be used as a signal for classifying the video as pornographic video. Manual inspection of the pornographic videos confirms the hypothesis as 94 % pornographic videos are marked as age-restricted video by YouTube.

4.3.2 Botnet Video Response Detection

For each video response, we extract all the videos uploaded by the uploader and compute TDUV: time difference between uploaded videos(sort the videos by their uploaded time in ascending order and compute the time difference between each subsequent video). We found that, for videos posted by an automatic script, difference between uploaded time of subsequent videos is very less(less than few seconds). We hypothesize that low value(negligible) of TDUV signals botnet behaviour because it is manually infeasible for a person to upload multiple videos with time difference nearly equal to 0 seconds (refer figure 4.6). We have also found that videos posted by an automatic scripts are of exactly same duration. We hypothesize

[4] http://www.mywot.com/wiki/API

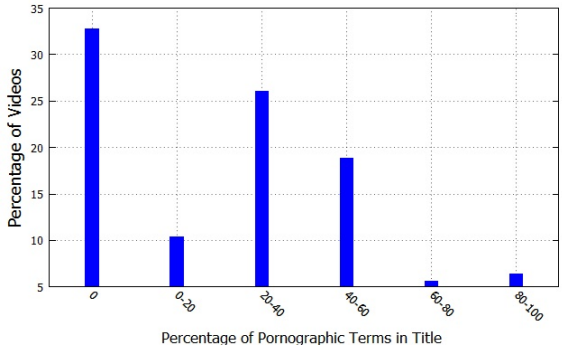

Figure 4.1: PPTT

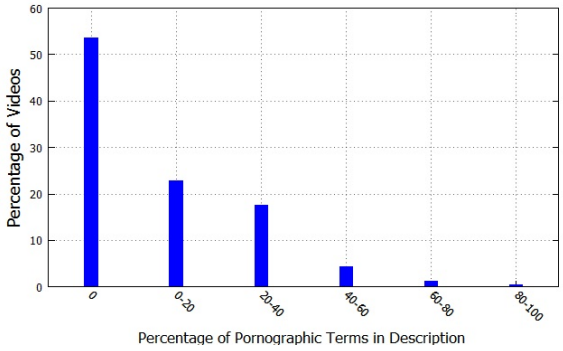

Figure 4.2: PPTD

that constant duration of the videos uploaded by an uploader can be used as a signal to classify videos as botnet videos because it is impractical for a human being to upload all the videos of exactly same duration. We notice several botnet profiles having exactly same duration of all the videos which proofs the hypothesis. We also focus on characteristics of the uploaderś profile such as number of subscribers and video popularity features such as number of comments and by manual analysis, we found that number of subscribers and number of comments of botnet profiles are usually less than those contributed by a legitimate user.

4.3.3 Promotional or Commercial Video Response Detection

Manual analysis of promotional video responses shows that on average around 7-8% video responses of top rated videos are promotional video responses. People post promotional videos as video response to most popular videos to gain popularity or to popularize their sites or products. These videos are neither related to any legitimate video nor do these videos have any content. Presence of such videos as a response to any legitimate video indicates the spam behaviour. The aim of this section is to present all discriminatory contextual features to classify promotional video responses.

Linguistic Features

Percentage of Commercial Terms in Title (PCTT) and Description (PCTD): We hypothesize that presence of PCTT and PCTD shows the promotional behaviour of the video. We confirm the effectiveness of this phenomenon by manual inspection that more than 80% videos contain commercial terms in their description and around 35% videos contain commercial terms in their title (refer figure 4.4 4.3). A standard dictionary of commercial terms taken from the web [5] is used to match predefined terms in title and description.

Temporal and Popularity Based Features

Number of subscribers, views, comments etc shows the popularity of the video. We fetch number of subscribers, views and number of comments (NCYV) of the promotional videos and compute ratio of number of subscribers by number of views (RSBV). We hypothesize that promotional videos are not popular among users so low value of RSBV and NCYV signals commercial behaviour of the video.

[5] http://www.puzzlers.org/pub/wordlists/ospd.txt

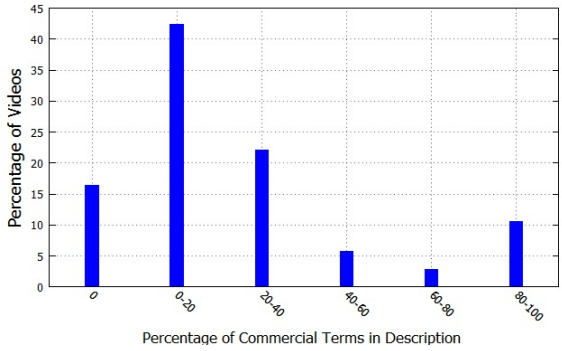

Figure 4.3: PCTD

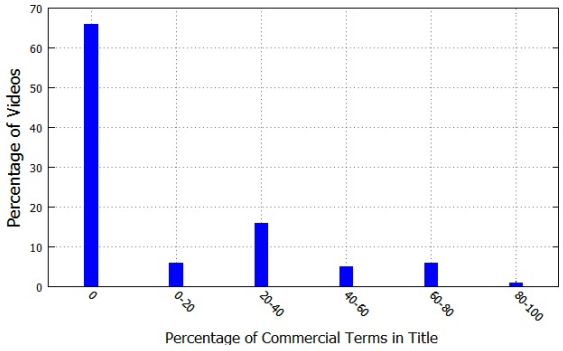

Figure 4.4: PCTT

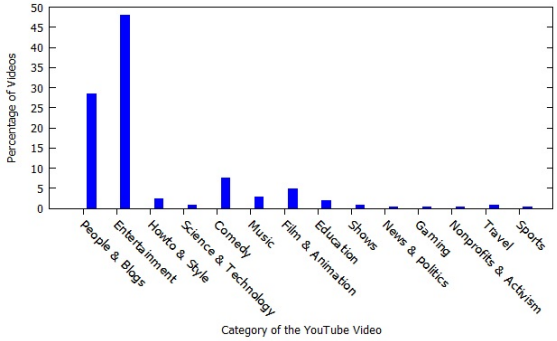

Figure 4.5: CatV

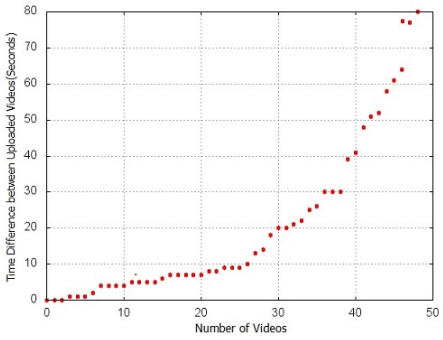

Figure 4.6: TDUV

Time Based Features

Manual inspection of the promotional videos shows that more than 70% of the promotional videos are posted by an automatic script. We fetch the uploader id of the promotional video and then fetch all videos uploaded by the uploader. We found that duration of all the videos uploaded by the uploader is nearly constant.Hence constant value of the duration (DYTV) can be a signal to recognize promotional videos.

Trust feature

Visual inspection of the description of the promotional videos reveals the fact that more than 80% videos contains some links in their description. Around 20% videos contains greater than 5 links in their description (refer figure 4.10). Presence of large number of links can be used as an indicator to recognize commercial behaviour of the video.
Table4.3 summarizes the discriminatory contextual feature set along with feature type.

Summary Figure 4.1 4.2 4.4 4.3 4.7 4.8 4.5 4.6 4.11 shows the analysis of the discriminatory behaviour of YouTube spam videos across multiple features. Figure 4.1 and 4.2 reveals that on average more than 50% pornographic videos contains some pornographic terms in their title and description. We observe several videos with their number of subscribers, likes and views and found that more than 70% videos have very less value of RSBV and RLBV. Figure 4.7 and 4.8 clearly shows that pornographic video responses have very less number of subscribers and likes as compared to their number of views. Figure 4.5 reveals that out of total 14 YouTube video categories, more than 75% videos fall under only two categories. Figure 4.6 shows that botnet users posting a large number of videos in a very small

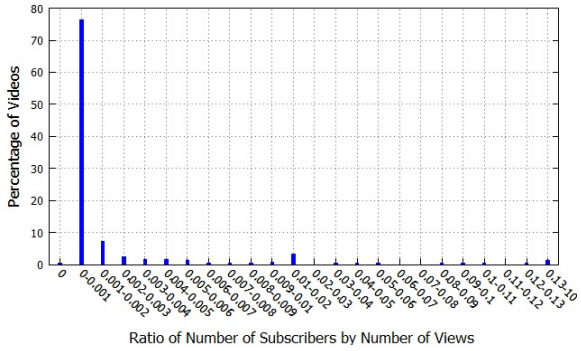

Figure 4.7: RSBV

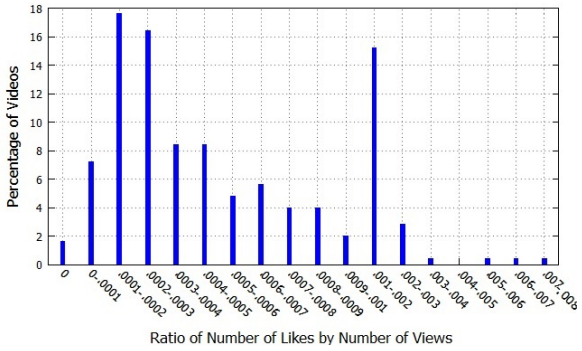

Figure 4.8: RLBV

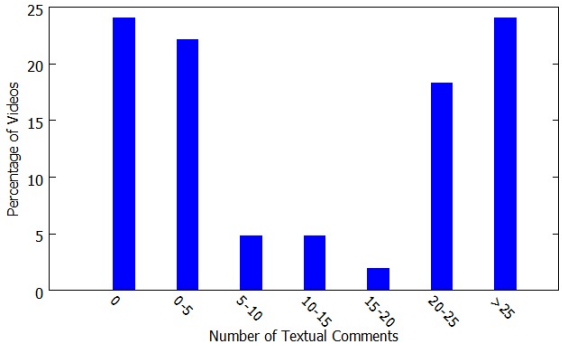

Figure 4.9: Number of comments of the video

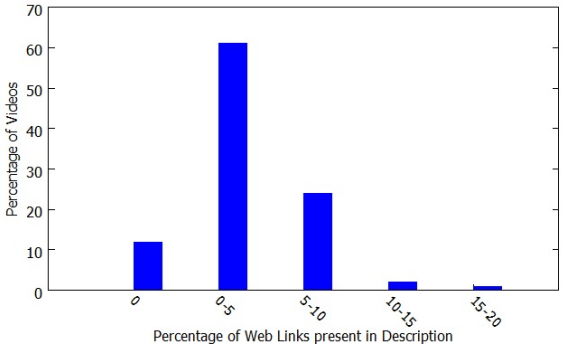

Figure 4.10: Number of links in description

time interval. Horizontal dots parallel to x-axis on same y-axis value shows the number of videos posted in same time duration. Figure 4.11 is a plot of duration of pornographic and promotional videos and reveals that more than 30% pornographic videos are less than 50 seconds in duration and more than 20% promotional videos are less than 50 seconds in duration. Figure 4.3 reveals the fact that 83% videos contain commercial terms like free, subscribe in their description which shows the commercial behaviour of the video. Figure 4.10 shows the presence of large number of links in description shows the promotional behaviour as around 90% videos contains some links in their description.

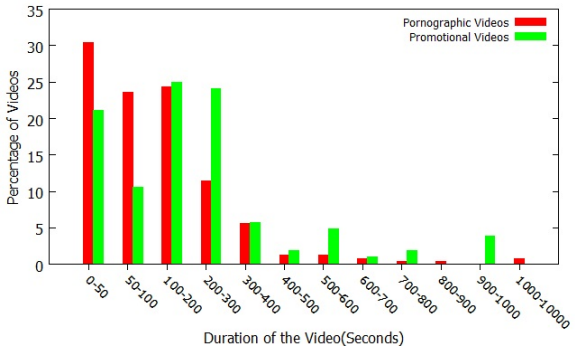

Figure 4.11: Duration of the Video

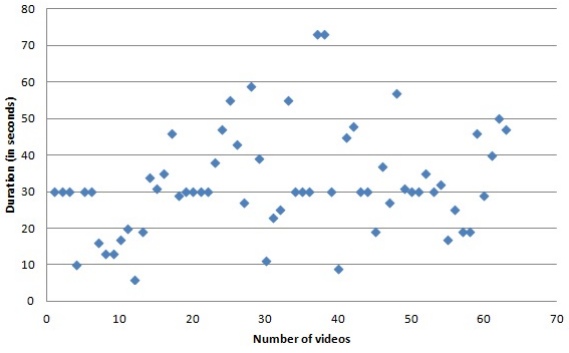

Figure 4.12: Constant time of uploaded videos

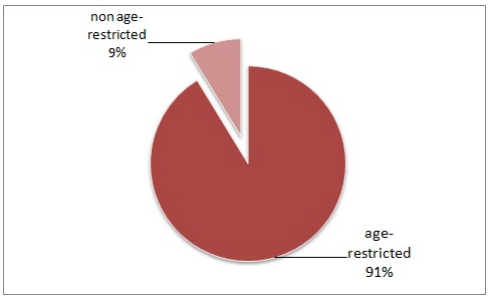

Figure 4.13: Safe search

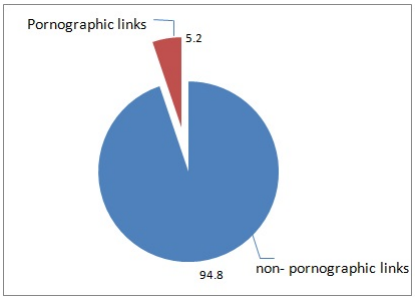

Figure 4.14: Web of Trust

4.4 Classifier Accuracy Results

We experiment with 8, 4 and 6 features for the three classification tasks PVRD, BVRD and CVRD respectively. We apply Algorithm 1 to compute the influence of each feature on the classifier performance. Figures 4.15, 4.16 and 4.17 reveals the impact of each discriminatory feature on the accuracy result of the classifier. The X-axis of Figures 4.15, 4.16 and 4.17 represents the features and Y-axis represents percentage decrease in classification accuracy in absence of the test feature. Feature DYTV and RLBV are two most influential features for the PVRD classifier. As sown in Figure 4.17, the percentage decrease in accuracy is more than 60% in absence of PCTC feature for the CVRD classifier. We learn that while all the 18 features are relevant and has some influence, the degree of influence for various features varies considerably (ranging from less than 10% to more than 70%). The outcome of empirical analysis on feature-impact is estimation of weights for each of the feature in the similarity function. Table 4.4, 4.5 and 4.6 shows the weight matrix for pornographic video response detection, botnet video response detection, and commercial video response detection respectively (the sum of the weights for each classifier is 1). The lesser the weight, the more influential the feature. Table 4.7 (confusion matrix) reveals the accuracy for each of the three one-class classifier. Experimental evaluation shows that the accuracy of the proposed solution is approximately 86% pornographic video responses, 87% of botnet video responses and 83% of commercial video responses. The accuracy results demonstrates the correlation between proposed features or markers and the target class. Figure 4.15 4.16 4.17 shows the effect of each feature on the accuracy of the system. Higher the decrease in accuracy of the system, more important the feature is. To present the classifier accuracy, confusion matrix is used. Table 4.7 shows the accuracy results of the proposed solution approach. The percentage value indicates the recall in each subcategory. Accuracy results shows that approximately 86% pornographic video responses, 87% of botnet video responses and 83% of commercial video responses are correctly classified by the classifier.

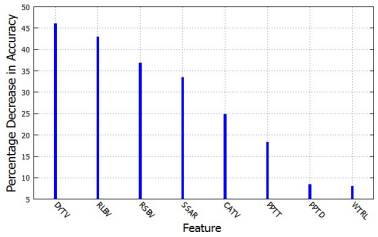

Figure 4.15: Effect of Individual feature on Accuracy of the PVRD

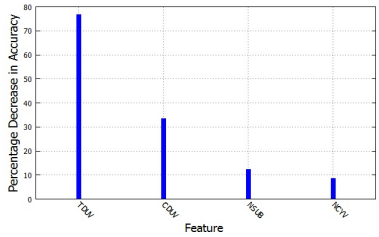

Figure 4.16: Effect of Individual feature on Accuracy of the BVRD

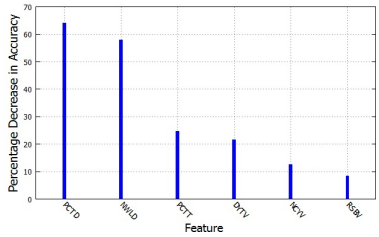

Figure 4.17: Effect of Individual feature on Accuracy of the CVRD

Tab;e 4.4 shows the weight matrix of PVRD. Weight matrix of pornographic video response detection shows that duration is the most important feature to detect pornographic video responses.Ratio of num-

Table 4.4: Weight matrix for pornographic video response detection

PPTT	PPTD	CatV	WTRL	RSBV	RLBV	DYTV	SSAR
0.16	0.2	0.12	0.2	0.09	0.08	0.05	0.1

ber of subscriber by number of views and ratio of number of likes by number of views are important features as compared to number of views, number of number of subscribers and number of likes is very less for pornographic videos, hence the ratio is very less. Percentage of pornographic terms in description and web of trust links are the least influential features. By manual inspection, we have found that around 40% pornographic videos does not contain any description. So in that case, value of PPTD and WTRL in zero.

Table 4.5: Weight matrix for botnet video response detection

TDUV	NSUB	NCYV	CDUV
0.05	0.4	0.35	0.2

Table 4.5 shows the weight matrix of BVRD. Weight matrix shows that time difference between uploaded videos is the most influential feature to detect botnet video responses because almost 99% botnet videos have TDUV value less than 10 seconds hence least weight is given to TDUV feature.

Table 4.6: Weight matrix for commercial video response detection

PCTT	PCTD	NWLD	RSBV	DYTV	NCYV
0.15	0.05	0.075	0.325	0.15	0.25

Table 4.6 shows the weight matrix of CVRD. Weight matrix shows that percentage of commercial terms in description (PCTD) and number of links present in description (NWLD) are the most important features for commercial video response detection because main aim of commercial videos is to promote the sites or products, so description of the video contains large number of commercial terms and web links.

Table 4.7 shows that the proposed classifier recognizes the spam video responses with more than 80% accuracy. The accuracy of the classifier is not 100% because there exist some spam videos, whose meta data does not contain any spam content but while looking at the main content of the video, it is spam. For example, by manual inspection we have found some pornographic videos does not contain any pornographic term in title and description, category of the video is either science and technology or travel and events (which are not related to pornographic behaviour), duration of the video is greater than 3 minutes but while looking at the content of the video, the video was a pornographic video. Presence of such videos decreses the accuracy of the classifier.

Table 4.7: Accuracy Results

Pornographic Video			
		Predicted	
		True	**U**nknown
AC	**T**rue	85.8% (212/250)	14.2% (38/250)
	Unknown	9.2% (69/750)	90.8% (681/750)

Botnet Video			
		Predicted	
		True	**U**nknown
AC	**T**rue	87.18% (34/39)	12.82% (5/39)
	Unknown	1.4029% (47/3350)	98.597% (3303/3350)

Commercial Video			
		Predicted	
		True	**U**nknown
AC	**T**rue	83% (166/200)	17% (34/200)
	Unknown	12.125% (97/800)	87.875% (703/800)

Chapter 5

Conclusion

We present an approach based on a one-class classifier framework to detect video response spam on YouTube. Our findings and performance evaluation results (80% accuracy on an experimental dataset) indicate presence of discriminatory features and reliable indicators in video meta-data which can be exploited for automatically recognizing video response spam. We propose 18 features based on our manual analysis and visual inspection: 8 (pornographic video detection), 4 (botnet or automated script uploader detection) and 6 (promotional video detection) respectively. Our results show that certain features are more informative and influential. We conclude that video meta-data (contextual information and non content based features) can be exploited to recognize video response spam with a reasonable accuracy.

Chapter 6

Future Work

The system presented in this thesis report for video response spam detection detects the spam video responses with more than 80% accuracy. Still to achieve the 100% accuracy, the future work will be to improve the proposed classifier and analyse more contextual features to improve the accuracy of the system.
The work presented in this thesis report is limited to the area of video response spam detection. We did some manual analysis of related video spam also. Related videos are the videos marked as related to the main video. By manual inspection, we have found that, spam exists in related videos also. Presence of spam as a related video of a legitimate video is undesirable and cause several problems like bandwidth waste, decreased user experience, degraded quality. The future work will be to build a tool which will automatically detect related video spam.

Bibliography

[1] Benevenuto, Fabrício, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. "Detecting spammers and content promoters in online video social networks." *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval.* ACM, 2009.

[2] Benevenuto, Fabrício, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, and Keith Ross."Video interactions in online video social networks" *ACM TOMCCAP* 5.4 (2009): 30.

[3] Benevenuto, Fabricio, Fernando Duarte, Tiago Rodrigues, Virgilio AF Almeida, Jussara M. Almeida, and Keith W. Ross. "Understanding video interactions in YouTube." *In Proceedings of the 16th ACM international conference on Multimedia*, pp. 761-764. ACM, 2008.

[4] Benevenuto, Fabricio, Tiago Rodrigues, Virgilio Almeida, Jussara Almeida, Chao Zhang, and Keith Ross. "Identifying video spammers in online social networks." *In Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, pp. 45-52. ACM, 2008.

[5] Sureka, Ashish. "Mining user comment activity for detecting forum spammers in YouTube." *International workshop on USEWOD co-located with WWW2011.*

[6] Heymann, Paul, Georgia Koutrika, and Hector Garcia-Molina. "Fighting spam on social web sites: A survey of approaches and future challenges." *Internet Computing, IEEE* 11.6 (2007): 36-45.

[7] Yang, Yiming, and Jan O. Pedersen. "A comparative study on feature selection in text categorization." *MACHINE LEARNING-INTERNATIONAL WORKSHOP THEN CONFERENCE-*. MORGAN KAUFMANN PUBLISHERS, INC., 1997.

[8] Niu, Yuan, Yi-Min Wang, Hao Chen, Ming Ma, and Francis Hsu. "A quantitative study of forum spamming using context-based analysis." *Proc. of 14th NDSS* (2007).

[9] Cha, Meeyoung, Haewoon Kwak, Pablo Rodriguez, Yong-Yeol Ahn, and Sue Moon. "I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system." *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.* ACM, 2007.

[10] Khan, Shehroz, and Michael Madden. "A survey of recent trends in one class classification." *Artificial Intelligence and Cognitive Science* (2010): 188-197.

[11] Xie, Yinglian, Fang Yu, Kannan Achan, Rina Panigrahy, Geoff Hulten, and Ivan Osipkov. "Spamming botnets: signatures and characteristics." *In ACM SIGCOMM Computer Communication Review*, vol. 38, no. 4, pp. 171-182. ACM, 2008.

[12] Markines, Benjamin, Ciro Cattuto, and Filippo Menczer. "Social spam detection." *In Proceedings of the 5th International Workshop on Adversarial Information Retrieval on the Web*, pp. 41-48. ACM, 2009.

[13] Yang, Yiming. "An evaluation of statistical approaches to text categorization." *Information retrieval* 1.1 (1999): 69-90.

[14] Thomason, Adam. "Blog spam: A review." *Proceedings of Conference on Email and Anti-Spam (CEAS).* 2007.

[15] Jensen, Lee S., and Tony Martinez. "Improving text classification by using conceptual and contextual features." *PhD diss., Brigham Young University. Department of Computer Science*, 2000.

[16] Hayati, Pedram, Vidyasagar Potdar, Alex Talevski, Nazanin Firoozeh, Saeed Sarenche, and Elham A. Yeganeh. "Definition of spam 2.0: New spamming boom." *In 4th IEEE International Conference on Digital Ecosystems and Technologies,* pp. 12-15. 2010.

[17] Wang, Ke, and Salvatore Stolfo. "One-class training for masquerade detection." (2003).

[18] Manevitz, Larry M., and Malik Yousef. "One-class SVMs for document classification." *the Journal of machine Learning research* 2 (2002): 139-154.

[19] Siersdorfer, Stefan, Sergiu Chelaru, Wolfgang Nejdl, and Jose San Pedro. "How useful are your comments?: analyzing and predicting youtube comments and comment ratings." *In Proceedings of the 19th international conference on World wide web,* pp. 891-900. ACM, 2010.