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

Vidushi Chaudhary
Indraprastha Institute of Information Technology (IIIT), Delhi
New Delhi, India
Email: vidushi1116@iiitd.ac.in

Ashish Sureka
Indraprastha Institute of Information Technology (IIIT), Delhi
New Delhi, India
Email: ashish@iiitd.ac.in

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 and commonly used feature 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 of the YouTube video), 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.

I. RESEARCH MOTIVATION AND AIM

The web 2.0 is now building up enormously which consist of search engines, social networking sites, video sharing and photo sharing sites. Specially social networking sites such as Facebook, Twitter, YouTube, Flickr have increased a lot since the last decade which specializes in micro-blogging, video sharing, photo sharing and discussion forums. 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 business discussion, political debates, educational tips etc. YouTube is one of the most popular and wildly used video sharing website (with social networking features) on the Internet.

YouTube provide us functionality to post a textual comment (some text typed by the user) on a particular video. With textual comment, it also has a feature called video response, which is a YouTube function that allow users to comment on video content through video. Video response is most widely used feature of YouTube. We analyzed that on average most viewed and most discussed videos contain more than 1000 video responses which shows the wide use and importance of the feature. Research shows Web 2.0 platforms and social media websites are easy target for spammers and users with malicious intent (because of low barriers to post content and anonymity). Previous studies show that spam as video pollution, video response to uploaded videos, forum comments is prevalent on YouTube and YouTube has taken several measures to counter the spam problem.

To demonstrate the presence of video response spam on YouTube, we manually analyze YouTube video responses. Our manual inspection shows that spam exist as video response to most viewed and most discussed videos. Figure 1 is the screenshot of “All Time Most Viewed” video, and commercial and botnet videos are posted as video response with the intention to promote their site or product. Figure 2 is the screenshot of a child game video with a pornographic video posted as video response.

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

- Large amount of data streams on YouTube every minute, presence of spam in such case cause bandwidth waste (on user side).
- Posting spam on YouTube has several disadvantages like bandwidth waste, 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, child game videos. Presence of porno-
graphic video as video response to a kids video is not legitimate and will have negative impact on kid’s mental growth.

- 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. An automatic script cannot analyse the content of a video [5]. These examples show the importance of the problem.

The focus of the work presented in this paper 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].

The research aim of the work presented in this paper 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.

## II. RELATED WORK AND RESEARCH CONTRIBUTIONS

The work presented in this paper belongs to the area of spam video response detection on YouTube. In this section, we discuss closely related work (to the experiment presented in this paper) 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.

### A. 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].

### B. 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 websites like Wikipedia, 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].

Jansohn et al. presented a framework that combines conventional keyframe-based method with analysis of MPEG-4 motion vectors. They followed two approaches to describe motion vectors. Their experiments demonstrate the effectiveness of the approach [7].

### C. 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[8].

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[9].

D. Research Contributions

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

1) The work presented in this paper 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[1], 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 II) offers a fresh perspective and a novel research contributions of this paper.

2) There has been work done in the direction of analyzing content based features (image analysis, motion and video analysis[2], analyzing IP address) to recognize pornographic, promotional and botnet videos on YouTube. Our study is the first step in the direction of analyzing whether the meta data of YouTube video is also discriminatory enough to recognize spam video responses on YouTube.

3) 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. Empirical analysis of YouTube meta data is the novel research contribution of this paper.

III. PROPOSED SOLUTION APPROACH

In this section, we first define our problem statement and then discuss our novel solution approach for video response spam problem.

A. Problem Statement

1) Let $v^i_{V_i}$ denote a video response $v^i$ of YouTube video $V_i$. Let $\{F_1,...,F_n\}$ be the contextual features (meta data) of YouTube videos $\{V_1,...,V_n\}$. Given a tuple $(v^i_{V_i}, F)$, where $v^i$ is a video response containing feature set $F: \{F_1,...,F_n\}$. Our goal is to find discriminatory feature set $f \subseteq F$ s.t. tuple $(v^i_{V_i}, f)$ can recognize type (spam or unknown) of the video response.

2) Let $v^j_{V_j}$ denote a video response $v^j$ of YouTube video $V_j$. Let $\{f_1,...,f_m\}$ be the feature set, originated due to discriminatory behaviour. Given a tuple $(TS_{n},f_m)$ where $f_m$ is the feature set for each video response present in training dataset $TS_n$. Our goal is to find $(w_{f_1},...,w_{f_m},s_{f_1},...,s_{f_m})$ to implement one-class classifier, where $w_{f_m}$ and $s_{f_m}$ be the weight (influence of each feature to recognize spam videos) and score(similarity ratio of each feature) with training dataset) of feature $f_m$ and finally given a tuple $(v^i_{V_i},w_{f_1},...,w_{f_m},s_{f_1},...,s_{f_m})$, find the label of $v^i_{V_j}$.

B. Proposed Approach

Figure 3 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, the proposed solution approach is a four step process. The first step consists of acquiring positive class training dataset (using YouTube APIs[6] for the purpose of training the classifier. As we are using one-class classification approach, our training dataset contain instances of only positive class i.e. spam class. 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). The next step consists of characterization and identification of discriminatory features. Characterization is the process by which feature can reveal the identity of the object. In our problem, object is a YouTube video and we are trying to analyze whether the meta data of YouTube video can reveal the type of the video. We conduct an in-depth manual analysis and visual inspection of the meta-data 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. 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. 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. The final output is either the test object belongs to the target class or it is unknown (because training dataset does not contain any negative class instances so any object not belongs to the target class is unknown to the classifier).

http://code.google.com/apis/youtube/overview.html
C. Classifier

In one class classification problem, either negative class is not present or it is not properly sampled. We employ one-class classification approach due to the nature of the problem (a recognition task). As conventional multi-class classification algorithms aim to classify an unknown object into one of several pre-defined categories but in our problem context there could be multiple types of videos (number is not defined), so the problem arises when the unknown object does not belong to any of those pre-defined categories. In one-class classification, one class (positive or target class) is well sampled by instances in the training data and the aim is to recognize whether the unknown object belongs to the target class [11] [12]. In this problem we are only concerned about "Is the given video response a spam video response?" If it is not spam it is unknown to us and treated as outlier. 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:

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

\( n = 8 \) 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.

\[
TD = ((x_1, x_2, x_3, \ldots, x_n), y_j), \text{ where } j = \text{Size of Training Dataset, } y_j = \text{Class Label}.
\]

\( n = 8, j = 250 \) in case of PVRD, \( n = 6, j = 250 \) in case of CVRD, \( n = 4, j = 39 \) in case of BVRD.

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

\[
TS = ((x_1, x_2, x_3, \ldots, x_n)), \text{ where } k = \text{Size of Testing Dataset}.
\]

\( n = 8, k = 10,018 \) in case of PVRD, \( n = 6, k = 9,256 \) 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 the recognition of spam video responses. Let \( W_i = \text{Weight of the feature } i \text{ s.t.} \)

\[
\sum_{i=1}^{n} W_i = 1 \quad (1)
\]

Algorithm 2 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 spam video response detection. We are considering lower the weight, more important the feature is.
Input: A list L of features.
Result: Weight of each feature.

initialization:
Assign equal weight to each feature s.t
\[ \sum_{i=0}^{n} W_i = 1 \]  
(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.
  \[ \sum_{\forall i \neq j} W_i = 1 \]  
  (3)
  Run the classifier and check the accuracy of the system, say accuracy2.
  if \((|\text{accuracy}_1 - \text{accuracy}_2| > \text{threshold})\) then
    Removed feature is an important feature and weight corresponding to this feature should be low;
    \[ \text{feature}_{\text{weight}} = \text{feature}_{\text{weight}} - 0.01; \]  
    (4)
  else
    Removed feature is not an important feature and weight corresponding to this feature should be high;
    \[ \text{feature}_{\text{weight}} = \text{feature}_{\text{weight}} + 0.01; \]  
    (5)
  end
end

Algorithm 1: Algorithm for Weight computation of each feature.

One class classification approach is based on similarity computation; we need to find the similarity of the new object with the existing training 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.

\[ S_i = \text{Score of the feature i s.t. } 0 < S_i <= 1 \]

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.

If j is the size of training dataset, then equation to score of numerical feature is:

\[ \text{Score}_{nf} = \frac{\sum_{i=0}^{j} \left| (\text{new}_{value} - TS[i]) / n \right|}{n} \]  
(6)

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.

\[ \text{Score}_{nf} = 1 - \left( x / 100 \right) \]  
(7)

For categorical features, percentage of videos fall into the specific category contributes in finding the score of the feature. Let y = Percentage of videos fall in the particular category

\[ \text{Score}_{cf} = 1 - \left( \frac{y}{100} \right) \]  
(8)

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, cvalue which is the resemblance of the feature with the target class and recognizes the Spam behaviour of the video.

\[ \text{cvalue} = \sum_{i=0}^{n} W_i \times S_i \]  
(9)

IV. EMPIRICAL ANALYSIS AND PERFORMANCE EVALUATION

A. Experimental Dataset

We acquire experimental data using YouTube API. We download all the meta-data (available using YouTube API) of 50 most viewed 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 they 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 11072 videos. Three annotators manually analyzed each video and if two out of three annotators label video as spam video (pornographic, botnet, or commercial) then only respective video is considered as spam and added into the experimental dataset with respective label i.e. pornographic video, botnet video, commercial video or unknown video. We classified 489, 704 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 4.98% of the video responses were pornographic and 7.4% 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. Our training dataset contain videos of only positive class i.e. spam video responses (because of one-class classification approach) and hence normal distribution techniques (40-60 ratio etc) to build training and testing dataset can not be applied in this case. So the size of training dataset is very small compared to testing dataset as testing dataset contain videos of spam class as well as unknown class also to find false positives and true negatives also. The size of the training and testing dataset for the three class of videos are shown in table[1].

<table>
<thead>
<tr>
<th></th>
<th>Training Dataset</th>
<th>Testing Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>PVRD</td>
<td>250</td>
<td>10018</td>
</tr>
<tr>
<td>CVRD</td>
<td>250</td>
<td>9256</td>
</tr>
<tr>
<td>BVRD</td>
<td>61</td>
<td>3389</td>
</tr>
</tbody>
</table>

TABLE I: Experimental Dataset

Our dataset and evaluation results for video response spam detection is publically available at [http://www.iiitd.edu.in/ashish/VRSDataset/](http://www.iiitd.edu.in/ashish/VRSDataset/).

Whole dataset is divided into three different files, one for each category (pornographic video response, promotional video response, botnet video response).

http://www.youtube.com/charts
B. Evaluation Metric

To evaluate the effectiveness of the proposed solution approach, 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.

C. Empirical Analysis

1) Pornographic Video Response Detection: Pornographic video response is considered as spam video response if it is posted as a video response to a non-pornographic video (like kids rhyme video, music video, educational video). Presence of pornographic videos as video response to non-pornographic video shows 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 features 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.

a) Linguistic features: Percentage of pornographic terms in title and description (PPTT and PPTD): We hypothesize that presence of pornographic terms in title and description is an indicator to recognize pornographic videos. Our hypothesis is based on the observation that more than 60% pornographic videos contain some pornographic terms in title while more than 95% non-pornographic videos do not contain any pornographic term in title (refer to figure 4) and around 45% videos contain some pornographic terms present in description while around 92% non-pornographic videos does not contain any pornographic terms in description (refer to figure 5) which shows the discriminatory behaviour of the feature. A standard dictionary of pornographic words taken from the web is used to match predefined terms in title and description.

b) YouTube Basic Features: Category of the video (CatV) is the feature which shows the type of the video i.e. education, entertainment, music etc. A visual inspection of multiple pornographic videos across their category shows the discriminatory behaviour of the feature as out of total 16 YouTube video categories, 48% pornographic videos fall under the category entertainment and 28% pornographic videos are of category people & blogs (refer fig 6).

c) Temporal and Popularity Based Features: YouTube features whose value change with time and shows the popularity of the videos (such as number of subscribers, likes and views) comes under the category Temporal and Popularity. By manual inspection we observe that ratio of number of subscribers by number of views (RSBV) and number of likes by number of views (RLBV) 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 RSBV and RLBV value. We hypothesize that low value of RSBV and RLBV signals pornographic behaviour. We confirm the effectiveness of this phenomenon in the evaluation dataset wherein the RSBV and RLBV exhibit a low value as around 92% pornographic videos have RSBV value less than 0.01 while around 90% non-pornographic videos have RSBV value greater than 0.01 (refer to figure 7) and around 79% pornographic videos have RLBV value less than 0.001 while 78% non-pornographic videos have RLBV value greater than 0.001 (refer figure 8).

d) 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 while around 75% non-pornographic videos have duration greater than 100 seconds (refer to figure 9).

e) 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 to detect trustworthiness of the links according to

---

8 http://urbanoalvarez.es/blog/2008/04/04/bad-words-list/
9 http://home.teleport.com/stevena/scrabble/expurg.html
10 http://www.mywot.com/wiki/API
<table>
<thead>
<tr>
<th>Abbr.</th>
<th>Feature Title</th>
<th>Feature Type</th>
<th>Remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>PPTT</td>
<td>Linguistic</td>
<td>Percentage of pornographic terms in video title. Terms like sex, kiss, xxx present in title.</td>
</tr>
<tr>
<td>P2</td>
<td>PPTD</td>
<td>Linguistic</td>
<td>Percentage of pornographic terms in video description. Terms like sex, kiss, xxx present in description.</td>
</tr>
<tr>
<td>P3</td>
<td>CatV</td>
<td>YouTube Basic</td>
<td>YouTube video category (selected by uploader) such as music, sports, gaming and movies.</td>
</tr>
<tr>
<td>P4</td>
<td>W1KL</td>
<td>Trust</td>
<td>WoT calculates Reputation and Confidence of the links according to child safety. Reputation is an estimated reliability of the reputation. Ref: <a href="http://www.mywot.com/wiki/API">http://www.mywot.com/wiki/API</a></td>
</tr>
<tr>
<td>P5</td>
<td>RSBV</td>
<td>Temporal + Popularity</td>
<td>Ratio of number of subscribers by number of views. 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.</td>
</tr>
<tr>
<td>P6</td>
<td>RLBV</td>
<td>Temporal + Popularity</td>
<td>Ratio of number of likes by number of views. 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.</td>
</tr>
<tr>
<td>P7</td>
<td>DYTV</td>
<td>Time based</td>
<td>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.</td>
</tr>
<tr>
<td>P8</td>
<td>PVARF</td>
<td>Trust</td>
<td>Rate 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.</td>
</tr>
<tr>
<td>B1</td>
<td>TDUV</td>
<td>Time Based</td>
<td>Time difference between uploaded videos. 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.</td>
</tr>
<tr>
<td>B2</td>
<td>NSUB</td>
<td>Temporal + Popularity</td>
<td>Number of subscribers of the user. Botnet videos usually do not contain any useful content so Number of Subscribers (authentic and permanent users) of a Botnet video are very less.</td>
</tr>
<tr>
<td>B3</td>
<td>NCYV</td>
<td>Temporal + Popularity</td>
<td>Number of comments of the YouTube video. Number of Comments is the textual response posted by the viewers to the YouTube video. Botnet Videos contains very less (negligible) Number of Comments.</td>
</tr>
<tr>
<td>B4</td>
<td>CDUV</td>
<td>Time Based</td>
<td>Constant duration of the uploaded YouTube videos. Most of the Videos are of same exactly duration which shows the Botnet behaviour as usually for a human being it is infeasible to maintain such symmetry to post all videos of exactly same duration but an automatic script can do this.</td>
</tr>
<tr>
<td>C1</td>
<td>PCTT</td>
<td>Linguistic</td>
<td>Percentage of commercial terms in title. Terms like free, win, subscribe, click present in title which shows the commercial purpose of the Video.</td>
</tr>
<tr>
<td>C2</td>
<td>PCTD</td>
<td>Linguistic</td>
<td>Percentage of commercial terms in description. Terms like free, win, subscribe, click present in Description.</td>
</tr>
<tr>
<td>C3</td>
<td>NWLD</td>
<td>Trust</td>
<td>Number of web links present in description. 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.</td>
</tr>
<tr>
<td>C4</td>
<td>RSBV</td>
<td>Temporal + Popularity</td>
<td>Ratio of number of subscribers by number of views. 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. So ratio of subscriber by view is very less compared to the non-promotional videos.</td>
</tr>
<tr>
<td>C5</td>
<td>DYTV</td>
<td>Time Based</td>
<td>Duration of the YouTube 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.</td>
</tr>
<tr>
<td>C6</td>
<td>NCYV</td>
<td>Temporal + Popularity</td>
<td>Number of comments of the YouTube video. 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.</td>
</tr>
</tbody>
</table>
child safety. The WoT service computes value of reputation and confidence using information received from users and other sources. Reputation is "what is the reputation of website in the society" and confidence "with how much confidence, WoT is saying reputation is high or low". 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 recognizing pornographic videos. Manual inspection of the pornographic videos confirms the hypothesis as 94% pornographic videos are marked as age-restricted video by YouTube.

2) Botnet Video Response Detection: Botnet video responses are those posted by an automatic script, and not by a human being. Previous research shows that botnet response are considered as spam responses because an automatic script can not analyse the content of an uploaded video and thus can not reply to any video by posting a video response. By manual inspection, we observed that botnet videos do not contain any legitimate content and mostly used for commercial purpose and hence considered as spam. 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 observed 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. Figure 11 is a profile of a botnet uploader and shows that botnet users posting a large number of videos in a very small time interval. Horizontal dots parallel to x-axis on same y-axis value shows the number of videos posted in same time duration. We have also found that videos posted by an automatic scripts are of exactly same duration. We hypothesize 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 unlikely for a human being to upload all the videos of exactly same duration i.e. 30 seconds which proves the hypothesis. We also focus on the characteristics of the uploader’s profile and analyze video popularity features such as number of subscribers and number of comments and by visual inspection, we found that number of subscribers and number of comments of botnet profiles are usually less than those contributed by a legitimate user. As figure 12 shows that around 24% videos does not have any comment get posted and 47% videos have number of comments less than 5 which shows the discriminatory behaviour of a botnet video.

3) Promotional or Commercial Video Response Detection:
Manual analysis of promotional video responses show that on average around 7-8% video responses of most viewed videos are promotional video responses. People post promotional videos as video response to most viewed and discussed videos either to gain popularity or to popularize their sites or products because most viewed and discussed videos have large number of users. These videos are neither related to any legitimate video nor do these videos have any legitimate content. Presence of such videos as a response to any non-commercial video indicates the spam behaviour. The aim of this section is to present all discriminatory contextual features to recognize promotional video response on YouTube.

a) Linguistic Features: Percentage of Commercial Terms in Title (PCTT) and Description (PCTD): We hypothesize that presence of commercial terms in title and description shows the promotional behaviour of the video. While manual analysis, we found presence of certain terms like free, click and win, subscribe for free etc. are present in title and description of commercial videos. 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 some commercial terms in their title while 99.2% videos do not conyain any commercial term in title (refer figure 13 and around 83% commercial videos contain some commercial term in description while around 77% videos does not contain any commercial term in description (refer fig14) which shows the discriminatory behaviour of the feature. A standard dictionary of commercial terms taken from the web is used to match predefined terms in title and description.
b) 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 NCYV and RSBV signals commercial behaviour of the video.

c) Trust Feature: We hypothesize that presence of large number of web links in description of the video indicates the promotional behaviour. Visual inspection of the description of the promotional videos reveals the fact that more than 85% videos contains some links in their description. Around 20% videos contains more than 5 links in their description (refer figure 15). Presence of large number of links can be used as an indicator to recognize commercial videos.

d) Time Based Features: Manual inspection of the promotional videos show that more than 60% 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. Figure 10 shows the profile of a commercial botnet uploader which shows that around 45% videos are of exactly same duration and hence it is a discriminatory feature. Table 11 summarizes the discriminatory contextual feature set along with feature type. By empirical analysis, we finally come up with 18 discriminatory features (8 for pornographic, 4 for botnet, and 6 for commercial) which help us in recognizing spam video responses and separating outliers.

TABLE III: Weight matrix for pornographic video response detection

<table>
<thead>
<tr>
<th>PTTI</th>
<th>PPTD</th>
<th>CatV</th>
<th>WTRL</th>
<th>RSBV</th>
<th>RLBV</th>
<th>DYTV</th>
<th>SSAR</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.16</td>
<td>0.2</td>
<td>0.12</td>
<td>0.2</td>
<td>0.09</td>
<td>0.08</td>
<td>0.1</td>
<td>0.05</td>
</tr>
</tbody>
</table>

TABLE IV: Weight matrix for botnet video response detection

<table>
<thead>
<tr>
<th>TDUV</th>
<th>NSUB</th>
<th>NCYV</th>
<th>CDUV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>0.4</td>
<td>0.35</td>
<td>0.2</td>
</tr>
</tbody>
</table>

TABLE V: Weight matrix for commercial video response detection

<table>
<thead>
<tr>
<th>PCTI</th>
<th>PCTD</th>
<th>NWLD</th>
<th>RSBV</th>
<th>DYTV</th>
<th>NCYV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.05</td>
<td>0.075</td>
<td>0.325</td>
<td>0.15</td>
<td>0.25</td>
</tr>
</tbody>
</table>

D. Classifier Accuracy Results

We experiment with 8, 4 and 6 features for the three classification tasks pornographic video response detection (PVRD), botnet video response detection (BVRD) and commercial video response detection (CVRD) respectively. We apply Algorithm 1 to compute the influence of each feature on classifier performance. Figures 16 through 18 reveals the impact of each discriminatory feature on the accuracy result of the classifier.

To present the classifier accuracy, confusion matrix is used. Table VI shows the accuracy results of the proposed solution approach. The percentage value indicates the recall in each subcategory. Accuracy result shows that approximately 84% pornographic video responses are correctly classified while approximately 16% get mis-classified, 87% of botnet video responses and approximately 86% of commercial video responses are correctly classified and approximately 17% get mis-classified by the classifier. The accuracy results demonstrates the correlation between proposed features or markers and the target class.

V. 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
TABLE VI: Accuracy Results

<table>
<thead>
<tr>
<th></th>
<th>Pornographic Video</th>
<th>Botnet Video</th>
<th>Commercial Video</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Predicted</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>True</td>
<td>Unknown</td>
<td>True</td>
</tr>
<tr>
<td>AC</td>
<td>84.10/4%</td>
<td>15.89/6%</td>
<td>88.52/2%</td>
</tr>
<tr>
<td>Unknown</td>
<td>4.05/72%</td>
<td>(937/9779)</td>
<td>1.40/29%</td>
</tr>
</tbody>
</table>

are more informative and influential. Some interesting findings of our research are:

1. Some pornographic videos fall under the category science & technology, education, travel & events.
2. Out of total 16 categories of YouTube videos, only 2 categories (Entertainment, People & Blogs) are most important categories and more than 75% porn videos fall under these two categories.
3. Some pornographic videos does not contain any pornographic terms in their title and description and hence these are not the most influential features.
4. Some commercial videos have more than 100 links present in description which show the commercial behaviour and hence it is one of the most important feature.
5. Accuracy result shows that only the meta data of the YouTube video is also discriminatory enough to recognize spam video responses on YouTube.
6. Not all features are equally important to recognize spam video response and weight is given to each feature based on their influential behaviour.

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. Our evaluation result shows that the proposed solution approach correctly able to recognize spam videos and outliers with more than 80% accuracy. The accuracy is not 100% because of the following reasons:

1. Training dataset size is small, we manually collect spam video responses and label them as spam or unknown which is a very time consuming task.
2. Spammers marked category of spam (pornographic) videos as education, science & technology which is completed unrelated to the type of the video.
3. Some spam videos does not contain any description, hence text classification technique can not be applied on such videos.
4. Currently we are only using English language dictionary to match presence of spam words in title and description. Some spam videos have their title and description in some other language and hence in such cases text classification technique can not be applied.

VI. ACKNOWLEDGEMENT

The authors would like to thank Akshay Rajwanshi (B.Tech student at IIIT Delhi) for help in manual analysis of the dataset.

REFERENCES