# NIRMALYA: Separating the Wheat from the Chaff in YouTube*


## ABSTRACT

YouTube is the leading social media platform for sharing videos. As a result, it is plagued with spam campaigns that includes spreading malicious links through video description or comments, disseminate adult or illegal content and generate artificial traffic through click baits. We tackle the problem of detecting misleading videos – those having description and title unrelated to the posted content. We show several insights of misleading (spam) behavior modeled through textual and temporal analysis of comments on the videos and by using the characteristics of the uploader and the uploaded video. We develop Nirmalya- a supervised learning framework to detect spam videos that can help prune search recommendations to contain only the legitimate videos. We evaluate our system on a novel manually annotated data set curated from a large corpus of 500K videos. It achieves mean F-score of 0.82 in detecting spam videos with a recall of 0.83.



## KEYWORDS

YouTube Spam, Misleading Metadata, YouTube Spam Detection via Textual and Temporal Features , Social Media Spam


## 1 INTRODUCTION

Since the inception of YouTube in 2005, it has seen an exponential growth in terms of video content as well as user engagement in a variety of fields including entertainment, advertisement, publicity, news, education, etc. Video-based information sharing is gaining leverage over text-based dissemination and has deep social impact. [1] This fact is corroborated by Alexa's web traffic statistics that lists YouTube as the second most visited page globally.

The growing popularity and associated economic opportunities for content providers has triggered the creation and promotion of spam campaigns on these platforms. There are various dimensions of this act: posting advertising links, increasing the view count of a video which in turn is monetized by the incentive schemes common in the video sharing sites, sharing illegal content or copyrighted content, disseminating adult content with misleading description, and promoting campaign that may be spurious and meant for phishing. In particular, let us consider this last motivation behind spam videos. One example is fake computer support companies posting videos purporting to be from reputed anti-virus/anti-malware companies and encouraging users to download some software for scanning and disinfecting their computing devices[2]. Another threat vector is users watching a YouTube video falling victim to drive-by-download attacks whereby malware such as Trojans are being downloaded to the user's device[3].

YouTube, itself, has classified spam videos into many different categories based on their policy[4]. These include video and comment spam, artificial traffic spam, misleading metadata, misleading or racy thumbnails, scams, and blackmail or extortion. According to YouTube: *"Metadata refers to any and all additional information provided on a video. This includes the title, description, tags, annotations, and thumbnail."* In this paper we focus on detecting spam related to subset of misleading metadata i.e. non coherence between video content and its title and description. *The unsolved problem till now is how to automatically and accurately classify spam videos with sufficient accuracy and robustness.*

This form of spam video classification is important because it is a predominant mode of launching the sample attacks discussed above that aims to trick naïve users into clicking unsafe links. Detecting spam videos is a challenging task as compared to detecting scam or artificial traffic spam. This is because these videos are created with an intent to deceive both the consumers as well as the anti-spam algorithms of YouTube. It is not surprising to see that there are several such videos still at large on the platform and have not yet been flagged even after a period of more than six months, e.g., see Figure 1(a) which from its metadata seems to indicate that it is related to the season finale of the highly popular television series "Game of Thrones" but the actual video has only content from previous seasons. In fact, our dataset contains spam videos from the period of 2013 to 2014 that are still not removed.

Most of the work in this domain has been related to detecting spammers or promoters of spam content [4–6, 14] or detecting if response (also video content) to a video is a spam [4]. However, to the best of our knowledge, we are the first to tackle the problem of identifying spam videos with misleading description and title. This covers two of the five categories of spam activity specified by YouTube on its website. We approach the problem by first characterizing spam and legitimate videos through human annotation and creating a rich dataset, which we release publicly to spur further research in this field [5]. Then, we extract a rich set of features for the videos which we categorize into three categories—comments, characteristics of the video, and characteristics of the channel of the posted video—and utilize the set of features in a supervised learning approach for detection. We find that video level and channel level information are good indicators of spam characteristics in many

*In the title *chaff* comes from the Middle English *chaf* which means husk. When extracting wheat from grains, the chaff has to be removed. Similarly, we aim to remove the spam from the legitimate content on a video sharing site such as YouTube. The word NIRMALYA means to be pure and that will be a possible and desired outcome of our system.

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

[2] https://blog.malwarebytes.org/threat-analysis/2013/12/tech-support-scammers-spam-youtube-with-robot-like-warnings/

[3] http://www.spamfighter.com/News-18855-Malware-Attacks-Users-PCs-While-Enjoying-YouTube-Videos-Bromium.htm

[4] https://support.google.com/youtube/answer/2801973?hl=en

[5] https://tinyurl.com/n65ew78




*CIKM'17, November 2017, Marina Bay, Singapore*





ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

cases, however, they are not sufficient by themselves. The comment activity on a video provides very interesting insights as it captures human feedback on the video veracity, relevance and popularity. For this paper, we exclude features related to the actual video content. Our supervised detection system NIRMALYA is trained and


(a)

**Figure 1: A spam trailer for a popular TV series with almost 1M views. It has misleading tags which will cause it to appear in search results for the actual trailer (which appeared at a later date).**

validated on a manually annotated dataset of spam and legitimate videos extracted by crawling YouTube. The high level design of NIRMALYA is shown in the Figure 2. A corpus of videos is collected and then filtered to increase the proportion of spam videos. This is then given to human annotators to flag as spam or legitimate. The final labeled set is used to train an ensemble classifier, which is then used at runtime to predict if new videos are spam or legitimate. In summary, the main contributions of this work are:

- To the best of our knowledge, we are the first to tackle the problem of identifying spam content on video sharing platforms with misleading description and title. We do this by creating a novel dataset dataset of 1690 videos containing 161 spam videos curated from a large corpus of 500k.
- We characterize spam activity in terms of comment activity (temporal patterns and textual information) and statistical features describing the video and the uploader's channel. On a balanced set, NIRMALYA achieves the mean F-score of 0.82 with a recall of 0.83 using 4-fold cross-validation compared with three baselines: (1) 24% higher than a classifier that always predicts the legitimate class with F-score of 0.66 (2) 52% higher than multi-variate Gaussian estimation with an F-score of 0.54. (3) 64% higher than random predictor with F-score of 0.5.
- Finally, we test the robustness of NIRMALYA by training and testing on non-overlapping datasets, calculate precision and recall on spam and legitimate classes and do error analysis of our technique.

The rest of the paper is organized as follows. In Section 2 we present relevant prior work. We provide a description of the data collection and processing methodology in Section 3. Section 4 describes the spamming behavior indicators used in our model. Section 5 gives a statistical analysis of proposed features. Section 6 gives our experimental evaluation. We conclude the paper with some thoughts about future work in Section 7.

## 2 RELATED WORK

Spam detection in web and social media has been a widely researched topic in the academic community. There has been a lot of work to detect and identify spam content in social media [16, 17]. In [9], the author describes a model to detect spam in tagging systems.

Many existing approaches to combating spamming activity are based on extracting evidence from the content of a text. However, as observed in later research [11, 16] these features are not always the best indicators. Mukherjee et al. [11] devise a clever network model which uses group behavior of spammers to tap spamming activities. Viswanath et al. [16] in their attempt to detect anomalous users on social network create an unsupervised technique to model the legitimate and anomalous behavior. There is also some work to identify extremist videos [15] and to unearth unsafe (as in "not appropriate for kids") videos [8].

For video sharing platforms, most of the work has been concentrated on finding spam comments [1, 13] or spammers [4, 5, 7, 14]. The methodology used in [4] tries to identify non co-operative users by analyzing parameters like tags, user profile, the user posting behavior and the user social relations. In [13, 14], the authors present a method to automatically detect comment spammer in YouTube and detect spam comments based on mining comment activity log of a user. In [13], the author aims at detecting comments which are likely to represent spam considering some indicators: a discontinuous text flow, inadequate and vulgar language or not related to a specific context. In [5] authors address the issue of detecting spam in response to a video. In their approach, they manually annotate a large number of potential YouTube spammers and then use user based, video based and social network features to classify a response as spam or legitimate. Related to the above work is [7], in which the problem of video response spam is divided into three sub categories. Although we are also concerned with video spam, we aim to classify any video uploaded as being spam or legitimate according to the criteria that it has misleading title or description, while the above works have only concentrated on videos which are posted as a response to some other video.

A work close but different to ours is seen in the [2] where the authors look at relevance of tags to the video content and we do not use tags in the annotation process as they are not visible to our human annotators. Also, we are only dealing with description and title in metadata. The work closest to our is seen in [6], where the authors build a model to identify videos which are fraudulently promoted using third party promoters e.g. *fiverr.com*. This work only looked at promoted videos, which are distinct from spam videos in that there are external sites through which these videos are promoted, and often by commercial entities with the business

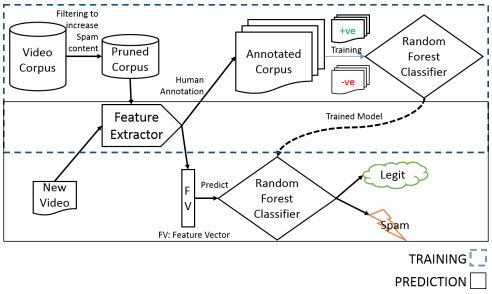



**Figure 2: NIRMALYA: Workflow showing the training phase and the prediction phase. The training phase, done in batch, generates the Random Forest model, which is then used in the prediction phase, on each newly posted video to determine if it is legitimate.**

model of promoting videos on YouTube. Although they study some videos characteristics similar to ours (such as likes and dislikes), NIRMALYA is modeled to identify videos which spam through a rich set of linguistic and temporal features.

## 3 BUILDING A REFERENCE DATASET

Since there is no publicly available dataset which contains spam videos suiting our problem definition of videos with misleading meta-data, we have created a manually annotated data-set for this task. This section describes the data collection and the annotation process.

### 3.1 Crawling

Crawling was done using the YouTube REST data API v3. Videos uploaded between September 2013 and October 2016 were crawled. The annotation and the subsequent manual inspection clearly revealed that there were spam videos still present on the site, e.g., a video with the title "PROOF Obama is a Member of the Muslim Brotherhood"[6] was posted on 23rd Jan, 2014. It has 130$k$ views, 800 likes and 520 dislikes and a manual inspection of the video reveals that the narrator has no actual proof to prove the point made in the description and the video only serves to attract anxious viewers with the controversial title.

YouTube API segregates the videos into several broad categories out of which we chose a few which we thought would be interesting to study [7]. The region code selected for crawling purpose was set to "US", i.e. all the videos were available in USA. A total of 503,824 videos were crawled, which had 60,962,704 comments on them. The parameters collected for each video as well as the comments are shown in Table 1. Apart from video characteristics, the channel information of the video was also crawled and the information available for the channel are also shown in Table 1. The information contained in the comments crawled for each video is shown in Table 1. It is important to mention here that the API does not allow us to crawl all comments for a particular video. Instead it provides an option to crawl comments based on their relevance to the video or in chronological order and provides a sampling based on this criteria. There is a limit to the number of comments available (of the order a thousand) however we did not find it to be fixed across videos. To overcome this sampling limit, we developed a comment crawler in CSS and python to download all the comments associated with a video to ensure integrity of the features studied. The results of the classifier with all comments proved to be better than the results obtained by using only the comments provided by the YouTube API.

[6]https://www.youtube.com/watch?v=BhDVqrjxwUE

[7]Categories: Film & Animation, Autos & Vehicles, Music, Shows, Pets & Animals, Sports, Travel & Events, Gaming, People & Blogs, Comedy, Entertainment, News & Politics, Howto & Style, Education, Science & Technology, Nonprofits & Activism

**Table 1: Dataset Description**

| |
|---|
| **Video Description** |
| view count, comment count, video duration, video licensed content dislike count, like count, type of thumbnails, publish date category, relevant topics, video channel, Up-loader's Google+ account URL |
| **Channel Description** |
| subscriber count, video count, view count, comment count Up-loader's Google+ account URL |
| **Comment Description** |
| Text, Comment Likes, Commenter's Google+ account URL, Upload Date Modified Date, Reply Count, Video Id, Can Rate |

### 3.2 Data Curation and Dataset creation

The desired properties of the dataset we want to use for our experiments are: 1. Have a representative set of videos which capture the demographics of YouTube *i.e.* represent videos that are very popular as well as ones that are moderately and mildly popular. However, bias should be towards popular content because they have more potential to cause harm. 2. Since the percentage of misleading content on YouTube is expected to be much lower in comparison to the legitimate content, our dataset should be appropriately biased to have sufficient number of spam videos. Below we describe our data curation process to create a final dataset that we will use for further processing.

With around 500K videos being crawled and possibly very low percentage of spam content, manually annotating each and every video and searching for spam videos was practically impossible. Random sampling from this set is not guaranteed to capture spam videos as number of spam videos will be much lower on such a highly managed platform. Also, a high percentage of videos would fall in the mildly popular category due to the sheer number of videos uploaded everyday on YouTube. Furthermore, randomly collected sample does not necessarily follow the class distribution in the dataset to achieve classification[18]. Hence, we decided to use some heuristics to narrow down the dataset keeping in mind the desired properties described above. First, we decided to create two non-overlapping datasets with respect to date of upload of videos with the motive to train our classifier on one dataset and test it on another. For our training dataset, we chose the videos uploaded between September 2013 to October 2014. The reason for choosing these dates were two fold: (1) we wanted to capture spam videos that were still at large and not detected by YouTube

anti-spam algorithms. (2) we wanted our classifier to be trained on data that capture sufficient statistics from metrics described in Table 1. For test dataset, we chose the date ranges from September 2015 to October 2016. After categorizing the training and testing datasets in two non-overlapping date ranges, we applied further heuristics on training and the test datasets. For training dataset, all the videos having less than 10,000 views, which is the average views per video for our dataset, were eliminated. This was done to train our classifier on videos having higher visibility and see the effect of testing it on videos with higher and lower visibility. Secondly, videos which had less than 120 comments, which is the average comments per video in our dataset, were eliminated.

On the remaining videos, a smart heuristic was applied to ensure that we remove videos which are more likely to be legitimate. In this method, we first manually identified a handful of spam videos. We then went through comments posted by different users on these videos and collected a few phrases that were commonly appearing in these comments. Few of these comments were "complete bullshit","fake fake fake" etc. Using these as the initial seed set, we looked for videos in our data set that contained at least 2 comments having phrases from the seed set. Using these new videos, we added commonly occurring phrases in our seed set and continued the process. After 3 iterations, we got a set of 4,284 videos which could potentially have nature similar to the spam content we want to identify. The intuition behind adopting this method is that videos which mislead users by posting spam content are more likely to have phrases which vent anger in the form of swear words and phrases. A similar method was adopted for clustering tweets belonging to a particular rumor chain on twitter in [19] with good effect. After this, another heuristic was used — ratio of dislike count:like count of the video for further filtering. The hypothesis to use such a heuristic was that for spam videos, the number of dislikes would be significant compared to the number of likes. Sorting the videos based on the ratio in non-ascending order and taking videos having ratio greater than 0.3 gave us a final set with 650 videos. For test dataset, we used similar methodology with some relaxed conditions: ratio of dislike count:like count was set to be greater than 0.1, average comments on the video were set to be greater than 50, comments had at least one comment from the seed set and number of views were set to be greater than 3k to capture videos of low popularity. Motivation behind relaxing these condition was to have a more realistic dataset that would closely resemble a random sample of YouTube. After applying these conditions, we got 1040 videos. This set now contains videos that had not caught the eye of large population as well as popular videos. It is important to note that after careful curation process, we have made our classification process harder because the filtering has removed many clearly legitimate videos.

### 3.3 Annotating the data

For the training dataset with 650 videos, we created an online annotation task where a user was given the link to a video and was asked to mark it as "spam", "legitimate". We also instructed to mark a video as "not sure" if some ambiguity is present in classifying as spam to minimize bias towards incorrect marking. We provided our annotators with the following instruction to identify a spam video: a video which has title or description not relevant to the content of the video is to be considered "spam". We made 33 separate surveys having 20 videos per survey(one having only 10). and gave it for annotation to 20 volunteering participants at our institutions. This task was repeated for a second round of annotation with the same set of annotators. During the experiments, precaution were taken so that no annotator marked the same set of 20 videos in the two rounds. The results of the two rounds of annotation can be seen in Table 2. On analyzing the results, we identified that there was lack of unanimous decision on several videos as seen in Table 3 which describes the (dis-)agreement between annotators in the two rounds of annotation. Each cell of Table 3 can be used to understand the annotator agreement for the different labels. For example, the first cell denotes that 70 videos which were marked spam in round one of annotation were also marked spam in round two of the annotation. We see that inter-annotator agreement was not perfect, an issue which has been reported repeatedly in prior works for annotation tasks in social media [3, 12]. We believe that this was due to fact that our dataset contains videos with at least 10k views and 120 comments. As a result we are dealing with several videos which at the outset may appear to be legitimate due to several reasons, e.g., there is a credible conversation thread on the video, the content looks legitimate at first glance but is actually morphed, or it is uploaded by a channel which looks reputable. Thus, the decision as to whether these videos are spam or legitimate, was subjective and challenging in nature. The discrepancies in the annotations were then resolved by another tie-breaker round. A graduate student volunteer then went through all the video content and annotations and if any ambiguity lied in characterizing the video as spam, the video was marked as "not sure". The statistics from the two rounds as well as the final annotation are shown in Table 2. The distribution of spam/legitimate in different categories of the pruned dataset is shown in Table 4.

**Table 2: Statistics from the two rounds of annotation**

| | Round1 | Round2 | Final Annotation |
|---|---|---|---|
| Spam | 158 | 130 | 123 |
| Legitimate | 400 | 422 | 423 |
| Not Sure | 92 | 98 | 104 |

For test dataset, we distributed 1040 videos among 52 volunteers with each volunteer getting to annotate 20 videos. After first round of annotations, we got 809 legitimate, 111 spam and 120 not-sure labels. Similar to the previous approach, we did a second round of annotation and used labels only for which both annotators agree. In the end of round 2, we ended up with 38 spam, 856 legitimate and 146 not sure videos. Presence of only 4% spam in our testing dataset down from 20% spam in our training set justifies our pruning process. If we relax the filtering conditions, we are likely to get very low percentage of spam in the dataset and it is extremely difficult to annotate larger datasets, in search of spam videos. From our datasets, there are some categories which were crawled but were completely eliminated in the pruning process because the crawled videos from those categories did not contain enough suspects of spam.

**Table 3: Annotator (dis-)agreement from Rounds 1 & 2**

| | Spam | Legitimate | Not Sure |
|---|---|---|---|
| Spam | 70 | 62 | 26 |
| Legitimate | 54 | 308 | 38 |
| Not Sure | 6 | 27 | 59 |

**Table 4: Category Wise distribution of Spam**

| | **Spam** | **Legitimate** |
|---|---|---|
| Entertainment | 66 | 262 |
| News & Politics | 41 | 95 |
| Film & Animation | 12 | 31 |
| Music | 4 | 22 |
| Shows | 0 | 13 |

# 4 FEATURES FOR DISTINGUISHING SPAM

Any spam detection technique relies on features which can help distinguishing spam and legitimate by modeling their behavior. In this section we propose three classes of features, which we believe can serve this purpose and help to build a robust classifier model.

## 4.1 Channel Level Indicators

These features capture the information about the channel through which a video was uploaded. The hypothesis is that channels which upload more legitimate videos would have much more positive likes, comments and subscribers etc.

- **Comment Count Ratio:** Ratio of total number of comments in that channel to the total number of videos, i.e., the average number of comments per video.
- **View Count Ratio:** Ratio of total number of view of that channel to the total number of videos, i.e., average number of views per video.
- **Subscriber Count Ratio (Video):** Ratio of total number of subscribers of that channel to the total number of videos.
- **Subscriber Count Ratio (View):** Ratio of total number of subscribers on that channel to the total number of views on videos of that channel.The last three features will all measure the average popularity of the videos posted on the channel.

## 4.2 Video Level Indicators

This category of features covers information about the video through the use of the YouTube API. Features picked for the study are following:

- **Video definition:** It can take only two values - *HD* (High Definition) or *SD* (Standard Definition). We hypothesize that legitimate videos would have a higher quality definition on average than spam videos.
- **Licensed Content:** It can take only two values - *true* or $false$. We hypothesize that legitimate videos would be more likely to have licensed content than spam videos.
- **Comment Count Ratio:** Ratio of total number of comments to the total number of views. This feature would measure the level of user interaction with the video.
- **Like Count Ratio:** Ratio of total number of likes for the video to the total number of views. This feature would measure the average likeness rating of the video per view.
- **Dislike Count Ratio:** Ratio of total number of dislikes for the video to the total number of views. This feature would measure the average dis-likeness rating of the video per view.
- **Dislike to Like Ratio:** Ratio of total number of dislikes in the video to the total number of likes in the video. This feature can be an important indicator of average approval rating of the video by its viewers. We hypothesize that spam videos are more likely to have a higher dislike to like ratio.

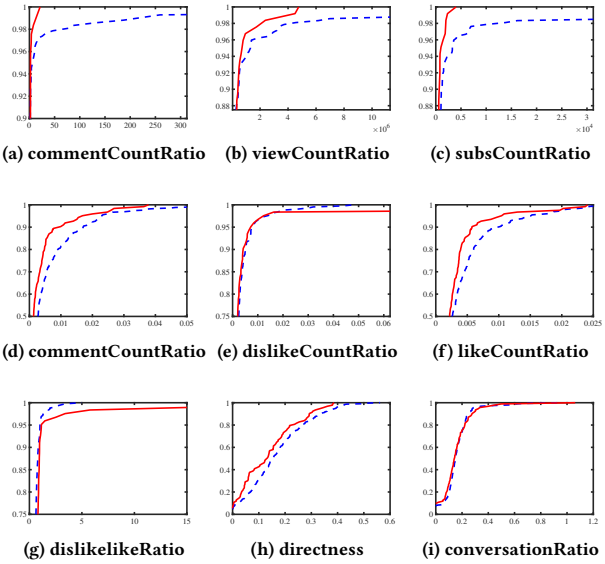
## 4.3 Comment Level Indicators

Comments allow users to express themselves more freely compared to expressing a binary opinion (like or dislike). Hence we believe that mining comments can give us several indicators to model spam behavior. We obtain all the comments using web scraping and also use the subset of comments available through the YouTube API. Here we describe two kinds of indicators based on textual or temporal patterns.

### *4.3.1 Linguistic Indicators.*

- **Inappropriateness score:** This score is used to capture the extent of inappropriateness in the comments, i.e., number of swear or cuss words used in the comments. To achieve this, we created a set of words using LIWC[8] swear words and their synonyms. We then count the number of uni-grams and bi-grams (concatenated as a single word) present in the set of swear words. The score is normalized by a factor of $2n-1$ where $n$ is the number of words in the comment. This is because there are $n$ uni-grams and $n-1$ bi-grams in a comment of $n$ words.
- **Directness:** This feature captures how directed a comment is toward the video as opposed to directed toward some user (up-loader or another user on the comment thread). Our hypothesis is that comments which were made as a response to some user and not directed toward the video have very little impact in helping to distinguish the legitimacy of the video itself. If the comment contains any user mention in the form <user-name>, then we assign a score of 0 signifying non-directness else a score of 1 signifying directness. Finally, we use what fraction of total comments are directed as the score for this feature.
- **Conversation Ratio:** It is the ratio of number of conversation comments to the total number of comments. A comment is defined to be a conversational if it has at least one reply to it or is a reply itself i.e. it is part of a conversation. We want to analyze whether conversation ratio feature proves to be helpful in categorizing a video as spam or not.

[8] http://liwc.wpengine.com/



(a) commentCountRatio (b) viewCountRatio (c) subsCountRatio

(d) commentCountRatio (e) dislikeCountRatio (f) likeCountRatio

(g) dislikelikeRatio (h) directness (i) conversationRatio

**Figure 3: Cumulative % of spam (solid red) and legitimate (dashed blue) groups vs. feature value. (a), (b), (c) are Channel level features. (d), (e), (f), (g) are Video level features. (h) & (i) are Comment level features. It can be observed that in all features denoting video popularity, legitimate group have higher feature values for same cumulative % as compared to spam group(and vice-versa for other features which denote non-popularity)**

- **Similarity score:** We create two bag-of-words models – for the video description and the comment in consideration. Video description contains the words used in video title, description and tags. We then compute the Jaccard similarity between these two sets of words and report it as the similarity score. We believe that this can also capture the relevance of a particular comment towards the video.

*4.3.2 Temporal Indicators.* Arrival rate of comments on videos can be a good indicator for spam and legitimate videos. The arrival rate behavior could be starkly different at different phases in the lifetime of a video based on whether it is spam or legitimate. For example, spam videos with racy content are more likely to have a lot of hits during the initial phase followed shortly by comments which refute the video content. On the other hand, the legitimate videos may also have hits in the initial phase but not a similar comment pattern. Legitimate videos are also expected to have a steadier decline in comment rate as compared to spam videos. This is because spam videos will likely have accumulated a high number of dislikes very quickly and hence people are less likely to visit and comment on them. This kind of behavior has also been observed and used in prior art to create models for detection of anomalous entities on social networks [16]. We created a feature vector of size 365 representing 365 days since the date of upload. The value for each feature is set to the comments posted on that day normalized by total number of comments on that video. Through this we aim to capture the growth of comments on a video over a period of one year.

**Binned Features** : For 'Inappropriateness' and 'Similarity' scores, we further wanted to capture the distribution of scores in the video of a particular class by creating bins. Each bin's index signifies a fraction of the range of the score and the value in that bin is the number of comments with its score in that range. We tested our classifier with different bin sizes ranging from 1 to 10 and optimal number of bins was found to be 3. This may be attributed to the fact that when we use more bins, the distribution across bins becomes more sparse and there is less distinction between the legitimate and the spam distributions across the bins.

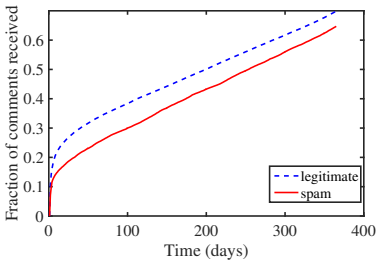



**Figure 4: Growth of comments with time for Legitimate and Spam videos**

**Comment Sets** The comments of a video are further segregated based on several parameters that we believed would be useful to gauge the importance of comments in classifying the spam and legitimate videos:

- **Relevant comments:** This set contains only the comments when crawled using the *relevance* parameter specified in the API. This parameter is used by YouTube to rank comments by relevance. It should be noted here that this parameter is not available when all comments are crawled using web scraping. But all comments crawled does contain the relevant comments as its subset and yield better results as compared to only relevant comments.
- **Liked Comments:** This set contains comments that had at least 1 like. We think that liked comments carry a sense of approval with them since they have been liked by peers.
- **Liked, Relevant Comments:** Comments that had at least 1 like and were also relevant as per the API.
- **All Comments:** The set of all comments crawled for a video.

## 5 DISCRIMINATING POWER OF INDIVIDUAL FEATURES

To check if the indicators described in the earlier section have a reasonable discriminating power in detecting spam, we perform empirical analysis to study the distribution of spam and legitimate groups across the feature dimensions. We call these microbenchmark experiments. Figure 3 shows the cumulative distribution function (cdf) for spam and legitimate videos individually for each feature, spanning the entire range of the feature's normalized values. This is performed for all video and channel level indicators. It is intuitive that

the number of comments, subscribers, and views towards a channel with more spam videos would be less compared to one with legitimate videos. Also, the channel which posts spam videos will always tend to post further spam videos. We can see that for all features that capture popularity/approval of the video (a, b, c, d, e, g), the spam group has much higher % of videos within a smaller score. Whereas, in the features which measure disapproval i.e. *dislikeCountRatio* and *dislikelikeRatio* (f, h), the legitimate group has most videos with smaller score and the spam group has comparatively much higher scores. From Table 6, we can see that the percentage of spam videos having *HD* definition is much less than the percentage of legitimate video having *HD* definition. This would imply that the up-loaders of spam video are less concerned with the quality of video. Similarly, the percentage of spam videos having licensed content is relatively less compared to that in the legitimate videos. With this empirical validation, we believe that the video and channel level indicators will be helpful in the classification task.

We also plotted the distribution of comments of spam and legitimate groups for the inappropriateness and similarity score but they did not yield a discriminating power. This may be due to a similar writing trend on both types of videos and use of profane words have become commonplace in social media. In Table 5, we see the comparison of some of the comment level indicators for the two groups for the two sets of comments - *AllComments* and *LikedComments* (comparison with other sets is omitted for brevity). We find that *%Directness* is much higher for spam group. Our hypothesis is that in spam videos the viewers direct most comments at the videos than at other peers. *ConversationRatio* also has discriminating power with legitimate videos having higher conversation ratio. We will do top feature selection test in section 6.2 to corroborate this empirical analysis.

**Table 5: Contrast of mean values of some comment level features for spam and legitimate groups . "CTP" stands for Comment Temporal Pattern**

| | All comments | | Liked comments | |
|---|---|---|---|---|
| Features | Spam | Legitimate | Spam | Legitimate |
| % Directness | 22 % | 16.56 % | 12 % | 6.56 % |
| Conversation Ratio | 0.081 | 0.084 | 0.251 | 0.279 |
| CTP-Day 1 | 0.079 | 0.118 | 0.149 | 0.206 |
| CTP-Day 2 | 0.027 | 0.042 | 0.0330 | 0.048 |
| CTP-Day 3 | 0.021 | 0.022 | 0.0245 | 0.025 |
| CTP-Day 7 | 0.005 | 0.009 | 0.00575 | 0.010 |

As can be seen from Figure 4, the rate at which the comments arrive for legitimate videos are higher than for spam videos. This is because people are more interested to watch and comment on legitimate content. In Table 5 we also see that more than 11% of the total comments are written on the first day for legitimate videos whereas the number is lower for spam videos. Similar trend can be observed for the Liked comments also. We hence believe that temporal indicators can be very helpful to distinguish spam and legitimate behavior whereas linguistic features, contrary to our initial hypothesis based on intuition, are not very discriminating. We find through our classifier evaluation (described next) that even the features that appear to be discriminating in these microbenchmarks, when used individually fail to generate a classifier with high enough accuracy to be usable in practice.

## 6 EXPERIMENTAL RESULTS

In this section, we describe our experimental framework in detail and present the classifier results on our training set for each class of features and finally with all the features combined. We then train our classifier using training dataset and evaluate its performance on our test dataset in section 6.6. We now present the results for training dataset in subsequent sections. For all our experiments, we created a balanced dataset by randomly selecting 123 videos from 422 videos marked legitimate and 123 spam videos that were annotated as such. We experimented with several learning methods using the scikit-learn library including SVM, Decision Trees, AdaBoost, Gradient Boosting and Random Forest (RF). The RF algorithm consistently performed better than the rest. All our experiments were subsequently performed using RF as the classification tool. Further, all the experimental results are presented using 4-fold cross validation. In each run, original set is partitioned into 4 sets with 3 used for training and 1 used for testing. In 4-fold validation, this is repeated four times with testing on each subset exactly once and then averaging the results. We repeated 4-fold validation experiments for 50 times with random seed for shuffling each time and then averaging the results. We have presented the best and the mean results for classification.

| Features | Spam | Legitimate |
|---|---|---|
| % HD definition | 57% | 73% |
| % Licensed content | 65% | 77% |

**Table 6: Contrast of mean values of features that measure video quality for spam and legitimate videos**

### 6.1 Evaluation Metrics and Baselines

For the training and test datasets, we use precision, recall, and F-score as the evaluation metrics for classification which are used according to their standard definitions. We will also look at the confusion matrix resulting from these metrics for test dataset. Since there was no prior baseline for comparison, we have created a synthetic baseline comparison algorithm. For this, we assumed that both spam and legitimate video are generated from different multivariate normal distributions. For training, we use the Gaussian Kernel Density Estimation method and learn the density functions from which these two classes may potentially arise. For testing we calculated the probability of each data point arising from spam kernel density function and legitimate kernel density function and assigned it to the class with the higher probability. After 4-fold cross validation we got Precision and Recall for spam class to be 0.5 and 0.58 with a F-Score of 0.54. In addition to this baseline, we also used the classifier that always predicts the majority class, which gives an F-score of 0.66. We also used the random predictor that gives a baseline F-score of 0.5.

## 6.2 Macro Evaluation and Comparison with Baselines

We experimented by using the different categories of features explained in Section 4 and finally using all the features together to form the macro benchmark evaluation. The results are reported in Table 7. For the mean F-score reported, standard deviation was within 2% for all the models except for the Comment Liked + Relevant feature for which it was 4%. From the table we see that all variants of NIRMALYA have mean F-scores better than the random baseline. This includes using Channel features, Video features, and Comment features individually.

**Table 7: Classification Results for different models. 'F', 'P' and 'R' refers to F-score, Precision and Recall respectively. For comments, 'L' means Liked subset and 'R' means Relevant subset.**

| Model | F | P | R |
|---|---|---|---|
| Random Prediction Baseline | 0.5 | 0.5 | 0.5 |
| Majority Class Prediction Baselie | 0.66 | 0.5 | 1.0 |
| Gaussian Baseline | 0.54 | 0.5 | 0.58 |
| Video Features | 0.58 | 0.60 | 0.56 |
| Channel Features | 0.62 | 0.62 | 0.62 |
| Comment Liked Features | 0.52 | 0.52 | 0.53 |
| Comment Relevant Features | 0.57 | 0.62 | 0.53 |
| Comment Liked + Relevant Features | 0.51 | 0.50 | 0.52 |
| All Comments | 0.557 | 0.566 | 0.555 |
| Video, Channel, Liked + Relevant (first month comments only) | 0.686 | 0.699 | 0.715 |
| **Video, Channel, All Comments (All available comments first year)** | **0.82** | **0.83** | **0.83** |

In the comment features, we use various classes of comments, such as 'Liked', 'Relevant' etc. We see that using all comments, which were not available from YouTube API but rather, only through our web scraping technique, yields better results than only using the Liked + Relevant comments from the YouTube API. This confirms our hypothesis that all comments does capture a deeper representation of user interaction in determining spam or legitimate content. But still, none of these features alone yield significantly better results than the baselines described above. Finally, we see that the best results are obtained by combining all comments from 1st year with Video and Channel features that yield a mean F-score of 0.82 with high precision and recall. This shows that any feature themselves are not enough and by combining all features together, our classifier performs significantly better than all baselines.

## 6.3 Feature Importance Ranking

In order to verify the importance ranking of these features we compute top ten features from standard RF feature selection method provided by scikit learn library. We compute the importance score of each feature in every iteration of cross-validation. We then average the scores for each feature and then provide the top ten features in order of importance in Table 8. Previously, we observed that including comment features significantly improves the results. This is also evident if we look at the importance ranking of the features in table 8. Two of the comment features, *ratioDirected* and *ratioConversation*, are among the top 3 important features as per the RF feature ranking. It confirms our intuition that videos with spam content are less likely to have conversations but will have most comments directed towards the video criticizing it. Table 8 also shows that the like and dislike statistics of the video are good indicators of the spamicity of the video because the dislike-to-like ratio and average number of dislikes per video feature high in the list. However, it is important to note that a legitimate but unpopular video may also have such statistics and hence video features are not sufficient by themselves.

**Table 8: Feature Importance Ranking. 'ch','v','sub','conv' refers to channel, video, subscriber and conversation respectively.**

| Importance Ranking | RF Feature Selection |
|---|---|
| 1 | videoDislikeLikeRatio |
| 2 | commentRatioConversation |
| 3 | commentRatioDirected |
| 4 | videoCommentCountRatio |
| 5 | VideoDislikeCountRatio |
| 6 | SubscriberCountRatio(video) |
| 7 | commentDay1 |
| 8 | commentDay324 |
| 9 | videoLikeCountRatio |
| 10 | channelViewCountRatio |

## 6.4 Temporal Characterization of NIRMALYA

From the last section, we saw that using comment features helped improve our detection model. We now study the temporal behavior of NIRMALYA with respect to comments. An important issue in such a problem is detecting spam at an early stage of upload and thwarting them. However, the video statistics which performed quite well in the previous section may not be available for early detection (say within 1 month) since the like, dislike and comment statistics for a video will only stabilize after several months of upload.

Here we characterize the performance of our classifier first by using all comments features only and then by considering the comments for a specific period of time from the upload date of the video. Again 4-fold cross validation is ran 10 times and the results are reported in Table 9. Standard deviation lies within 2%. It can be observed that in 10 days NIRMALYA achieves an F-score of 0.61 which is 13% better than the Gaussian baseline protocol. One use case of this temporal study is that content providers can characterize spam videos during early days of upload and manually investigate the content of such videos before it causes more harm. Still, only using the comments does not perform better than our majority class prediction baseline. This temporal characterization strengthens the fact that we need to use all the features to get best classification results.

**Table 9: Temporal Performance using only All Comments features.**

| Days from Upload | F-score | Precision | Recall |
|---|---|---|---|
| 1 | 0.58 | 0.60 | 0.56 |
| 5 | 0.59 | 0.61 | 0.58 |
| 10 | 0.61 | 0.62 | 0.60 |
| 15 | 0.57 | 0.59 | 0.55 |
| 30 | 0.57 | 0.59 | 0.54 |
| 60 | 0.51 | 0.54 | 0.53 |
| 90 | 0.55 | 0.56 | 0.54 |
| 180 | 0.57 | 0.56 | 0.59 |
| 360 | 0.58 | 0.57 | 0.59 |

## 6.5 Category-wise performance of Nirmalya

We also want to evaluate the robustness of using a system like Nirmalya on the different categories of videos present in YouTube for spam detection. We use the same categories as shown in Table 4. For such an evaluation we decided to use a leave-one-out policy — train a model using videos from all categories except a particular category (say *News & Politics*) and then test on videos of that category. For the same category we also train a model containing videos from all categories (including itself) and then compare the results. We show the results in Table 10. For both experiments the RF learner with the same settings as in earlier experiments was trained and tested for the top 2 categories (categories were ranked by the significant number of videos in our dataset). The other categories had too few videos to be useful for any statistical significance. The results are shown in Table 10. It can be seen that a model trained without including a particular category and testing on that categories performs much worse than the model trained on all categories. We think that this could be explained by the observation that the nature of channel, video and comment characteristics change significantly across categories and therefore a model which is not trained on a particular category does not perform optimally on it. This phenomenon has been observed in a different yet related problem of detecting fake product reviews on Amazon as studied in [10]. The authors address a similar problem in developing a well generalizable model across "Hotel", "Restaurant" and "Doctor" categories and find that the same model does not apply across these three categories. We therefore are able to see that Nirmalya performs better when the training set includes videos from that category.

**Table 10: Category Level Performance of Nirmalya on the top 2 categories. P, R, F represent the best Precision, Recall and F-Score for leave-one-out policy. P*, R*, F* is the best result obtained by training on all categories. For both experiments a Random Forest model was used with 4-fold cross-validation.**

| Category | P | R | F | P* | R* | F* |
|---|---|---|---|---|---|---|
| News & Politics | 0.62 | 0.63 | 0.63 | 0.76 | 0.70 | 0.73 |
| Entertainment | 0.58 | 0.56 | 0.57 | 0.60 | 0.61 | 0.60 |

## 6.6 Evaluation on Test Dataset

To test the robustness of Nirmalya, we train it on the training dataset and test it on the testing set. Instead of using all features, we will use the results of the features here related to the Video, Channel and All comments for first month. We limit the comments to one month, because for the videos uploaded in second half of 2016, we do not have first year comments available yet. So, to keep uniformity across the dataset, we will use the results of the said feature in this section. Our test set contains 856 legitimate videos and 38 spam videos (only 4%). We trained Nirmalya on imbalanced training dataset (123 spam + 422 legitimate videos) and it predicted all the videos to be legitimate in the test dataset. If we use the original set without balancing, then classifier will favor legitimate videos because there are more of them. So, Nirmalya was trained on balanced dataset of 246 videos (123 spam + 123 legitimate) from the training data set. We ran it ten times, because in each iteration, the sub-sampled videos from the balanced dataset will be different. For ten iterations, if a spam video is predicted as spam in at least 6 runs out of 10, we mark it as correctly predicted. The average results for the ten iterations are presented in the Table 11. It can be seen that the classifier is able to catch 53% of spam videos and misclassify around 47% of spam videos. On spam class, Nirmalya achieves an F-score of 0.12 which is 71.4% higher than the baseline classifier that always predicts legitimate class. Recall on spam class is 53% and precision is 75% higher than the baseline of 0.04. On the other hand, Nirmalya classifies 67% of the legitimate content at the cost of misclassifying 33% legitimate videos as spam. It achieves a recall of 67%, precision of 97% and an F-score of 0.79 on legitimate class.

The fact that our training dataset contains only 4% of spam videos with a lot of videos having insufficient statistics in terms of low number of views, low number of comments etc., Nirmalya is able to achieve in an accuracy of 66%. Also, using all comments from first year will further improve the results. Practically, Nirmalya can be used by the content providers as starting point to reduce their efforts of manually flagging the videos. They can initially flag all spam classified videos to warn the viewers of possible spam content. And after manually checking, they can clear the flags from the videos which are actually legitimate. System owners or content provider can characterize around two third of legitimate videos correctly. The goal might be to review the statistics of legitimate content or provide ads on those legitimate videos for monetary benefits.

**Table 11: Classification of Spam and Legitimate Videos on Test dataset**

| | | **Predicted** | |
|---|---|---|---|
| | | **Legitmate** | **Spam** |
| **True** | **Legitimate** | **571(67%)** | 285(33%) |
| | **Spam** | 18(47%) | **20(53%)** |

## 6.7 Error Analysis

Referring to Table 11, we see a misclassification rate of 47% for spam videos. In order to analyze why our classifier was unable to catch

those spam marked videos, We had to manually investigate the uploaded videos. Some interesting insights were gained from the manual inspection. For example let us look at "The Flat Moon over the Flat Earth" video [9] , then we will provide general observations for all the spam videos that were misclassified as legitimate. Here the video uploader is claiming that moon is flat. This channel has around 106k subscribers. The reason for misclassification becomes clear when we investigate the statistics in context of our classifier's top features given in Table 8. First, the video has around 2100 likes and only around 650 dislikes. That gives a dislike to like ratio of of 0.3. Second, the channel is extremely popular with most videos having views greater than 10K. Third, we can see that most comments have long conversations and most of the comments are actually corroborating the uploader's claims. So, analyzing these factors, we come to a conclusion that our top features namely vDislikeLikeRatio, commentRatioConv, commentRatioDirected, vDislikeCountRatio and vLikeCountRatio and feature related to channel are behaving exactly the way that they should behave for legitimate videos. That is why our classifier is unable to classify this video as spam. Another observation for misclassified videos was that they contained very high number of dislikes as compared to likes but channel was extremely popular (with greater than 100K subscribers). In this case, dislike to like ratio (although the top discriminating feature) alone cannot help classify the video as spam because other channel, video and comment features dominate the decision of classifier. Some misclassified videos had roughly equal number of likes and dislikes and the channel was not very popular. They also contain very small conversation ratio and directed comments thus giving insufficient statistics to extract meaningful features to classify them correctly.

After doing error analysis on all misclassified spam videos, we also wanted to gain insights on why legitimate videos were classified as spam. Among the 285 legitimate videos classified as spam, we randomly sampled 5 videos and extracted general insights from them. Some of them were fan made videos of popular films or games. Although the video uploader described it in the description, but still those videos garnered very high number of dislikes as compared to likes. Also for some videos, channel features, commentRatioConv and commentRatioDirected had the similar behavior as observed for spam thus misleading the classifier into believing them as spam. However, further analysis is required to develop deeper insights.

## 7 CONCLUSIONS AND FUTURE WORK

In this paper, we presented Nirmalya, a supervised learning framework to detect spam videos in online video sharing portals such as YouTube. Spam videos are defined as those having misleading metadata, in terms of title and description being unfaithful to the content of the video. Extensive experiments confirm that using a variety of video, channel, and comment features, Nirmalya could detect the spam videos with a recall of 0.83 and an F-score of 0.82. Future work will involve doing a deeper analysis of the comments' contents to extract features and patterns indicative of the spam videos. We also plan to utilize processing of video frames to help in spam content detection. Further improvement can be made by analyzing transcripts of speech from YouTube API's and combine fact checking techniques to help improve the detection purpose.

[9] https://www.youtube.com/watch?v=fH7BjIzXWOg

## REFERENCES


[1] Ahmad Ammari, Vania Dimitrova, and Dimoklis Despotakis. 2011. Semantically enriched machine learning approach to filter YouTube comments for socially augmented user models. *UMAP* (2011), 71–85.

[2] Payal Bajaj, Mridul Kavidayal, Priyanshu Srivastava, Md Nadeem Akhtar, and Ponnurangam Kumaraguru. 2016. Disinformation in Multimedia Annotation: Misleading Metadata Detection on YouTube. In *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*. ACM, 53–61.

[3] Hila Becker, Mor Naaman, and Luis Gravano. 2010. Learning similarity metrics for event identification in social media. In *Proceedings of the third ACM international conference on Web search and data mining*. ACM, 291–300.

[4] Fabrício Benevenuto, Tiago Rodrigues, Virgílio Almeida, Jussara Almeida, and Marcos Gonçalves. 2009. Detecting spammers and content promoters in online video social networks. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. ACM, 620–627.

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

[6] Vlad Bulakh, Christopher W Dunn, and Minaxi Gupta. 2014. Identifying fraudulently promoted online videos. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 1111–1116.

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

[8] Rishabh Kaushal, Srishty Saha, Payal Bajaj, and Ponnurangam Kumaraguru. 2016. KidsTube: Detection, Characterization and Analysis of Child Unsafe Content & Promoters on YouTube. *arXiv preprint arXiv:1608.05966* (2016).

[9] Georgia Koutrika, Frans Adjie Effendi, Zoltán Gyöngyi, Paul Heymann, and Hector Garcia-Molina. 2007. Combating spam in tagging systems. In *Proceedings of the 3rd international workshop on Adversarial information retrieval on the web*. ACM, 57–64.

[10] Jiwei Li, Myle Ott, Claire Cardie, and Eduard H Hovy. Towards a General Rule for Identifying Deceptive Opinion Spam.. In *ACL (1)*. Citeseer, 1566–1576.

[11] Arjun Mukherjee, Bing Liu, and Natalie Glance. 2012. Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on World Wide Web*. ACM, 191–200.

[12] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 309–319.

[13] Cristina Radulescu, Mihaela Dinsoreanu, and Rodica Potolea. 2014. Identification of spam comments using natural language processing techniques. In *Intelligent Computer Communication and Processing (ICCP), 2014 IEEE International Conference on*. IEEE, 29–35.

[14] Ashish Sureka. 2011. Mining user comment activity for detecting forum spammers in youtube. *arXiv preprint arXiv:1103.5044* (2011).

[15] Ashish Sureka, Ponnurangam Kumaraguru, Atul Goyal, and Sidharth Chhabra. 2010. Mining youtube to discover extremist videos, users and hidden communities. In *Asia Information Retrieval Symposium*. Springer, 13–24.

[16] Bimal Viswanath, M Ahmad Bashir, Mark Crovella, Saikat Guha, Krishna P Gummadi, Balachander Krishnamurthy, and Alan Mislove. 2014. Towards detecting anomalous user behavior in online social networks. In *23rd USENIX Security Symposium (USENIX Security 14)*. 223–238.

[17] Alex Hai Wang. 2010. Don't follow me: Spam detection in twitter. In *Security and Cryptography (SECRYPT), Proceedings of the 2010 International Conference on*. IEEE, 1–10.

[18] Gary M Weiss and Foster Provost. 2001. The effect of class distribution on classifier learning: an empirical study. *Rutgers Univ* (2001).

[19] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. 2015. Enquiring minds: Early detection of rumors in social media from enquiry posts. In *Proceedings of the 24th Int. Conference on World Wide Web*. Int. WWW Conferences Steering Committee.