"Video Unavailable": Analysis and Prediction of Deleted and Moderated YouTube Videos

Maram Kurdi[†*], Nuha Albadi[‡*], Shivakant Mishra[*]
[†]Department of Computer Science, Taif University, Taif, Saudi Arabia
[‡]Department of Computer Science, Taibah University, Medina, Saudi Arabia
[*]Department of Computer Science, University of Colorado Boulder, Boulder, USA
{maram.kurdi, nuha.albadi, mishras}@colorado.edu

***Abstract*—YouTube strives to moderate its content by censoring, demonetizing or removing videos that allegedly violate their community guidelines. Such strategies, especially if seen as unjust by the affected users, could be met with resentment, anger, and in some cases, violence. In addition to YouTube removing videos, uploaders sometimes delete their videos for a variety of reasons such as paraphrasing or preserving online self-image. In this paper, we provide a detailed analysis of deleted/removed videos on YouTube. To do this, we tracked over 73,000 recent YouTube videos for one week and identified those that got deleted or removed. We have then conducted a large-scale analysis of this data and reported on the most informative features that distinguish deleted/removed videos from the ones that remain available. Based on our analysis, we have developed machine learning prediction models that predict videos that will get deleted/removed at different stages of a video's lifetime, viz., at the time of posting, and after up to seven days have elapsed. Our findings indicate that we can predict video deletion/removal with high accuracy even at the time of posting—a strategy that could help users perceive the removal of their videos as fair as well as reduce public and moderators exposure to problematic videos.**

***Index Terms*—deletion, prediction, deleted videos, YouTube**

I. Introduction

Social media platforms strive to moderate their content to make it healthy and safe for their users. However, the immense volume of user-generated content that is being shared on social media every day makes it infeasible to rely only on human moderators [1, 2]. As a result, most social media platforms adopt machine learning models to flag and take down posts that allegedly violate their community guidelines [3, 4]. Researchers have also contributed to this automated moderation strategy by developing models that can detect specific types of toxic content such as bullying [5], hate [6], and radicalization [7, 8, 9]. However, this automated content moderation process is bound to suffer from errors, which can cause users to perceive this practice as unjust [10, 11]. It is not only social media platforms that may remove content, users themselves sometimes delete their own postings for various reasons such as regret, typos, paraphrasing, bullying, and/or preserving online self-image [12, 13].

Researchers in HCI are increasingly interested in the topic of online content moderation and deletion in various online communities including Reddit [11, 14], Twitter [15, 16], Facebook [13], and Instagram [17]; YouTube is an exception here in that there is a lack of any such studies for it.

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

Every minute, over 500 hours of videos are uploaded to YouTube, some of which are by people who rely on YouTube as their means of income. While a large number of videos are uploaded to YouTube, those that are deemed inappropriate may get censored, demonetized, and/or removed. For example, YouTube reported that they removed over nine million videos during the second quarter of 2019 [18]. However, many YouTube content creators have expressed their frustration towards what they perceive as unfair censorship and demonetization practices by YouTube [19], with some users unfortunately resorting to violence to express their discontent regarding their content getting removed or censored [20]. Given the sensitivity of the moderation task and how it may negatively impact people's psychology and/or earnings, content removal and restriction should be exercised with great care and accuracy.

Despite the immense popularity of YouTube and the possible negative impacts of deleted content on users, to the best of our knowledge, there hasn't been any prior work to examine content moderation and deletion of YouTube videos. Hence, a large-scale analysis to investigate content deletion on YouTube is of paramount importance. Furthermore, an ability to provide any pre-posting alert about possible deletion of a video would be extremely valuable for YouTube users. For example, it was reported recently that a majority of the users expressed frustration after their posts got removed by the community moderators on Reddit [11]. Further, it was reported that the users who did not suspect the removal of their posts ahead of time, perceived their post removal as unjust and were reluctant to post again. On the other hand, users who suspected that their post will get removed, perceived the removal as just and their posting behavior remained unaffected [11]. This suggests that any type of pre-posting alert to the YouTube users, e.g. that their post violates the guidelines, and allowing them to edit their post to avoid future removal would be highly valuable.

Pre-posting alert addresses another important issue about problematic posts. According to YouTube, 32.2% of the problematic flagged YouTube videos (e.g, spam, hate, harassment, cyberbullying, child abuse) received views before they were taken down [18]. Exposing YouTube users and moderators to such problematic and possibly disturbing content can negatively affect their mental well-being [21]. A pre-posting alert can help reduce the amount of problematic content seeing the light of day and hence benefit the community at large.

In this paper, we address these concerns by developing a pre-posting machine learning prediction model that can predict

at the time of posting which videos will very likely get deleted . To do so, we first need to understand the features that distinguish deleted videos from other existing ones. In particular, we examine the following research questions:

- RQ1: What characteristics distinguish deleted videos on YouTube from the ones that do not get deleted?
- RQ2: Is it possible to predict video deletion with high accuracy? How early can we make such a prediction?

To address these research questions, we first collected a total of 73,983 recent YouTube videos and kept track of their updated metadata (e.g., comments, views) and status (i.e., available or unavailable) over time for a period of one week. We found that deleting videos is not a rare occurrence on YouTube, with 17.3% of the collected videos being deleted within a week of their posting. We then performed a detailed statistical analysis of this dataset to identify the characteristics that clearly distinguish unavailable videos from available ones. Based on our analysis, we have developed supervised machine learning classifiers trained on informative textual, content, temporal, and statistical features to predict video deletion at the time of posting and after posting a video. Our results show that we can accurately predict video deletion at a very early stage of the video's lifetime, even at the time of posting.

Our work offers important insights into video deletion on YouTube and marks a first step towards automatic prediction of unavailable videos. The findings presented in this paper suggest that is feasible and important for YouTube to alert uploaders at the time of posting their videos about any possible forthcoming deletion. Such a pre-posting alert tool can help reduce frustration or regret that users may face. Additionally, platform designers can benefit from adopting such a tool by reducing moderators' and public's exposure to disturbing videos that can be caught early by the proposed system.

II. Related Work

A significant body of work has examined content deletion and content moderation in online communities. To situate our study, we review prior work in these two areas of research.

While online communities moderate their content to preserve a healthy and safe community for its users, moderation sometimes may have a negative effect on people's online social experience. For example, Jhaver et al. [11] found that the majority of users who encountered content moderation felt frustrated about their posts being removed. As well as others have been adversely affected professionally, socially, and emotionally [10]. In contrast, Srinivasan et al. [14] found that non-compliant comments rate reduced immediately after removing the problematic comment.

A large body of work demonstrated that online community users delete their posts as a way to protect their privacy, or deal with regrets. For example, Almuhimedi et al. [15] found that tweets with residence-tagged locations are more likely to get deleted than other locations. Wang et al. [13] found that posts deletion is a common strategy among Facebook users to handle regrettable posts. Tufekci [22] concluded that female Facebook users tend to untag themselves from Facebook posts and delete information from their profile more often than male.

Considerable efforts have been devoted on developing models to predict online posts that will eventually be deleted. Zhou et al. [23] were able to accurately predict tweets that will get deleted due to regrets based on tweet content and user's historical deletion behavior. Another study by Chancellor et al. [17] analyzed deleted pro-eating disorder Instagram posts and built a classifier to distinguish them from the still available ones. Their results showed that captions and tagging can predict pro-eating disorder post status (deleted or still alive).

While content deletion has been studied on Twitter, Facebook, and Instagram, content deletion and removal is significantly understudied on YouTube. However, there exists a substantial body of work that analyzes YouTube for other purposes including characterizing and predicting YouTube videos popularity [24], inappropriate videos targeting kids [25], and videos promoting extreme ideologies [26]. YouTube-hosted comments has also been studied to analyze commenter demographics [27] and predict their comments popularity [28].

Given the lack of research on content deletion and moderation on YouTube, our work begins to fill this gap by providing insights into what distinguishes deleted and moderated videos from those still available. Additionally, we propose a systematic approach to flagging problematic content at a pre-posting stage, which would serve the online community as a whole (i.e., moderators, uploaders, and general users).

III. Data Collection

One key challenge in analyzing deleted videos and perhaps the reason for why there hasn't been any work analyzing such videos is that there are no repositories of unavailable YouTube videos. So, our first task was to develop a methodology for collecting deleted videos. So, we tracked a range of videos after they were posted and recorded all the reactions to those videos along with all metadata. So, by the time some of those videos got deleted, we had already collected their information.

We leveraged YouTube Data API v3 [1] to collect the videos' metadata. Unfortunately, YouTube Data API does not provide the option of retrieving random videos. To tackle this, we provided keywords to retrieve videos identified by those keywords. YouTube has a predefined list of sixteen categories. To collect videos from all these categories, we carefully curated English words and proper names from different websites, and then we randomly selected up to 100 keywords for each category to minimize selection bias.

We were mainly interested in tracking videos within one hour of their posting. YouTube API does not have the option of retrieving recent videos directly. Instead, we obtained recent videos by setting two API parameters (publishedAfter and publishedBefore) to capture videos published within the last one hour. YouTube videos can sometimes be blocked due to some copyright claim depending on the geographical location where a video is being retrieved. To ensure that we analyze

[1] https://developers.google.com/youtube/v3/

only deleted videos and not blocked ones, we used the default regionCode parameter (default value is US) to retrieve videos that were available in US. So, the set of deleted videos we analyze in this paper are the ones that were actually deleted and not blocked for any reasons. Also, we made sure that we only collect the data of unique video ids, as in some cases one video id could be retrieved by more than one keyword. For each retrieved video id, we collected the following metadata associated with it: video comments, topic details, video statistics, and content details.

Due to the limited number of quotas enforced by YouTube and to ensure that we collected videos published at different days of a week, the data collection was done in four phases spread over about six weeks. In each phase, we started by retrieving videos posted within last hour and then tracked those videos over seven days. Overall, we tracked 18,501 videos in Phase 1 starting on Feb 7, 2019, 18,600 videos in Phase 2 starting on Feb 16, 2019, 18,541 videos in Phase 3 starting on Feb 27, 2019, and 18,341 videos in Phase 4 starting on Mar 9, 2019 for a total of 73,983 videos in four phases. The same keywords were used in every phase to retrieve up to a maximum of 50 recent videos for each keyword.

A. *Video Deletion*

Due to the restrictions imposed by YouTube API on video access frequency, we checked the status of each video five times a day (instead of continuously) for one week. Each time, we first checked to see if the video still exist or not as the API does not provide a deletion notice. We did this by checking the ”items” property in the JSON object that is returned when querying YouTube API using a video ID. The ”items” property is empty for a video that has been deleted from YouTube. In this case, we labeled the video as deleted and no longer checked for updates. Otherwise, we inserted a new record in the video database reflecting the updated video statistics.

B. *Ethics*

To conduct this study as ethically as possible, we have collected only the metadata of publicly accessible videos on YouTube through the official YouTube API. In particular, we did not downloaded the video itself nor its thumbnail. We approached the dataset with an “eyes off” methodology where our analysis was done only by the algorithms we developed. Finally, We have not disclosed any personal identifying information about the uploaders such as username or video ids.

IV. Video Analysis

The analysis in this section pertains to all data collected in the four phases, unless otherwise stated. We refer to the dataset with deleted and undeleted videos as *”our dataset”*, to the one with only deleted videos as *”the deleted dataset”*, and to the dataset with only undeleted videos as *”the undeleted dataset”*.

A. *General Statistics*

Table I summarizes some of the general statistics of our dataset. Of the 73,983 videos that we tracked, 17.3% were deleted within seven days of their posting, which is quite significant—approximately one out of every six videos is deleted. 23.7% of these deleted videos had their comment section disabled. On the other hand, only about 1.8% of the undeleted videos had their comment section disabled.

TABLE I: General statistics of videos in our dataset.

# of videos	73,983
# of deleted videos	12,808
# of undeleted videos	61,175
# of videos with comments disabled	4,150
# of deleted videos with comments disabled	3,040
# of undeleted videos with comments disabled	1,110

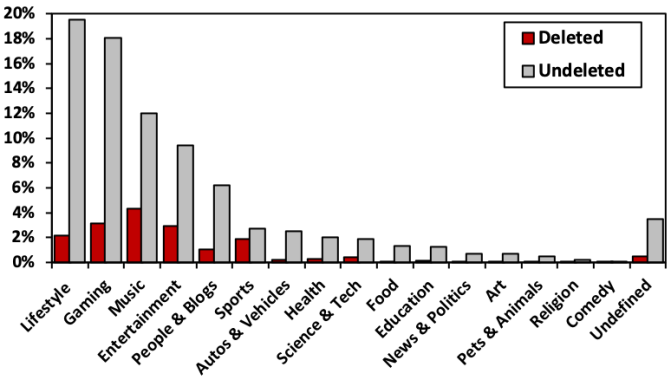

Fig. 1: Category proportions for deleted vs. undeleted videos.

B. *Video Category*

Figure 1 shows the distribution of the categories for deleted/undeleted videos. Our findings indicate that the distribution of deleted/undeleted videos differs significantly among the sixteen categories (χ^2 = 3,176, df= 15, $p < .001$). The main observation is that the music category is the most common category in the deleted dataset, even though it is the third most common category in our dataset. The most common category in our dataset, namely lifestyle, ranks fourth category in the deleted dataset. Gaming is the second most common category in both the deleted dataset and our dataset.

To determine whether a relationship exists between the video categories and deleted/undeleted videos, we built a regression model to analyze the size and the direction of the relationship between the dependent class (i.e., deleted or undeleted) and the independent variables (i.e., video categories). We then used logistic regression to investigate the relationship between our categorical dependent and independent variables. Table II shows top five per topic regression coefficient; a positive coefficient indicates an association with the deleted class, a negative coefficient indicates association with the undeleted class, and the absolute value of the coefficient represents the strength of association. As shown in Table II, videos related to sports, music and entertainment are most likely to get deleted, while videos related to religion, food, pets and animals are the least likely to get deleted.

TABLE II: Per topic regression's coefficient.

Category	Coef.	Category	Coef.
Sports	+1.55	Religion	-1.05
Music	+0.92	Food	-0.99
Entertainment	+0.78	Pets&Animals	-0.84
Science&Technology	+0.43	Autos&Vehicles	-0.47
People&Blogs	+0.19	Education	-0.41

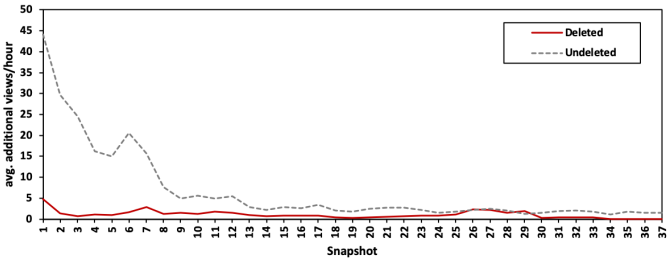

Fig. 2: Average additional views per hour per snapshot (about 5-hour interval) for deleted and undeleted videos.

C. *Video Statistics*

*1) **Video Engagement**:* Here we look at the statistics that describe audience engagement such as number of views, likes, and dislikes. We found that on average deleted videos receive less views (μ = 311, std = 10,773, max = 840,717) than undeleted ones (μ = 2,072, std = 58,416, max = 8,158,053). Similarly, on average, deleted videos received less likes (μ = 5, std = 91, max = 5,431) than undeleted videos (μ = 60, std = 1928, max = 394,604). Interestingly, we found that on average deleted videos got less dislikes (μ = 0.4, std = 8, max = 585) than undeleted ones (μ = 3, std = 73, max = 9,888). We found this difference in views, likes, and dislikes to be statistically significant (Mann-Whitney U = 32x10^7, 21x10^7 and 29x10^7 respectively, $p < .001$) indicating that deleted videos in general attract less audience engagement than the undeleted ones.

However, to truly understand the difference in audience engagement between deleted/undeleted videos, we need to take into account *survival bias*, which is that undeleted videos have more time to accumulate engagement than the deleted ones. So, we compute average engagement *per hour* for deleted and undeleted videos within each snapshot of our dataset. Each snapshot is about a five-hour time interval, and only videos that were alive in that snapshot were considered for calculating the averages per hour. In the following subsections, we provide this fine-grained comparison.

*2) **Video Views, Likes and Dislikes**:* We investigated the variation in number of views per hour per snapshot that deleted/undeleted videos received over time (See Figure 2). We found that the variation in the average number of views for both deleted/undeleted videos follow the same pattern over time, with the number of views being significantly higher in the earlier snapshots and falls down quite rapidly in later snapshots. However, in the earlier snapshots, the average number of views of undeleted videos is significantly higher than of the deleted videos. The average number of views of deleted videos drops to nearly zero views as the video age approaches about 180 hours mark. In fact, not getting much views could be a reason for uploaders deleting their videos.

We further looked at the average number of likes and dislikes per hour per snapshot for the deleted and undeleted videos in our dataset. We notice that average number of likes and dislikes for both deleted and undeleted videos follow the same pattern over time as with average number of views (Figure 2). The number of likes and dislikes is significantly higher in the earlier snapshots and falls down quite rapidly in later snapshots. Further, in the earlier snapshots, the average number of likes and dislikes received for undeleted videos is higher than the average number of like and dislikes received for deleted videos respectively.

*3) **Video Duration and Tags**:* On average, we found that deleted videos tend to be longer in duration (μ = 941 sec, std = 1,999 sec, max = 42,901 sec) than undeleted videos (μ = 767 sec, std = 2,085 sec, max = 43,020 sec). This difference in duration was found to be significant (Mann-Whitney U = 35x10^7, $p <$.001). We also looked at the number of tags provided by the uploader when uploading a video. These tags are important for ranking videos in YouTube search results. On average, deleted videos used 11 (std = 9, max = 72) tags per video, while undeleted videos used 12 (std = 10.4, max = 95) tags per video. The most common tags used in the deleted videos are: *Hack, Roblox, Sims, Free, Free play*, respectively. While the most common tags used in the undeleted ones are: *News, Game, Entertainment, PlayStation, Fortnite*, respectively. We conclude that gaming tags were the most used tags for both deleted/undeleted videos.

D. *Video Comments*

We looked at the comments for deleted/undeleted videos to get more insight into whether discussions under a video have any association with the video getting deleted. On average, deleted videos received 1.1 comments (std = 13.8, max = 742) whereas undeleted videos received 10.4 comments (std = 295, max = 65,491). Thus undeleted videos in general receive more comments. Note that this finding is in line with our earlier finding about video engagement, i.e. deleted videos receive much less audience engagement than the undeleted ones.

*1) **Comment Sentiments**:* For our sentiment analysis, we applied VADER [29], that was designed for analyzing social media sentiments, to each comment in our dataset and captured their compound score, which is then normalized between -1 (extremely negative) and +1 (extremely positive).

We found that deleted videos tend to have lesser positive comments and more negative comments than undeleted ones. The distribution of comment sentiments among the deleted and undeleted videos also differs significantly (χ^2 = 137, df = 2, and $p < .001$). This is a somewhat expected result, where a video that receives more negative comments and less positive comments is deleted.

*2) **Comment Counts**:* To investigate the variation in comment counts for deleted and undeleted videos over time taking into account the potential survival bias, we adopted the same

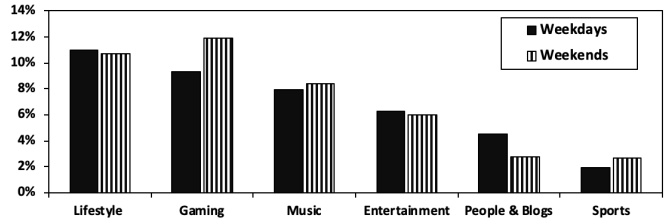

Fig. 3: Distribution of video categories in our dataset that were posted on weekdays vs. weekends.

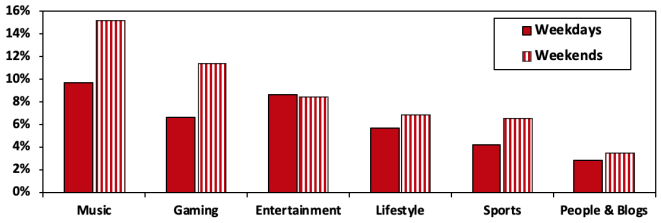

Fig. 4: Distribution of categories of videos in the deleted dataset that were posted on weekdays vs. weekends.

methodology described in Section IV-C1 where we computed the average comment counts per hour per snapshot for deleted and undeleted videos We see a similar pattern in average comment counts between the deleted and the undeleted videos over successive snapshots. For both deleted and undeleted videos, most of the comments are posted in the beginning and as time goes by, comment rates decrease, and the average comment counts itself is higher for the undeleted videos than the deleted videos at any point in time.

*3) **Comment Lengths**:* To deepen our understanding of the differences between the deleted/undeleted videos, we looked at their comment lengths (number of characters per comment). We found that comment lengths differ significantly between deleted and undeleted videos (Mann-Whitney U = 25x10^7, $p < .05$). The average length of comments for deleted videos is 58.15 (std = 93, max = 1,945), while it is 68.52 (std = 129, max = 8,617) for undeleted videos. In other words, deleted videos tend to receive shorter comments than undeleted ones.

E. Temporal Analysis

Our data collection took place during both weekdays and weekends. Thus, we investigate the differences between videos posted over weekdays versus the videos posted over weekends.

Given that gaming is the second most deleted category, we hypothesized that videos posted on weekends would have a higher proportion of deletion than videos posted on weekdays as young adults interested in gaming usually have more time on weekends to spend on games and social media. Although 50% of the videos in our dataset were posted during weekdays and the other 50% were posted on weekends, the percentage of deletion was indeed higher (20%) for videos posted on weekends than videos posted on weekdays (15%).

Next, we looked at the video categories of the videos posted on weekdays and weekends. Figure 3 shows the distribution of the most common video categories in our dataset for videos posted on weekdays and weekends. We found that gaming and sports videos tend to be posted more on weekends than weekdays, whereas people & blogs videos tend to be posted more on weekdays than weekends. As for lifestyle and entertainment videos, there was no significant difference in the number of videos posted on weekdays and those posted on weekends. The difference in distribution between categories on weekdays and weekends was found to be statistically significant (χ^2 = 880, df= 16, $p < .001$).

We then looked at the distribution of deleted videos per category that were posted on weekdays and weekends (See Figure 4). The difference in distribution between deleted videos categories over weekdays and weekends is quite significant with higher percentage of deleted videos on weekends than the weekdays for nearly all categories. Indeed, this difference is statistically significant (χ^2 = 186, df= 16, $p < .001$).

The main takeaway from these observations about video postings/deletion over weekdays and weekends is that videos posted on weekends are more likely to get deleted than videos posted on weekdays. Of all the deleted videos, about 57% were posted on weekends, while 43% were posted on weekdays.

V. Predicting Deleted Videos

Based on all our observations about deleted videos, we have built two classifiers to predict videos that get deleted on YouTube. The main motivation for building such classifiers is to develop a tool that would warn users about the possibility of a video that they post getting deleted in future. We first build a general prediction model to predict deletion using all available metadata after a video had been posted on YouTube and learn about audience engagement (i.e., views, likes, comments). Since, we would like to warn users as early as possible, we next investigate how accurate it would be to predict deletion even at the time the users click on the publish button.

A. Methods

We trained a random forest classifier based on extracted features from the videos in our dataset. We used 80% of the videos in our dataset to train the classifier, and the other unseen 20% for testing the classifier performance.

*1) **Feature Extraction**:* We identified three sets of features: video-related, comments-related, and text-related features (See Table III). For audience engagement features, we considered the last record in the video database that reflects the updated video statistics (e.g., view counts, comments) and then for each feature, we calculated the average per hour for each video.

For comment-related features, comment sentiments were obtained using VADER [29]. Text-related features, such as video title, description, and comments, were preprocessed by removing punctuation, numbers, white spaces, and English stop words and converting all letters to lower case. After prepossessing, word n-gram features (n = 1,2) with term frequency-inverse document frequency (TF-IDF) weights were

TABLE III: Summary of features used for training the classification model. (* Represents audience engagement features).

Feature	Description
Video-related features:	
VideoAvgViews*	The average # of views per hour
VideoAvgLikes*	The average # of likes per hour
VideoAvgDislikes*	The average # of dislikes per hour
VideoAvgFavorites*	The average # of favorites per hour
VideoTagsCount	The # of tags
VideoDuration	The length of the video in seconds
VideoTitleLen	The # of characters in video title
VideoCategory	The predefined video category
VideoDefaultLanguage	The language used for the video title & description
VideoDefinition	Indicates whether the video is available in hd or sd
VideoLicensedContent	Indicates whether the video represents licensed content
VideoLiveBroadcast	Indicates whether the video is an upcoming/active live broadcast
VideoDimension	Indicates whether the video is available in 3D or in 2D.
Weekend	Indicates whether the video was uploaded on weekend or weekday
Comments-related features:	
CommentSentiment*	A sentiment score from -1 to +1
CommentsAvg*	The average video comments received per hour
Text-related features:	
VideoTitle	Represented by word n-grams (n = 1,2) weighted by TF-IDF
VideoDescription	Represented by word n-grams (n = 1,2) weighted by TF-IDF
VideoTags	Represented by word n-grams (n = 1,2) weighted by TF-IDF
VideoComments*	Represented by word n-grams (n = 1,2) weighted by TF-IDF

TABLE IV: Evaluation results of classification model for predicting video deletion after posting a video.

Features	F_1	Precision	Recall	AUC	Accuracy
Video related	0.718	0.891	0.601	0.793	0.918
+ Comments related	0.737	0.915	0.617	0.802	0.921
+ Text related	0.764	0.952	0.637	0.815	0.933

extracted. To reduce the high dimensionality of n-gram features, only the top 30 features ordered by TF-IDF weights were considered. We trained the classifier on successive sets of features presented in Table III to explore prediction performance based on each set of features and to assess whether or not using more features would improve prediction performance.

B. Results

Here we present the results of the two classification models that we developed to predict video deletion: pre-posting and post-posting classifiers. The main difference between the two is that the pre-posting classifier doesn't rely on audience engagement features that are captured after posting a video.

*1) **Deletion Prediction After Posting a Video**:* Table IV compares the classifier performance based on successive sets of features used to train the classifier. Training the classifier only on video-related features resulted in an F_1 score of 0.72 and the most informative features for this classifier are avg. views, video duration, video title length, and avg. video likes. Training the classifier further on comment-related features, improved performance by about two points in terms of F_1 score. Most important features in this model are video duration, avg. views, video title length, avg. comments, and avg. video likes, respectively. Finally, the best prediction performance was achieved after combining all previous features with text-related features. This model achieved a 0.76 F_1 score in predicting video deletion. The most informative features for this classifier are avg. video views, video duration, video title length, avg. video likes, and video title, respectively.

*2) **Model Evaluation**:* Here we compare the performance of our best prediction model with the performances of four baseline models that we created using four different classification methods. To our knowledge, there hasn't been any prior work on predicting content deletion on YouTube that we could use to compare our prediction performance with. We refrained from direct performance comparisons with prediction models developed for predicting content deletion on other platforms such as Twitter [23] and Instagram [17], as the experimental setup may be fundamentally different. We followed [30] in creating two baseline models: 1) majority vote model, in which the classifier simply predicts the majority class (undeleted) for all examples in the test dataset; and 2) random vote model, in which the classifier randomly predicts one of the two classes for each video in the test dataset. The other two baseline models were developed using a lexicon-based approach [31] in which we use one of two feature selection methods, chi-square (χ^2) and Pointwise Mutual Information (PMI) to create two lexicons consisting of terms found in the video title along with weights representing their association strength toward one of the classes. Terms with positive weights indicate an association with the deleted class, whereas ones with negative weights indicate an association with the undeleted class. To classify new unseen examples of videos, we sum up the weights of the title's terms that exist in the lexicon. If the summation result is positive, the video would be classified as deleted, otherwise undeleted. We term these two baseline models as Lexicon-PMI baseline model and Lexicon-Chi baseline model, respectively.

Table V compares the performances of the four baseline models against our proposed deletion prediction model. We found that using a lexicon-based approach performed much better than the random vote baseline model with respect to all metrics, which is somewhat expected given the naive nature of the random vote classifier. Within the lexicon-based approach, Lexicon-Chi achieved better results than Lexicon-PMI in all metrics except for the recall. Lexicon-PMI has the best recall among all classifiers, but it achieved the second poorest precision among all models. In contrast, our proposed deletion prediction model outperformed all baseline models by a large margin in terms of F_1, precision, and accuracy.

Although our model achieved the second best recall among all baseline models, it achieved the best precision. While there are some situations where recall may be more important than precision (e.g., predicting cancer), we believe that in our case (predicting video deletion), precision is more important than recall. This is because false positive cases may have a negative impact on uploaders [20, 32], whereas missing true positive cases would still go through other filtering mechanisms such as flagging problematic content by YouTube users.

*3) **Deletion Prediction at the Time of Posting a Video**:* We now investigate video deletion based on the metadata available at the time of posting a video. This is important because such a prediction would allow us to build a tool to warn users ahead of time the likelihood of their videos being deleted after posting. To do this, we trained a random forest classifier based on the all features presented in Table III, except audience engagement features. Table VI compares the classifier performance based on training it on successive

TABLE V: Performance metrics for the evaluated baselines and for our proposed post-posting prediction model.

Model	F_1	Precision	Recall	Accuracy
Majority Vote	-	-	-	0.826
Random Vote	0.253	0.171	0.494	0.497
Lexicon-PMI	0.458	0.332	**0.739**	0.703
Lexicon-Chi	0.538	0.491	0.594	0.826
Proposed Model	**0.764**	**0.952**	0.637	**0.933**

TABLE VI: Evaluation results of classification model for predicting video deletion at the time of posting a video.

Features	F_1	Precision	Recall	AUC	Accuracy
Video related	0.639	0.798	0.533	0.752	0.894
+ Text related	0.721	0.898	0.602	0.794	0.921

sets of features. The best performance was achieved (0.72 F_1) when training the classifier on video, and textual features combined. The most informative features are video duration, video title length, video title, and video tags count. This is quite a revelation that it is possible to predict video deletion with an accuracy of 92% at an early stage even at the time of posting a video.

VI. Discussion and Conclusion

This paper presents insights into the phenomenon of video deletion on YouTube. Our analysis reveals that video deletion is not a rare event for uploaders, with 17.3% of the collected videos being deleted/removed. We identified several features that differentiate deleted videos from undeleted ones: (1) Sports, music, and entertainment videos are more likely to get deleted than religion, food, and pets and animal videos; (2) Deleted videos receive less audience engagement than undeleted ones; (3) Disabling comments is a more frequent behavior among deleted videos than undeleted ones; (4) Comments for the deleted videos tend to be shorter in length than those that are for undeleted videos; and (5) Finally, deleted videos tend to be longer in content than undeleted ones.

This paper also presents classifiers that can predict video deletion with a high degree of accuracy (92%) at an early stage, at the time of publishing a video. By adapting this early deletion prediction strategy, YouTube can reduce any frustration that may result from taking late action toward videos found in violation of YouTube's guidelines. Further, it can be used to warn users about the possibility of their videos being deleted at the time or soon after they post their video.

Prior research has studied content deletion on other social media platforms, and their results are likely to apply to YouTube. For example, Almuhimedi et al. [15] have found that tweet deletion rate was higher on weekends than weekdays, which was found to be true for YouTube as well. We found that videos uploaded on weekends got deleted more often than those that were uploaded on weekdays. Petrovic et al. [30] found that tweets containing swear words are more likely to be deleted. We found that deleted video comments have a higher percentage of negative sentiments, mostly due to swearing and negative words, than undeleted video comments. The findings of Figueiredo et al. [33] suggest that deleted videos get most of their views earlier in their lifetime. We found a similar pattern with respect to comments where deleted videos received 75% of their comments during the first 24 hours.

Our work has important implications for platform designers. The current strategy adopted by YouTube is that it depends on both humans and automated detection models to flag and take down inappropriate videos after these videos go alive for awhile [34]. We showed that it is possible to predict video deletion even at the time of publishing a video, which can reduce the community's exposure to likely disturbing videos. This early deletion prediction strategy can also have a positive impact on users psychological and emotional well-being as it has been shown that there is a link between user's perceptions of fairness and post removal anticipation [11].

Limitations and Future Work. As is the case for many studies that involve collecting data from social media, our data sample is limited by the collection method provided by YouTube. YouTube API quotas restrictions doesn't allow for collecting a large representative sample of data and only videos that were available within the US were considered in this study. Although visual features, e.g. the thumbnails of the videos, have been shown to be effective in other prediction tasks on YouTube [25], we refrained from using them for ethical reasons described in III-B. Finally, it could be that some of the videos in the undeleted dataset got deleted after the one week tracking period of this study. However, our analysis shows that within the first three days, 84% of the videos were deleted soon after they were posted, while only 2.7% of them were deleted in the last day of the one-week tracking period.

While this paper presents useful insights into video deletion and prediction, it leaves many analysis opportunities for other researchers to explore. For example, it will be valuable to investigate the reasons behind users deleting videos and whether or not regret plays a role in that, as regret has been identified as a deletion reason for online posts [13]. Another important future direction is to understand the characteristics of videos removed by YouTube and those deleted by the user. Understanding of such characteristics would help in building more specific prediction models of unavailable YouTube videos that distinguish between videos that may get deleted by the user or by YouTube, which would then allow the system to provide an explanation of the guidelines that the videos expected to be removed by YouTube violate. Educating the online community users about their mistakes rather than removing their posts without providing an explanation have been found to be effective for better understanding of the community norms and reducing future removal [11].

References

[1] J. Preece and D. Maloney-Krichmar, "Online communities: focusing on sociability and usability," *Handbook of human-computer interaction*, pp. 596–620, 2003.

[2] R. L. Williams and J. Cothrel, "Four smart ways to run online communities," *MIT Sloan Management Review*, vol. 41, no. 4, p. 81, 2000.

[3] Twitter, "Evolving our twitter transparency report: expanded data and insights," shorturl.at/ijQX1, 2018.

[4] M. Bickert, "Publishing our internal enforcement guidelines and expanding our appeals process," shorturl.at/fhiZ0, 2018.

[5] D. Soni and V. K. Singh, "See no evil, hear no evil: Audio-visual-textual cyberbullying detection," *CSCW*, vol. 2, no. CSCW, p. 164, 2018.

[6] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," in *ICWSM*. IEEE, 2017.

[7] N. Albadi, M. Kurdi, and S. Mishra, "Investigating the effect of combining gru neural networks with handcrafted features for religious hatred detection on arabic twitter space," *Social Network Analysis and Mining*, vol. 9, no. 1, p. 41, 2019.

[8] ——, "Are they our brothers? analysis and detection of religious hate speech in the arabic twittersphere," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2018, pp. 69–76.

[9] ——, "Hateful people or hateful bots? detection and characterization of bots spreading religious hatred in arabic social media," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–25, 2019.

[10] S. Myers West, "Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms," *New Media & Society*, vol. 20, no. 11, pp. 4366–4383, 2018.

[11] S. Jhaver, D. S. Appling, E. Gilbert, and A. Bruckman, "Did you suspect the post would be removed?: Understanding user reactions to content removals on reddit," *CSCW*, vol. 3, no. CSCW, p. 192, 2019.

[12] J.-M. Xu, B. Burchfiel, X. Zhu, and A. Bellmore, "An examination of regret in bullying tweets," in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013, pp. 697–702.

[13] Y. Wang, G. Norcie, S. Komanduri, A. Acquisti, P. G. Leon, and L. F. Cranor, "I regretted the minute i pressed share: A qualitative study of regrets on facebook," in *Proceedings of the seventh symposium on usable privacy and security*. ACM, 2011, p. 10.

[14] K. B. Srinivasan, C. Danescu-Niculescu-Mizil, L. Lee, and C. Tan, "Content removal as a moderation strategy: Compliance and other outcomes in the changemyview community," *CSCW*, vol. 3, no. CSCW, p. 163, 2019.

[15] H. Almuhimedi, S. Wilson, B. Liu, and A. Sadeh, "Tweets are forever: a large-scale quantitative analysis of deleted tweets," in *CSCW*. ACM, 2013, pp. 897–908.

[16] P. Bhattacharya and N. Ganguly, "Characterizing deleted tweets and their authors," in *ICWSM*, 2016.

[17] S. Chancellor, Z. J. Lin, and M. De Choudhury, "This post will just get taken down: characterizing removed pro-eating disorder social media content," in *CHI*, 2016.

[18] Google, "Youtube community guidelines enforcement," shorturl.at/ixU89, 2019.

[19] J. Dry, "Youtube creators cry censorship as 'inappropriate' content is no longer monetizable on the platform," shorturl.at/gkGH8, 2016.

[20] D. Coldewey and T. Hatmaker, "Police say shooter's anger over youtube policies appears to be the motive," shorturl.at/LQV08, 2018.

[21] J. Crossfield, "The hidden consequences of moderating social media's dark side," shorturl.at/twZ16, 2019.

[22] Z. Tufekci, "Facebook, youth and privacy in networked publics," in *ICWSM*, 2012.

[23] L. Zhou, W. Wang, and K. Chen, "Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones," in *WWW*, 2016, pp. 603–612.

[24] T. Trzciński and P. Rokita, "Predicting popularity of online videos using support vector regression," *IEEE Transactions on Multimedia*, vol. 19, no. 11, pp. 2561–2570, 2017.

[25] K. Papadamou, A. Papasavva, S. Zannettou, J. Blackburn, N. Kourtellis, I. Leontiadis, G. Stringhini, and M. Sirivianos, "Disturbed youtube for kids: Characterizing and detecting inappropriate videos targeting young children."

[26] R. Ottoni, E. Cunha, G. Magno, P. Bernardina, W. Meira Jr, and V. Almeida, "Analyzing right-wing youtube channels: Hate, violence and discrimination," in *WEBSCI*, 2018, pp. 323–332.

[27] P. Schultes, V. Dorner, and F. Lehner, "Leave a comment! an in-depth analysis of user comments on youtube." *Wirtschaftsinformatik*, vol. 42, pp. 659–673, 2013.

[28] J. B. Walther, D. DeAndrea, J. Kim, and J. C. Anthony, "The influence of online comments on perceptions of antimarijuana public service announcements on youtube," *Human Communication Research*, vol. 36, no. 4, pp. 469–492, 2010.

[29] C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *ICWSM*, 2014.

[30] S. Petrovic, M. Osborne, and V. Lavrenko, "I wish i didn't say that! analyzing and predicting deleted messages in twitter," *arXiv preprint arXiv:1305.3107*, 2013.

[31] N. Kaji and M. Kitsuregawa, "Building lexicon for sentiment analysis from massive collection of html documents," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, 2007.

[32] E. Rosenfeld, "Shooter allegedly targeted youtube hq because she 'hated' the company for blocking her videos," shorturl.at/xzNT2, 2018.

[33] F. Figueiredo, F. Benevenuto, and J. M. Almeida, "The tube over time: characterizing popularity growth of youtube videos," in *WSDM*, 2011, pp. 745–754.

[34] Google, "Report inappropriate content," shorturl.at/isxEM, 2019.