2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Crowdsourcing-based Copyright Infringement Detection in Live Video Streams

Daniel (Yue) Zhang, Qi Li, Herman Tong, Jose Badilla, Yang Zhang, Dong Wang
Department of Computer Science and Engineering
University of Notre Dame
Notre Dame, IN, USA
{yzhang40, qli8, ktong1, jbadilla, yzhang42, dwang5}@nd.edu

Abstract*—With the increasing popularity of online video sharing platforms (such as YouTube and Twitch), the detection of content that infringes copyright has emerged as a new critical problem in online social media. In contrast to the traditional copyright detection problem that studies the static content (e.g., music, films, digital documents), this paper focuses on a much more challenging problem: one in which the content of interest is from live videos. We found that the state-of-the-art commercial copyright infringement detection systems, such as the ContentID from YouTube, did not solve this problem well: large amounts of copyright-infringing videos bypass the detector while many legal videos are taken down by mistake. In addressing the copyright infringement detection problem for live videos, we identify several critical challenges: i) live streams are generated in real-time and the original copyright content from the owner may not be accessible; ii) streamers are getting more and more sophisticated in bypassing the copyright detection system (e.g., by modifying the title, tweaking the presentation of the video); iii) similar video descriptions and visual contents make it difficult to distinguish between legal streams and copyright-infringing ones. In this paper, we develop a crowdsourcing-based copyright infringement detection (CCID) scheme to address the above challenges by exploring a rich set of valuable clues from live chat messages. We evaluate CCID on two real world live video datasets collected from YouTube. The results show our scheme is significantly more effective and efficient than ContentID in detecting copyright-infringing live videos on YouTube.

I. Introduction

It has been a recent phenomenon that online social media (e.g., YouTube and Twitch) allow users to broadcast live videos to audience worldwide [1]. These video sharing platforms are fundamentally different from "static" content distribution platforms (e.g., Netflix and Hulu) because streams are generated and consumed in real-time. The live videos (e.g., real-time game play, live TV shows) create great revenues for both live stream uploaders (often referred to as "streamers") and the video sharing platforms. For example, according to a recent survey [2], the live video market is estimated to grow from 30.29 billion US dollars in 2016 to more than 70 billion US dollars by 2021. With such incentives, YouTube has attracted over 76,000 active streamers in March 2017 alone and has a projected growth of 330% new active streamers per month.

This prevalence of live stream platforms also opens the door for severe copyright infringement issues where users can stream copyrighted live events (e.g., TV shows, sport matches, Pay-per-view programs) without the permission of content owners. For example, a major anti-piracy agency claims 77 million people watched "Game Of Thrones" season 7 episode 1 via unauthorized live videos, causing an estimated total of 45 million US dollars of revenue loss to HBO, the legal copyright owner of the series [3]. One of the main reasons for such serious copyright infringement is the grass-root nature of the video sharing platforms: anyone can start a live video stream on the platform without going through a rigorous copyright screening process. This leaves room for "rogue accounts" to host illegal live streams.

Due to the increasing demand of blocking unauthorized video streams from copyright owners, the video sharing platforms have spent a significant amount of efforts addressing the copyright infringement problem. One of the most representative copyright detection tools for live videos is ContentID [4], a proprietary system developed by YouTube to detect copyright-infringing video uploads. In ContentID, each uploaded video stream is compared against a database of files provided by content owners to check for copyright issues. ContentID also uses the self-reports from content owners when they identify pirated videos. Unfortunately, ContentID has received heated criticisms from both video streamers and copyright owners due to its high false positives (i.e., falsely taking down legal streams) [1] and false negatives (i.e., constantly miss copyright-infringing videos) [2]. In fact, our empirical study showed that the ContentID failed to catch 26% of copyrighted videos after they have been broadcast for 30 minutes and shut down 22% video streams that are not copyright-infringing.

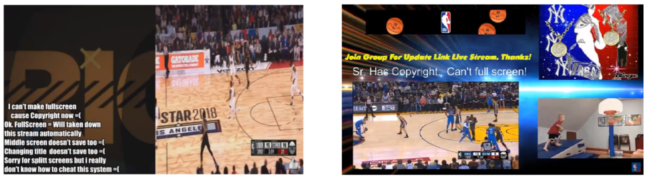
Several alternative copyright protection techniques (e.g., digital fingerprinting and forensic watermarking) can help track down pirated content effectively. However, such solutions require the original copy of the copyrighted content in advance to extract unique video features or embedded identifiers (e.g., digital watermarks or serial numbers) for tracking. Therefore, they are often applied on static content (e.g., eBooks, music and films) and are not suitable for live videos that are generated in *real-time*. Several tools have been developed to detect the copyright-infringing content by examining the video content (referred to as "video copy detectors") [5]. However,

[1] https://www.cbronline.com/news/youtube-blocked-vieos

[2] https://mashable.com/2017/08/16/youtube-live-streaming-copyrights/

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00 © 2018 IEEE

they cannot be applied to our problem because many streamers are sophisticated enough to change the video presentation and make it look very different from the original one (See Figure 1). Therefore, a system that can effectively address the copyright detection problem of live video streams has yet to be developed.


I can't make fullscreen
cause Copyright now =(
Ok. FullScreen = Will taken down
this stream automatically
Middle screen doesn't save too =(
Changing title doesn't save too =(
Sorry for splitt screens but i really
don't know how to cheat this system =(

Join Group For Update Link Live Stream. Thanks!
Sr. Has Copyright. Can't full screen!

(a) Video with Split Screen (b) Video with Camouflage

Figure 1: Videos modified by sophisticated streamers that successfully bypassed the ContentID system from YouTube

In this paper, we develop a novel Crowdsourcing-based Copyright Infringement Detection (CCID) scheme to capture copyright-infringing live videos. Our solution is motivated by the observation that the live chat messages from the online audience of a video could reveal important information of copyright infringement. For example, consider detecting copyrighted streams of a NBA match. If the audience of the live video are chatting about the current game status (e.g., "nice 3-pointer from Player A"), this video stream is likely to be copyright-infringing because the audience will only know these details of the game if the broadcast is real. Alternatively, if a video stream is copyright-infringing, the audience sometimes colludes with the streamers by reminding them to change the title of the stream to bypass the platform's detection system. However, such colluding behavior actually serves as a "signal" that the stream has copyright issues. In this paper, the CCID designs a novel detection system that explores the "clues" extracted from both live chats of the audience and the meta-data of the videos (e.g., view counts, number of likes/dislikes). It develops a *supervised learning* scheme to effectively track down copyright infringement in live video streams.

To the best of our knowledge, the CCID scheme is the first crowdsourcing-based solution to address the copyright infringement issues for live videos in online social media. It is robust against sophisticated streamers who can intentionally modify the description and presentation of the video, because CCID does not rely on the analysis of the actual content of the videos. Additionally, CCID performs the detection task on-the-fly without accessing the original copyrighted content. We evaluate the performance of CCID on two live stream video datasets collected from YouTube. The results show that our scheme is more accurate (achieving 17% higher in F1-Score) and efficient (detecting 20% more copyright-infringing videos within 5 minutes after the videos start) than the ContentID tool from YouTube.

II. Related Work

A. Copyright Protection

Due to the increasing popularity of online data sharing platform, protecting copyrighted content has become a critical problem in recent years [6]. Various techniques have been proposed to protect copyrighted music, text documents and videos. For example, Podilchuk *et al.* developed a robust watermarking technique that can covertly embed owner information into a digital image without affecting the perceived visual quality of the original content [7]. Low *et al.* proposed a data hiding technique to protect copyrighted text documents by slightly shifting certain text lines and words from their original positions to create unique identifiers for the original content [8]. Waldfogel *et al.* developed a music copyright protection scheme based on the observation that the original music often has superior quality than the unauthorized copies [9]. However, these techniques focus on the static contents and cannot be applied to live video streams where contents are generated in real-time.

B. Video Copy Detection

Video copy detection is one of the most commonly used techniques for detecting copyright infringement in video content. For example, Esmaeili *et al.* proposed a video copy detection system that compares the fingerprints (unique features extracted from the copyrighted content) of different videos to detect the copyright issues [10]. Nie *et al.* developed a near-duplicate video detection framework by combining comprehensive image features using a tensor model [5]. Chou *et al.* proposed a spatial-temporal pattern based framework for efficient and effective detection of duplicate videos [11]. However, these methods all require access to the original copy of the video in advance which is not practical in live video streams. More importantly, these content-based methods often fail when streamers are sophisticated enough to tweak the video presentations to bypass the detection system. In contrast, our scheme relies on the chat messages from the audience and video meta-data, which is independent of the video content.

C. Crowdsourcing in Online Social Media

Crowdsourcing-based techniques have been widely used in the analysis of online social media data [12], [13], [14], [15]. For example, Wang *et al.* developed a principled estimation framework that identifies credible information during disaster events by taking Twitter users as crowd sensors [16], [17], [18]. Zhang *et al.* applied machine learning based approaches to further study the profiles, sentiments, opinions, and dependencies of the crowdsourcing users on social media [19], [20], [21], [22], [23]. Steiner *et al.* developed a generic crowdsourcing video annotation framework that invites the users on YouTube to annotate the type of events and named entities of the videos they viewed [24]. Our work is different from the above schemes in the sense that it is the first crowdsourcing-based approach to address the *copyright infringement detection problem of live videos* on online social media.

III. Problem Statement

In this section, we present the copyright infringement detection problem of live video streams. In particular, we assume that a video hosting service has a set of live videos related to a piece of copyrighted content $y, 1 \leq y \leq Y$: $V(y) = V_1^y, V_2^y ... V_{N(y)}^y$ where $N(y)$ denotes the total number of live videos related to y. A video V_i^y is associated with a tuple, i.e., $V_i^y = (Meta_i^y, Chat_i^y, z_i^y)$ where $Meta_i^y$ is the metadata of the video (e.g., description, view count, likes/dislikes). $Chat_i^y$ is the live chat messages for the video. z_i^y is the ground truth label defined below:

- Copyright-Infringing (labeled as "True"): live videos that contain actual copyrighted content (e.g., a live broadcasting of a football game; a live stream of latest episode of "Game of Thrones".).
- Non-Copyright-Infringing (labeled as "False"): videos that do not contain actual live copyrighted content.

An example of the above definitions is shown in Figure 2. We observe that all four pictures are related to a copyrighted NBA game and claimed to broadcast the live events for free. However, only the last one (bottom-right) should actually be labeled as "True" (i.e., copyright-infringing) and the others should be labeled as "False". For example, the top-left video is a game-play video of an NBA 2K game. The top-right one is just a static image and the bottom-left one is broadcasting a recorded match.

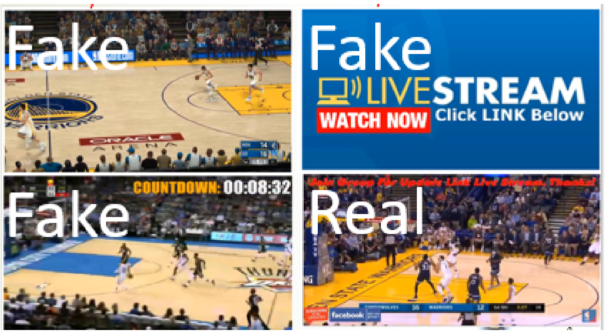


Figure 2: Live Videos on YouTube

We make the following assumptions in our model.

Real-time Content: the content of the copyrighted material is assumed to be generated in real-time and the content of the video cannot be acquired in advance.

Sophisticated Streamers: we assume streamers are sophisticated and they can manipulate the video descriptions and content to bypass the copyright infringement detection system.

Colluding Audience: we assume some of the audience can collude with streamers by reminding them of ways to cheat the copyright infringement detection system (e.g., change the title).

Given the above definitions and assumptions, the goal of copyright right infringement detection is to classify each live video stream into one of the two categories (i.e., copyright-infringing or not) by leveraging the live chat messages and the metadata of the videos. Formally, for every piece of copyrighted content $y, 1 \leq y \leq Y$, find:

$$\arg\max_{\tilde{z}_i^y} Pr(\tilde{z}_i^y = z_i^y | Meta_i^y, Chat_i^y), \ \forall 1 \leq i \leq N(y) \quad (1)$$

where $\tilde{z}_i^y$ denotes the estimated category label for V_i^y.

IV. Approach

In this section, we present the CCID framework to address the copyright infringement problem for live videos. It consists of four major components: i) a data collection component to obtain the live videos from online social media (e.g., YouTube); ii) a live chat feature extraction component to obtain the features from the live chats to which the audience of the video contributes; iii) a metadata feature extraction component to extract features from the descriptive information of each video; iv) a supervised classification component to decide if the video stream is copyright-infringing or not. We discuss each component in details below.

A. Obtaining Live Video Datasets

The data collection is challenging because: i) no existing live streaming video dataset is publicly available with chat content and ground truth labels; ii) many live streams about the same event are simultaneously broadcast (e.g., sport games, TV shows), which requires a scalable design for data collection. In light of these challenges, we developed a distributed live stream crawling system using Selenium [3] and Docker [4]. The system is deployed on 5 virtual machines hosted on Amazon Web Service. The crawling system collects the following items of a live video stream:

Video Metadata: The metadata of the video includes video title, video description, streamer id, view count, and the number of likes and dislikes.

Live Screen Shots: The real-time screenshots of the live video that are captured every 30 seconds.

Live Chat Messages: The real-time chat messages from the audience about the video.

Terminology Dictionary: A dictionary related to a piece of copyrighted content y. Examples of the terms in the dictionary include the names of the main characters in a TV show, the names of players and terminologies used in a sport event.

We refer to our system implementation [25] for more details.

B. Live Chat Feature Extraction with Truth Analysis

The goal of the live chat feature extraction component is to identify the key features from the audience's chat messages that are relevant to the copyright infringement of a live video. We observe that the chat messages often reveal important "clues" that help make inferences with respect to the copyright infringement of a video. For example, many viewers often complain about the quality of the video (e.g., resolution, sound quality) for a live stream that is copyright-infringing. Alternatively, the viewers are disappointed (e.g., by posting

[3] https://www.seleniumhq.org/

[4] https://www.docker.com/

Table I: Examples of Crowd Votes

Types of Crowd Votes	Example Chat Messages
Colluding Vote	edgar dlc: change title please so nba wont copyright Johnson's Baby Oil: Change the name when it starts
Content Relevance Vote	Joshua Colinet: As a lakers fan, I'm hoping the new look cavs flop so we can get a higher draft pick Moustache Man11: Can someone tell me who scored
Video Quality Vote	KING BAMA: Looks good but can we please get some sound buddy?!! Malik Horton: would be alot more fun too watch if it wasn't laggy
Negativity Vote	PHXmove: FAKE DO NOT BOTHER DaVaughn Sneed: They keep putting bullsh*t up I'm just trying to watch this game

cursing and negative comments in the chat messages) if the content of the video is actually fake. In the CCID scheme, we define these messages relevant to copyright infringement as a set of *crowd votes*.

DEFINITION 1. Crowd Votes: a crowd vote is a chat message that suggests whether a video is copyright-infringing or not. The vote reflects the viewer's observation about the "truthfulness" (copyright infringement) of the video. More specifically, we define four types of crowd votes:

- *Colluding Vote:* a live chat message from the audience to help the streamer bypass the copyright infringement detection system of the video platform. Examples of the colluding vote include chat messages that contain the keywords such as "change the title," "change the name," "change description."
- *Content Relevance Vote:* a live chat message that contains keywords that are directly relevant to the event. For example, the names of players in an NBA game, the names of main characters in a TV show. The relevance vote of a message is derived based on the overlap between the message and terms in the Terminology Dictionary described above.
- *Video Quality Vote:* a live chat message that contains keywords about the quality of the video (e.g., "lag," "resolution," "full screen," "no sound"). Normally, the more people care about the video quality, the more likely the video contains the real copyrighted content.
- *Negativity Vote:* a live chat message that contains direct debunking of the content (e.g., "fake, dislike, down vote, quit, go to my stream instead") and a set of swear words that express anger towards the streamer [5].

Table I shows a few examples of different types of crowd votes from our collected video datasets.

To better quantify the contribution of a crowd vote to the likelihood of a video being copyright-infringing, we further define the weight of a crowd vote as follows.

DEFINITION 2. Weight of Crowd Vote: the weight of a crowd vote is defined as the probability that a video is copyright-infringing given the type of the crowd vote about the video. Formally, it is defined as:

$$\phi_{i,k} = Pr(V_i = T | SV_k) \tag{2}$$

where $V_i = T$ denotes that the video V_i is copyright-infringing and SV_k denotes the crowd vote is of type k [6]. For the ease of notation, we omit the superscript of copyrighted content (i.e., y) in all equations in this section.

In the CCID scheme, we develop a principled model to compute the weight of each crowd vote using a Maximum Likelihood Estimation (MLE) approach. This approach is inspired by the recent development of the truth analysis technique [16], [26], [27] that jointly estimates the quality of sources (voters) as well as the label of the items (videos) being voted on. We first define a few important notations. Let $a_{i,k} = Pr(SV_k | V_i = T)$ denote the probability that a crowd vote of type SV_k appears in a copyright-infringing video. It can be derived from $\phi_{i,k}$ using Bayes' theorem: $a_{i,k} = \frac{\phi_{i,k} \times Pr(SV_k)}{\pi_T}$, where π_T is the probability of a randomly selected video being copyright-infringing (i.e., $Pr(V_i = T)$). Similarly, we define $b_{i,k} = Pr(SV_k | V_i = F) = \frac{(1-\phi_{i,k}) \times Pr(SV_k)}{1-\pi_T}$. It represents the probability that SV_k appears in a video with no copyright infringement. We further define a helper function $\chi(c,k)$ which returns 1 if a chat message c is of type SV_k and 0 otherwise.

Given the above definitions, we derive the likelihood function of observed data X (i.e., the videos $\{V_1, V_2, ...V_N\}$ and its corresponding comments $\{Chat_1, Chat_2, ..., Chat_N\}$) as:

$$\begin{aligned}\mathcal{L}(\Theta|X) = \prod_{i=1}^{N} \Bigg\{ & \prod_{c \in Chat_i} \prod_{k=1}^{K} a_{i,k}^{\chi(c,k)} \times (1-a_{i,k})^{(1-\chi(c,k))} \\ & \times \pi_T \times z_i + \prod_{c \in Chat_i} \prod_{k=1}^{K} b_{i,k}^{\chi(c,k)} \times (1-b_{i,k})^{(1-\chi(c,k))} \\ & \times (1-\pi_T) \times (1-z_i) \Bigg\}\end{aligned} \tag{3}$$

where z_i is a binary variable indicating whether a video stream V_i is copyright-infringing ($z_i = 1$) or not ($z_i = 0$).

In the above equation, the estimation parameters Θ are $\pi_T, \phi_{i,1}, ...\phi_{i,K}\}$. They can be estimated by maximizing the likelihood of the observed data.

$$\underset{\{\pi_T, \phi_1^T, ...\phi_K^T\}}{\arg\max} \; \mathcal{L}(\Theta|X) \tag{4}$$

[5] http://www.bannedwordlist.com/lists/swearWords.txt

[6] Without loss of generality, we assume a set of K types of crowd votes in our model, i.e., $SV = \{SV_1, SV_2..SV_K\}$. In this paper, we focus on the four types as defined above (i.e., $K = 4$).

Using the Bayesian estimation [28], we can derive the closed-form solution to the above estimation problem as follows:

$$\pi_T = \frac{\sum_{i=1}^{N} z_i}{N}, \quad \phi_{i,k} = \frac{\sum_{c \in Chat_i} z_i \times \chi(c,k)}{\sum_{c \in Chat_i} \chi(c,k)} \tag{5}$$

where the value of z_i will be learned from the training data.

Using the weights of the crowd votes from the above estimation, we can define the **Overall Crowd Vote** ($Chat_{ocv}$) feature for each video V_i. This feature represents the aggregated observations from the audience on whether the video stream is copyright-infringing. Formally, $Chat_{ocv,i}$ is derived as the aggregated weights of all crowd votes about video V_i.

$$Chat_{ocv,i} = \sum_{c \in Chat_i} \sum_{k=1}^{K} \phi_{i,k} \times \chi(c,k) \tag{6}$$

In addition to the $Chat_{ocv}$ feature, we also investigate other live chat features that are potentially relevant to copyright infringement detection. We summarize these features below:

Chat Message Rate ($Chat_{rateM}$): the average number of chat messages per minute.

Chat User Rate ($Chat_{rateU}$): the average number of distinct chatting users per minute.

Early Chat Polarity ($Chat_{polarity}$): The average sentiment polarity of the chat messages posted during the starting stage of the event (i.e., 0-3 minutes). The polarity refers to how positive/negative the chat messages are. Normally, the audience starts to curse and posts negative comments of a video after they find the live stream to be fake, which usually happens at the beginning of the stream.

C. Metadata Feature Extraction

We found the metadata of a video also provides valuable clues for copyright infringement detection. In our CCID scheme, we focus on the following metadata features.

View Counts ($Meta_{view}$): The number of viewers that are currently watching the live video stream. Intuitively, the more viewers, the more likely the video is broadcasting copyright-infringing content.

Title Subjectivity ($Meta_{subT}$): The subjectivity of the video's title. We derive a subjectivity score (a floating point within the range [-1.0, 1.0]) of each video using the subjectivity analysis of TextBlob [29]. Intuitively, a title with high subjectivity (e.g., "Super Bowl live stream for free!", "Best Quality Ever!") can potentially be a spam (non-copyright-infringing) since a copyright-infringing video normally keeps an objective and low-profile title (e.g., "NFL Super Bowl LII ") to minimize the chance of being caught by the platform's copyright infringement detection system.

Description Subjectivity ($Meta_{subD}$): The subjectivity of the video's description. It is chosen based on the same intuition as the title subjectivity.

Number of Likes/Dislikes ($Meta_{like}$, $Meta_{dislike}$): The total number of viewers who hit the "like"/"dislike" button. Intuitively, if the video contains copyrighted content, it may receive more "likes" from the audience. However, if the audience finds out that the video stream does not contain copyrighted content, they are more likely to hit the "dislike" button.

Note that we chose not to use the word-level features directly related to the content of titles and descriptions of the videos (e.g., using text mining to extract topic and Bag-of-Words (BoW) features). This is because the sophisticated streamers often manipulate the title and descriptions to bypass the platform's copyright detection system. For example, in one of the copyright-infringing live streams of an NBA game, the streamer modified the title as "1000 iphones!!!!". On the other hand, many legal video streams (e.g., a live game play of an NBA 2K video game) actually have very suspicious titles such as "2018 NBA All-Star Game LIVE!!!" in order to attract attention of audience. BoW or topic based feature extraction techniques are often not robust against sophisticated streamers and can easily lead to a large amount of false alarms [30].

D. Supervised Classification

After the live chat and metadata features are extracted from the collected data, CCID performs supervised binary classification using the extracted features to classify live videos as copyright-infringing or not. Rather than re-inventing the wheel, we use a set of the state-of-the-art supervised classification models in the CCID scheme. Examples include neural networks, boosting models, tree based classifiers and a support vector machine. These classifiers serve as plug-ins to our CCID scheme and the one with the best performance from the evaluation on training data will be selected. We present the detailed performance evaluation of CCID when it is coupled with these classifiers in Section V.

V. Evaluation On Real World Data

In this section, we evaluate the CCID scheme using two real-world datasets collected from YouTube. The results demonstrate that CCID significantly outperforms Content ID from YouTube, the only available copyright infringement detection tool for live videos at the time of writing.

A. Datasets

We summarize the two real-world datasets used for evaluation in Table II. The NBA dataset contains 130 live video streams related to NBA games from Dec. 2017 to Mar. 2018. 28.46% of the collected videos are found to be copyright-infringing. The Soccer dataset contains 226 live videos related to soccer matches in major soccer leagues worldwide from Sept. 2017 to Mar. 2018. 17.70% of these videos are copyright-infringing. We use the data crawler system (described in Section IV) to collect these live videos. The search terms we used to collect these videos are team names of the match (e.g., "Houston Rockets Detroit Pistons"). We leverage the advanced search filters provided by YouTube to ensure all collected videos are live video streams. For each video, we collect the stream for a duration of 30 minutes and we start the crawling process at the scheduled time of each game.

Table II: Data Trace Statistics

Data Trace	NBA	Soccer
Collection Period	Dec. 2017 - March 2018	Sept. 2017 - Mar. 2018
Number of Videos	130	226
% of copyright-infringing Videos	28.46%	17.70%
Number of Chat Users	2,705	4,834
Number of Chat Messages	61,512	94,357

To obtain the ground truth labels, we manually looked at the collected screenshots of a video stream to check if the video is copyright-infringing. This labeling step is carried out by three independent graders to eliminate possible bias. We sort the video streams by their chronological order and use the first 70% of data as the training set and the last 30% (latest) as the test set. In the training phase, we perform 10-fold cross validation to tune the parameters of the classifiers.

To build the terminology database we crawled the names of players and teams from the ESPN website [7] for the NBA dataset. For the Soccer dataset, we use an existing database to extract the names of players and teams of major soccer clubs [8]. We also built a set of terminologies and slogans related to these events (e.g., flop, 3-pointer, foul, dribble, header, hat-trick).

B. Classifier Selection and Baseline

We chose a few state-of-the-art supervised classifiers that can be coupled with the CCID scheme in our experiments. We summarize them below.

- **AdaBoost, XGBoost, Random Forest (RF):** AdaBoost [31], XGBoost [32], and Random Forest [33] are ensemble-based classification algorithms that combine a set of classifiers (we use 50 decision trees) to improve classification performance.
- **Linear Support Vector Machine (SVM):** Given labeled training data, the SVM algorithm outputs an optimal hyperplane to categorize new data samples [34].
- **Multi-layer Perceptron (MLP):** An artificial neural network based classification scheme that can distinguish data that is not linearly separable [35].

We compare the CCID system with the current copyright detection system (i.e., ContentID) developed by YouTube [4]. To evaluate YouTube's ContentID without direct access to its internal system (since it is a proprietary system), we estimate the effectiveness of ContentID as follows. In We label a video as copyright-infringing (detected by ContentID) if it i) went offline abruptly during the broadcasting, or ii) it was explicitly reported by the copyright owner and taken down. We observe that the latter case is rare (less than 10%) in the live streams we collected. This again demonstrates the importance and necessity of developing an automatic detection system like CCID to keep track of copyright-infringing content in live videos from online social media.

C. Results: Detection Effectiveness

In the first set of experiments, we evaluate the detection effectiveness of CCID when it is coupled with different classifiers and identify the best performed classifier for CCID. We then compare CCID with ContentID used by YouTube. The detection effectiveness is evaluated using the classical metrics for binary classification: *Accuracy*, *Precision*, *Recall* and *F1-Score*. The results are reported in Table III.

We observe that AdaBoost achieves the best performance among all candidate classifiers. We also observe adding the features extracted from live chat messages can significantly improve the detection performance of CCID. More specifically, CCID with AdaBoost achieved 6.8% and 17.2% increase in F1-Score in the NBA and Soccer datasets, respectively, compared to YouTube's ContentID. In fact, we observe ContentID has poor precision in both datasets due to high false positive rates (which will be further discussed in the next subsection). The high false positive rate leads to the unfair taking down of legal live videos and discourages streamers from uploading live video contents. In contrast, the CCID scheme exploits the chat messages from the actual audience of the videos to identify potential evidence of copyright infringement, making it more robust to false alarms.

D. Results: Detection Time

We then evaluate the detection time of both CCID and ContentID. The detection time is defined as the amount of time the system takes to detect the copyright infringement of a live video after it starts. We focus on two aspects of the detection system when we study the detection time: i) True Positive Rate: it characterizes the ability of the system to correctly identify a copyright-infringing video. This metric is important for *copyright owners* who would like to detect all illegal video streams; ii) False Positive Rate: it characterizes the ability of the system to suppress the misclassified copyright-infringing videos. This is particularly important to "streamers" who would like to keep their legal content from being falsely taken down.

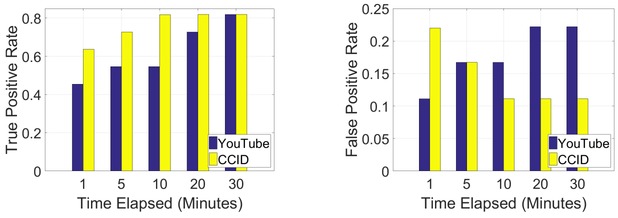
In the experiment, we tune the time window of the data collection from 1 to 30 minutes and only use the chat messages within the specified time window for CCID. We also chose the best-performed classifier (i.e., AdaBoost) for CCID. The results are shown in Figure 3 and Figure 4. We observe that, for the true positive rate, our scheme quickly outperforms YouTube at a very early stage of the event and keeps a consistently high performance for the rest of the event. Such results suggest our CCID scheme can catch copyright-infringing videos not only more accurately but also much faster than ContentID from YouTube. For the false positive rate, we observe that the CCID scheme has a higher false positive rate at the very beginning of the event (due to the lack of sufficient chat messages). However, our scheme quickly catches up and starts to outperform YouTube (ContentID) when the time window is longer than 5 minutes. We also observe that YouTube starts to mistakenly take down more and more legal videos (as copyright-infringing ones) as time elapses. Such increase can clearly discourage streamers with legal content from using the video sharing platform.

[7] http://www.espn.com/nba/players

[8] https://www.kaggle.com/artimous/complete-fifa-2017-player-dataset-global

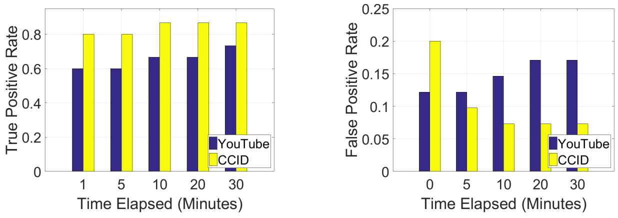
Table III: Classification Accuracy for All Schemes

		NBA				Soccer			
Algorithms	Features	Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
Adaboost (CCID)	**w/ chat features**	**0.8621**	**0.8182**	**0.8182**	**0.8182**	**0.9103**	**0.8125**	**0.8667**	**0.8387**
	w/o chat features	0.8276	0.7500	0.8182	0.7826	0.8750	0.7500	0.8000	0.7742
XGBoost	w chat features	0.7971	0.7777	0.6364	0.7000	0.8571	0.7059	0.8000	0.7500
	w/o chat features	0.7586	0.6667	0.7273	0.6957	0.8214	0.8571	0.4000	0.5455
RF	w/ chat features	0.7586	0.7000	0.6364	0.6667	0.8750	0.7857	0.7333	0.7586
	w/o chat features	0.6897	0.5714	0.7273	0.6400	0.8036	0.7000	0.4667	0.5600
SVM	w/ chat features	0.6207	0.5000	0.4545	0.4762	0.8214	0.6667	0.6667	0.6667
	w/o chat features	0.5862	0.4286	0.2728	0.3333	0.7679	0.6667	0.2667	0.3810
MLP	w/ chat features	0.6207	0.5000	0.4545	0.4762	0.7321	0.5000	0.5333	0.5161
	w/o chat features	0.4137	0.3636	0.7273	0.4848	0.6786	0.2000	0.0667	0.1000
YouTube (ContentID)		**0.7931**	**0.6923**	**0.8182**	**0.7500**	**0.8036**	**0.6111**	**0.7333**	**0.6667**



(a) True Positive Rate (b) False Positive Rate

Figure 3: NBA Dataset



(a) True Positive Rate (b) False Positive Rate

Figure 4: Soccer Dataset

E. Results: Feature Analysis

In addition to the holistic evaluation of CCID system, we also investigate which features are the most critical ones in our selected classifier. Table IV shows the ranking of features we used in the CCID scheme (with Adaboost) based on the *information gain ratio*, a commonly used metric in analyzing the feature importance for decision-tree based classifiers [31].

We found $Meta_{view}$, $Chat_{ocv}$ and $Meta_{subT}$ are the three most important features for both datasets. The first two features (i.e., $Meta_{view}$, $Chat_{ocv}$) are intuitive: the more viewers a video stream attracts, the more likely it is broadcasting copyrighted content (otherwise the viewers will simply quit and switch to another video stream). Similarly, crowd vote represents a strong signal from the audience to indicate if a video is copyright-infringing (via the overall crowd votes). For $Meta_{subT}$, we attribute it to the fact that a title with high subjectivity can potentially be spam (a copyright-infringing video normally keeps a low profile to minimize the chance of being caught.). We also observe that some intuitive features such as the number of likes/dislikes do not actually play a critical role in the classification process. This observation might be attributed to the fact that users may not even bother hitting the "like" or "dislike" button when they are watching the videos about live events (e.g., sports).

Table IV: Feature Importance for All Schemes

	NBA		Soccer	
Features	Ranking	Gain Ratio	Ranking	Gain Ratio
$\boldsymbol{Chat_{ocv}}$	**2**	**0.1568**	**2**	**0.2011**
$Chat_{rateM}$	5	0.1004	6 (tie)	0.0726
$Chat_{rateU}$	4	0.1129	6 (tie)	0.0726
$Chat_{polarity}$	8	0.0718	4	0.1006
$\boldsymbol{Meta_{view}}$	**1**	**0.2048**	**1**	**0.2179**
$Meta_{like}$	9	0.0588	8	0.0614
$Meta_{dislike}$	7	0.0723	9	0.0447
$Meta_{subD}$	6	0.0972	5	0.0894
$\boldsymbol{Meta_{subT}}$	**3**	**0.1132**	**3**	**0.1388**
$Meta_{enabled}$	10	0.0117	10	0.0009

Finally, we evaluate the parameter estimation of the live chat feature extraction module. As shown in Table IV, we observe the overall crowd vote ($Chat_{ocv}$) feature plays an important role (ranked 2nd) in detecting copyright infringement videos. It is interesting to further look into the decomposition of the crowd votes. The estimation of the weights of the crowd votes are shown in Table V. We observe that colluding, negativity, and quality votes are strong indicators of whether a video is copyright-infringing (i.e, the weights are either high or low). In contrast, the relevant vote seems to be a weak indicator. After a careful investigation, we find the main reason is that some live streams of the games are only in *audio* (thus non-copyright-infringing) but the audience still posts chat messages with soccer or NBA-related terms in those videos.

Table V: Parameter Estimation ($\phi_{i,k}$) for Crowd Votes

Datasets	Colluding	Negativity	Relevance	Quality
NBA	0.968	0.382	0.589	0.920
Soccer	0.890	0.257	0.497	0.872

VI. Conclusion

In this paper, we develop the first crowdsourcing-based solution (i.e., CCID) to address the copyright infringement detection problems for live video streams on online social media. The proposed scheme is robust against sophisticated streamers by leveraging the valuable clues from the unstructured and noisy live chat messages from the audience. Using two real-world live stream datasets, we have demonstrated that CCID can significantly outperform ContentID from YouTube by detecting more copyright-infringing videos and reducing the number of legal streams of being mistakenly taken down.

Acknowledgement

This research is supported in part by the National Science Foundation under Grant No. CBET-1637251, CNS-1566465 and IIS-1447795, Army Research Office under Grant W911NF-17-1-0409, Google 2018 Faculty Research Award. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] D. Wang, B. K. Szymanski, T. Abdelzaher, H. Ji, and L. Kaplan, "The age of social sensing," *arXiv preprint arXiv:1801.09116, to appear in IEEE Computer*, 2018.

[2] "Statistics of live stream market," https://www.researchandmarkets.com/research/8xpzlb/video_streaming, accessed: 2018-04-07.

[3] P. Tassi, "Game of thrones' sets piracy world record, but does hbo care?" *Forbes*, vol. 4, p. 15, 2014.

[4] D. King, "Latest content id tool for youtube," *Google Blog*, 2007.

[5] X. Nie, Y. Yin, J. Sun, J. Liu, and C. Cui, "Comprehensive feature-based robust video fingerprinting using tensor model," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 785–796, 2017.

[6] C.-Y. Lin, "Watermarking and digital signature techniques for multimedia authentication and copyright protection," Ph.D. dissertation, Columbia University, 2001.

[7] C. I. Podilchuk and W. Zeng, "Image-adaptive watermarking using visual models," *IEEE Journal on selected areas in communications*, vol. 16, no. 4, pp. 525–539, 1998.

[8] S. H. Low, N. F. Maxemchuk, and A. M. Lapone, "Document identification for copyright protection using centroid detection," *IEEE Transactions on Communications*, vol. 46, no. 3, pp. 372–383, 1998.

[9] J. Waldfogel, "Copyright protection, technological change, and the quality of new products: Evidence from recorded music since napster," *The journal of law and economics*, vol. 55, no. 4, pp. 715–740, 2012.

[10] M. M. Esmaeili, M. Fatourechi, and R. K. Ward, "A robust and fast video copy detection system using content-based fingerprinting," *IEEE Transactions on information forensics and security*, vol. 6, no. 1, 2011.

[11] C.-L. Chou, H.-T. Chen, and S.-Y. Lee, "Pattern-based near-duplicate video retrieval and localization on web-scale videos," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 382–395, 2015.

[12] D. Y. Zhang, C. Zheng, D. Wang, D. Thain, X. Mu, G. Madey, and C. Huang, "Towards scalable and dynamic social sensing using a distributed computing framework," in *Distributed Computing Systems (ICDCS), 2017 IEEE 37th International Conference on*. IEEE, 2017, pp. 966–976.

[13] D. Wang, T. Abdelzaher, and L. Kaplan, "Surrogate mobile sensing," *IEEE Communications Magazine*, vol. 52, no. 8, pp. 36–41, 2014.

[14] D. Y. Zhang, D. Wang, H. Zheng, X. Mu, Q. Li, and Y. Zhang, "Large-scale point-of-interest category prediction using natural language processing models," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1027–1032.

[15] D. Y. Zhang, D. Wang, and Y. Zhang, "Constraint-aware dynamic truth discovery in big data social media sensing," in *Big Data (Big Data), 2017 IEEE International Conference on*. IEEE, 2017, pp. 57–66.

[16] D. Wang, L. Kaplan, H. Le, and T. Abdelzaher, "On truth discovery in social sensing: A maximum likelihood estimation approach," in *Proc. ACM/IEEE 11th Int Information Processing in Sensor Networks (IPSN) Conf*, Apr. 2012, pp. 233–244.

[17] D. Wang, M. T. Amin, S. Li, T. Abdelzaher, L. Kaplan, S. Gu, C. Pan, H. Liu, C. C. Aggarwal, R. Ganti, X. Wang, P. Mohapatra, B. Szymanski, and H. Le, "Using humans as sensors: An estimation-theoretic perspective," in *Proc. 13th Int Information Processing in Sensor Networks Symp. IPSN-14*, Apr. 2014, pp. 35–46.

[18] D. Wang, T. Abdelzaher, L. Kaplan, and C. C. Aggarwal, "Recursive fact-finding: A streaming approach to truth estimation in crowdsourcing applications," in *Distributed Computing Systems (ICDCS), 2013 IEEE 33rd International Conference on*. IEEE, 2013, pp. 530–539.

[19] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "On opinion characterization in social sensing: A multi-view subspace learning approach," to appear in Distributed Computing in Sensor Systems (DCOSS), 2018 International Conference on. IEEE, 2018.

[20] J. Marshall and D. Wang, "Mood-sensitive truth discovery for reliable recommendation systems in social sensing," in *Proceedings of International Conference on Recommender Systems (Recsys)*. ACM, 2016, pp. 167–174.

[21] M. T. Al Amin, T. Abdelzaher, D. Wang, and B. Szymanski, "Crowd-sensing with polarized sources," in *Distributed Computing in Sensor Systems (DCOSS), 2014 IEEE International Conference on*. IEEE, 2014, pp. 67–74.

[22] Y. Zhang, N. Vance, D. Zhang, and D. Wang, "Optimizing online task allocation for multi-attribute social sensing," in *The 27th International Conference on Computer Communications and Networks (ICCCN 2018)*. IEEE, 2018.

[23] C. Huang and D. Wang, "Topic-aware social sensing with arbitrary source dependency graphs," in *International Conference on Information Processing in Sensor Networks (IPSN)*. ACM/IEEE, 2016, pp. 1–12.

[24] T. Steiner, R. Verborgh, R. Van de Walle, M. Hausenblas, and J. G. Vallés, "Crowdsourcing event detection in youtube video," in *10th International Semantic Web Conference (ISWC 2011); 1st Workshop on Detection, Representation, and Exploitation of Events in the Semantic Web*, 2011, pp. 58–67.

[25] D. Y. Zhang, J. Badilla, H. Tong, and D. Wang, "An end-to-end scalable copyright detection system for online video sharing platforms," in *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2018*, 2018, accpeted.

[26] D. Zhang, D. Wang, N. Vance, Y. Zhang, and S. Mike, "On scalable and robust truth discovery in big data social media sensing applications," *IEEE Transactions on Big Data*, 2018.

[27] D. Wang and C. Huang, "Confidence-aware truth estimation in social sensing applications," in *International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2015, pp. 336–344.

[28] D. D. Lewis, "Naive (bayes) at forty: The independence assumption in information retrieval," in *European conference on machine learning*. Springer, 1998, pp. 4–15.

[29] S. Loria, P. Keen, M. Honnibal, R. Yankovsky, D. Karesh, E. Dempsey *et al.*, "Textblob: simplified text processing," *Secondary TextBlob: Simplified Text Processing*, 2014.

[30] X. Wei and W. B. Croft, "Lda-based document models for ad-hoc retrieval," in *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 2006, pp. 178–185.

[31] T. G. Dietterich, "Ensemble methods in machine learning," in *International workshop on multiple classifier systems*. Springer, 2000.

[32] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 2016, pp. 785–794.

[33] A. Liaw, M. Wiener *et al.*, "Classification and regression by randomforest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[34] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "Liblinear: A library for large linear classification," *Journal of machine learning research*, vol. 9, no. Aug, pp. 1871–1874, 2008.

[35] D. W. Ruck, S. K. Rogers, M. Kabrisky, M. E. Oxley, and B. W. Suter, "The multilayer perceptron as an approximation to a bayes optimal discriminant function," *IEEE Transactions on Neural Networks*, vol. 1, no. 4, pp. 296–298, 1990.