Discovering Illegal Advertisements Hidden in Free Online Steaming Films

Shengyu Chen
School of Informatics,
Computing and Engineering
Indiana University Bloomington
Bloomington, Indiana 47408
Email: chen435@iu.edu

Xinran Dai
School of Informatics,
Computing and Engineering
Indiana University Bloomington
Bloomington, Indiana 47408
Email: xinrdai@iu.edu

Peng Wang
School of Informatics,
Computing and Engineering
Indiana University Bloomington
Bloomington, Indiana 47408
Email: pw7@iu.edu

***Abstract*—In recent years, online streaming services are becoming more and more popular. YouTube, one of the most popular online streaming websites, releases official editions of movies that people can purchase to watch. However, some people can take advantage of these movies without having the copyrights. These people upload videos without copyrights to various websites including YouTube, and they can insert advertisements in these videos. Videos like this might contain faulty advertisements such as fake luxury product shopping, gambling, or links to other websites.**

Our research aims to find those advertisements and establish an analysis on them. Specifically, we first look for videos on YouTube that might violate copyrights, and then by extracting the context of the frames, we can analyze the similarities between the advertising contents with such faulty advertising. In this research process, we utilize Natural Language Processing (NLP) techniques to find the similarities. Since this project can involve a large amount of data, we compare a few different NLP approaches in order to find one that can predict the best accuracy in the least time.

I. Introduction

A. Background

As technology develops, it is more and more common that people choose to watch movies or TV shows online. Online streaming services have freed people from carrying the physical disk and having to have a DVD player. Instead of looking for one particular DVD in the store, people can simply log in to one website and have access to a great variety of videos according to whatever they want to watch. It is much more convenient and saves people the time from going to the store and back. There are many famous websites that streams online movies such as Netflix and YouTube, where Netflix hits 55 million subscribers alone in the year of 2018, and YouTube has around 30 million active users per day. People can pay for official editions of videos. They can choose from millions of movies, entertainment shows, documentaries and TV series.

YouTube, in particular, not only provides users with official editions, but also serves as an enormous platform for regular users to upload their own videos. People who upload their videos are called YouTubers. It is common for popular video YouTubers to include advertisements in their uploads, and their videos and their advertisements are both under the policy and restrictions from YouTube. Once a YouTuber uploads a video of which the content or the advertisement violates YouTube policies, YouTube can take off their videos. YouTube has a strict screening and reviewing process, however, there are still some videos that violate these rules. In addition to this, since YouTube allows regular users to upload videos, some people can take advantage of movies or TV shows, uploading them to the public without having the copyrights. Moreover, the videos can contain faulty advertisements which violate YouTube policies. Whenever another user browses this kind of video, the user is exposed to advertisements that is not approved by YouTube or the producers. Here is an example of the advertisements on YouTube in such videos.

Fig. 1. Example for YouTube faulty advertisements, including luxury product shopping and gambling

This example is from a screenshot of a Chinese movie called Dying to Survive that was released in June, 2018 in China and there are advertisements in Mandarin Chinese added to the film. As shown in the picture, above the main frame of this movie, two lines of advertising about so-called luxury product for sale are added on top of the black background. It says that they are selling a hundred percent authentic luxury products from official stores and the products can be shipped to all over the country. Then there is contact information of the sellers. At the lower central frame, two new lines are added to the frame which says that the

upload of this movie is sponsored by a casino in Macau with a link to their website. In addition, it also says that they are giving away a million RMB at the casino.

B. Approaches

Our research focuses on this kind of faulty advertisements that are hidden in free streaming videos on YouTube and the goal is to establish a knowledge base about them since there are not many researches that have been done on this particular topic. We start by manually checking some of the videos to gain a basic understanding of the advertisements. In order to conduct a more thorough research, we ask a few questions to begin with:

- What types of videos might contain faulty advertisements?
- What languages are these videos originally in?
- How likely are the faulty advertisements to occur in the videos?
- At what position of the video does this kind of advertising occur, for example, is it at the beginning of the video?

The research starts with the implementation an automatic searching snippet to search for potential videos without authorization or copyrights. Once we can find a list of movies, we start playing the movies and taking screen-shots at time spots when the advertisements usually occur. After we have the frames of the videos, we extract the texts from the screenshots using computer vision strategies. We are using the open-source computer vision library OpenCV for this step to get the contents. During the research process, we compared two computer vision algorithms, convolutional neural network (CNN) and optical character recognition (OCR) and the comparison of the two approaches shows that OCR is a more accurate and faster approach for extracting texts. We store the texts into txt files for further use of NLP techniques.

Process.png

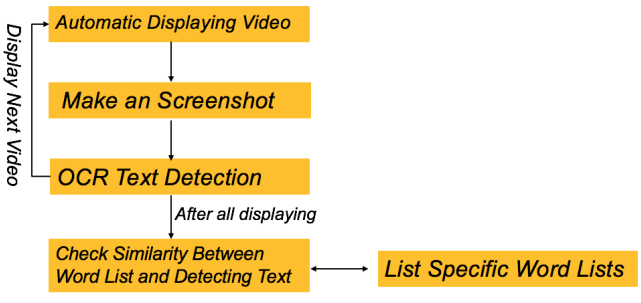

Fig. 2. Research process

When processing the extracted texts from the videos, we tried three different approaches of NLP in order to get which was the most accurate and the fastest way to get the similarities between extracted contents and key words of faulty advertisements. The three NLP techniques are term frequency (tf), term frequency-inverse document frequency tf-idf, and Jaccard Similarity. We find out that tf has the best performance for finding the similarities and tf-idf takes the longest time, while Jaccard Similarity algorithm returns the least accurate results for similarities. We will present more detailed explanations on how we analyzed the similarities, including the categories of the advertisements, related key words, and the performance of the techniques.

C. Findings

Our research is based on a variety of videos, and most of them are in Mandarin Chinese and Cantonese, including other videos of Japanese animations, English movies and TV shows. The research checked 300 videos in total, and 68 of them contained faulty advertisements. We find out that the among these videos, 54.41% of them contained luxury product shopping, and 66.18% of them contained advertisements about gambling. Among these advertisements, 19.12% are in Traditional Chinese and 80.88% are in Simplified Chinese. As for the details in regarding to the data source and the research process, we will explain in details later in the paper.

II. Background and Related Work

A. Background: YouTube Policy

The videos that the research studies upon violate at least one of the YouTube policies. The most common YouTube policy that gets violated is their policy about copyrights. According to YouTube policies on copyrights [1], users who upload videos should only upload videos to their own copyrights, or use the videos without infringing others' copyrights. Videos we find all violate this policy, and more specifically, they violate the policy about audiovisual works, which refer to TV shows, movies, and other online videos [2]. The channels that uploaded these videos did not give credits to the owners of the videos, and some of the videos were recorded by themselves which also violates YouTube copyright policies. For example, one of the videos we studied on was a Chinese movie called Hidden Man that was released in July, 2018. The video we found on YouTube was an over-an-hour long full movie of Hidden Man, and the account that uploaded this video was not an official account of the producer. The YouTube video page did not provide with any information of the producers or give any credits to them.

Moreover, YouTube dis allows links to external websites that may have concerning content such as pornography or copyright issues. Here is an example where the content of the video contains a link to an external website, and this external website has the movie which violates copyright of the movie

Christopher Robin. The advertisement is in simplied Chinese but the link is to an external website.

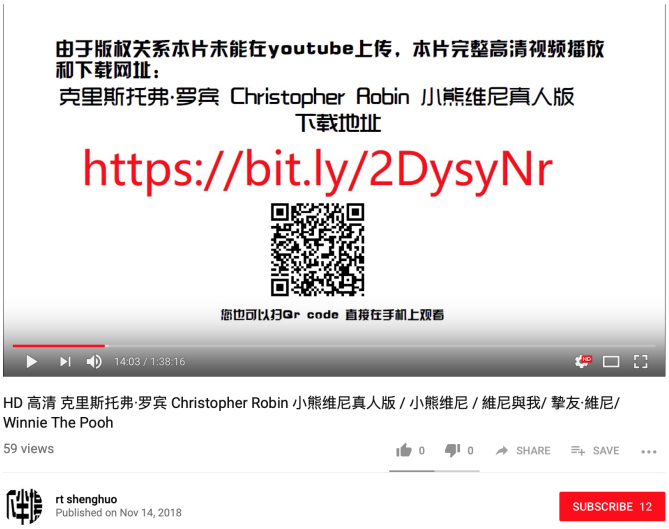

Fig. 3. Example of YouTube video containing external links which has content violating copyrights

As for videos that contain advertisements, such as luxury product shopping and gambling, they also violate another YouTube policy that is about spam, deceptive practices and scam [3]. YouTube policies on spam, deceptive practice and scam clarifies the dis-allowance of links in content to external websites if the websites containing pornography, malware and other contents that do not meet the requirements of YouTube Community. One example is as shown in Figure 1, where the advertisement says that the casino is giving away a million RMB for users. This can be a misleading advertising content as it does not provide with any other information and only tries to draw attention from the audience by saying there is a money give-away. The advertisement on luxury product shopping in the same figure might also violate the YouTube policy since this is not an advertisement from an official store or brand, and there is a possibility that this is another luxury goods fraud.

In addition, according to American Composites Manufacture Association (ACMA), gambling advertisements that target children are banned online [4]. As the study shows, the videos we discovered on YouTube were open to audience of any age group, which means that children can also see the gambling advertisements.

The advertising content discovered in the YouTube videos in this study is largely related to scam and gambling, and for those that do not contain advertisements like this, the videos violate YouTube policies on copyrights. Even though there is no direct hyper links that can redirect users to external websites, these advertisements on YouTube can still be a security issue since people can be mislead to scam and gambling, or they can follow the links that is printed on the frame. As the advertisements indicate, the links are to online interactive gambling websites which may cause financial loss of the users.

B. Related Work

In this study, one of the discovered advertising kinds is about gambling. According to the study conducted by Julia Braverman and Howard J. Shaffer in 2010 [5], online gambling can change people's behaviors dramatically and the effects are not on the positive side. They tracked the pattern change in people who started gambling on the Internet over the first month, and presented that seventy percent of the gamblers closed their financial account. This shows that gambling can be problematic because of people's impulsive gambling habits and the increasing amount of money they put in gambling. As Braverman and Shaffer's study suggests, Internet gambling can be problematic and this is one of the advertisements our study has found. Another study by Alissa Sklar, Jeffrey L. Derevensky et al [6] shows that on gambling shows that advertising on gambling can greatly affect underage youth. Their study focuses on the appealing effects of gambling advertisements and what impacts they have on underage youth. The results of their study suggests that teenagers and young adults can be drawn to gambling due to the exposure to gambling advertisements. As discussed in previous section, the videos we found on YouTube were open to any age group, making it also viewable to teenagers and young adults.

Our research establishes a way to detect the contents of videos on YouTube in order to find content in advertising that might

- violate YouTube policies or Federal laws
- contain illicit advertisements about gambling

The reason of the first goal is to help YouTube with discovering violations of their rules and policies. The reason to the second is that, as discussed above, advertising on gambling can be problematic and lead to high-risk gambling activities.

There are a lot of studies on free online streaming movies, especially on how some of the websites can contain malicious advertisements. However, these has not been any study or research on illicit advertising on YouTube. For example, the research conducted by M. Zubair Rafique, Tom Van Goethem et al [**?**] shows that free streaming services can contain many unwanted advertisements with pop-windows, and many of these advertisements contain links to downloading malicious content. There are rarely studies on detecting illicit YouTube advertising inserted by users, and especially not in simplified Chinese.

That being said, some previous studies suggest that YouTube has a great impact on the youth. This implies that any misleading content, one of which is gambling and the focus of our study, can have negative impacts on teenagers and young adults since YouTube has a great base of young users. According to statista.com, in the year of 2018, 96% of U.S internet users in between age 18-24 are YouTube users [8]. One of the researches is conducted by Milad Dehghani, Mojtaba Khorram Niaki et al which evalutes the influence of YouTube advertisements on young customers [9]. Their

results suggest that the advertisements on YouTube channels have a great positive impact on customers. The brands that are frequently displayed in YouTube advertising received more positive reviews and impressions. This shows the importance of our study as in the gambling advertisements that our study has discovered, Grand Lisboa Macau is the most frequent hotel brand that occurs in the advertisements. Advertisements like such can be high-risk and lead younger users to gamble or get involved with interactive internet gambling activities.

There are many papers and studies on security that utilizes natural language processing techniques. For example, in the research conducted by Xiaojing Liao, XiaoFeng Wang et al, the research team uses NLP techniques to build a connection between the illicit advertising words and website information on innocent top level domains such as .org and .edu [11]. More specifically, the researchers uses the skip-gram model of word embedding techniques which maps words to a higher dimension vector by a function in order to establish a connection between words. Their research paper can serve as a great example and model for us to execute our researching plan.

III. Acquiring video content from YouTube

In order to detect the advertisements in YouTube videos, the first step is to gain the content of the videos. Since YouTube has a strict policy in protecting their contents, it is in great difficulty to download video content or to intercept their internet packets. Our research group decided to take screen-shots of the videos, and use NLP techniques to extract texts from the frames.

Optical character recognition (OCR) techniques are widely used to recognize texts from printed documents, natural scenes or handwritten materials and transform them into machine encoded characters. To begin with, characters are input as images and OCR algorithm studies the images pixel by pixel in order to find the slightest pattern. OCR techniques learn patterns of them and then try to identify texts from input images. OCR uses deep learning algorithms such as k nearest neighbors to build the model.

In this study, we tried two of the most popular OCR software applications in python, PyTesseract and TensorFlow in order to find out which is the most accurate tool to use for OCR.

A. Challenges

The first challenge in this step the study faces is that texts are inserted on top of original films, and there can be many unnecessary texts from the original videos which can serve as noises. In the early stage of our research, we manually checked around 150 videos with comparison to other websites that also had illegal free streaming services for movies, TV shows and animations. We find out that most of these advertisements are presented on the top of the frame. Our approach to solve this problem is to take screen-shots only at the top of the frame. This way we greatly decrease the amount of noises in our data.

Furthermore, some of the videos may be in high resolution while others are in low resolution. Videos with low resolution can be harder for OCR to detect characters and can also take a longer period of time. Our study compares the use of TensorFlow package and PyTesseract so that the method we choose is the most effective.

B. Comparison

We tested two images with simplified Chinese characters using each of PyTesseract and TensorFlow. Ten of these images are with low resolution and the other ten are with high resolution. We compared the average running time and the average accuracy of using each software.

PyTesseract

	Time	Accuracy
High Resolution	0.34	92.54%
Low Resolution	1.128	74.77%

Fig. 4. Average time cost and average accuracy of using PyTesseract

TensorFlow

	Time	Accuracy
High Resolution	0.57	98%
Low Resolution	1.89	81%

Fig. 5. Average time cost and average accuracy of using TensorFlow

As shown in Figure 4 and Figure 5, the use of TensorFlow software produces a better result in accuracy, while the use of PyTesseract takes less time with both low resolution and high resolution. In order to find the best approach combined with NLP techniques, our teach later tested TensorFlow and PyTesseract each with three NLP techniques: term frequency, term frequency-inverse document frequency and Jaccard similarity.

IV. Natural Language Processing

In this research, we use natural language processing (NLP) techniques after we extract texts from the screen-shots of the videos in order to establish a connection between the texts and illicit advertisements. Our research takes a set of key words

that are closely related to three categories of the advertisements that are inserted into YouTube videos: (1) gambling, (2) money loaning, (3) luxury goods for sale. When our team was manually checking the advertisements, we found out that due to the limitations of the screen and because the video cannot be interfered with too many words, the advertisements are usually one or two sentences long, sometimes three but usually not longer. Thus, it is important to find the right key words to use for building the relationship and we should make use a NLP technique that is light and fast, suitable for this kind of situation.

As there are many NLP techniques being used for various purposes, we decided to first run a sample test on three NLP techniques that are the most popular and widely used, which are

- Term frequency
- Term frequency-inverse document frequency
- Jaccard similarity

A. Term frequency

Term frequency (tf) is one of the most widely used NLP techniques so far. The goal of using tf is to find out the level of reverence between certain queries and a chunk of documents by looking for the frequencies of the words within the documents. The tf technique assigns a weight for each word that has occurred in the document, where the more frequent the words appear, the heavier the weights are assigned. The relationship or the level of relevance is determined by the score which is the weight of the query gets assigned.

The formula for weight assigning that we use looks like this:

$$\mathrm{tf}(t,d) = 0.5 + 0.5 \cdot \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

Fig. 6. Weight assigning formula for term frequency method

This is the formula after double normalization in order to prevent bias which might occur in documents that are longer. $tf(t,d)$ stands for the term frequency of the term t in a document d. $f_{t,d}$ stands for the raw frequency or the number of occurrences of the term t in the document d, and it is divided by the number of occurrences of the most frequent term t' which belongs to the same document d.

B. Term frequency-inverse document frequency

Term frequency-inverse document frequency (tf-idf) is based on tf but it takes this approach one step further for a better accuracy. Tf-idf is a combination of the term frequency technique and the inverse document frequency (idf) technique.

The goal of using idf approach is to decide how much information can a term or a word contain. Instead of simply taking the frequency of a word into account, idf takes the result from dividing the number of total documents N by the number of documents d which contain the term t, and calculates the logarithmic value of this inverse fraction.

$$\mathrm{idf}(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$$

Fig. 7. Formula for inverse document frequency

The tf-idf technique takes the production of the result from using tf and the result of the use of idf.

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \cdot \mathrm{idf}(t,D)$$

Fig. 8. Formula for term frequency-inverse document frequency

The advantage of making use of tf-idf is that in order to assign a heavy weight for a certain term t, both of the result of using tf and the result of using idf have to be large values. According to the meaning of the two formulas, this means that the term t should be occurring at a rather high frequency but it is not so common that it appears in every document d of the total amount of documents D since idf needs to be low. The use of tf-idf makes sure to rule out common terms such as 'a', 'the' or 'have' that have very high frequencies but do not count as meaningful terms for building up a relationship or relevance.

C. Jaccard Similarity

The idea of Jaccard similarity, also known as Jaccard distance, was come up by Paul Jaccard, is a rather straightforward technique that has been widely used to find the distance between two objects.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

Fig. 9. The formula of calculating the Jaccard similartiy between two objects

As shown in Figure 9, Jaccard similarity calculates the portion between the size of the intersection between two objects and the total size of the two objects. In NLP, Jaccard similarity is calculated by dividing the number of common words shared by two documents by the number of words in

total from the two documents.

intersection.png

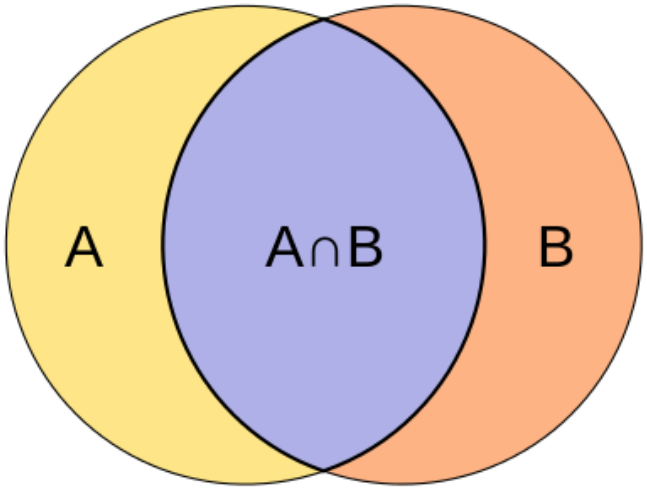

Fig. 10. The intersection between two objects forms the Jaccard similarity

D. Comparison among the three techniques

In order to find out which NLP technique is the most efficient to use, which means it has a rather high accuracy and takes a rather low running time, we first tested among the NLP techniques. We tested five sets of English queries that shares a moderate level of differences, which is ten queries in total, and five sets of English queries with high similarity, which is ten queries in total as well, using each of the three NLP techniques, tf, tf-idf and Jaccard similarity. For each test with the level of similarity, our team took the average performance result.

similarity for 3 methods.png

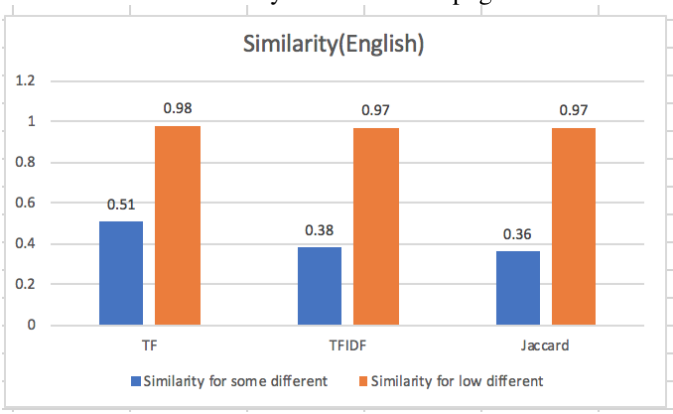

Fig. 11. The results of testing ten queries each for two levels of similarity in English using tf, tf-idf and Jaccard similarity

similarity for 3 methods.png

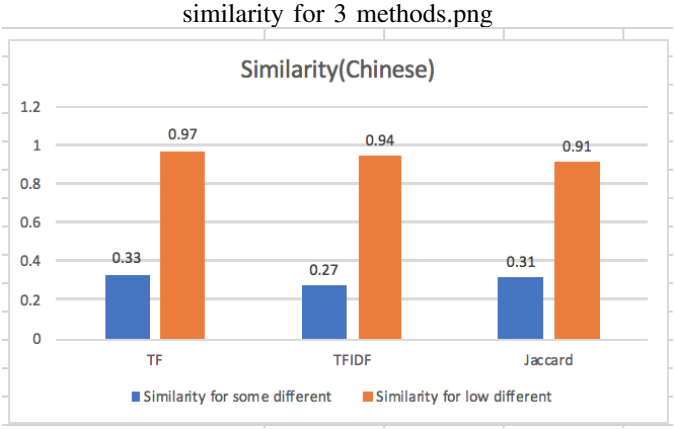

Fig. 12. The results of testing ten queries each for two levels of similarity in Simplified Chinese using tf, tf-idf and Jaccard similarity

As shown in Figure 11 and Figure 12, each of the results by running tf, tf-idf and Jaccard similarity on different sets of queries are very close to each other. At this point, it is hard to decide on which one of the three NLP techniques to use. Our team then conducted another set of tests on the combination of the three NLP techniques with the two OCR software.

E. Combination of OCR and NLP techniques

In order to get the best combination of OCR techniques and NLP techniques, we took 100 screen-shots of the test data setwhich are queries in Simplified Chinese that share a level of similarity. Our team choose to test on Simplified Chinese because most of the illicit advertisements are in Simplified Chinese. We set the number of testing data as 100 since processing two screen-shot takes less than one second and we increase the total number of test data set to let the program run at a longer time. This way, we can better observe which technique is the best to use. We set the accuracy threshold as 10%. All 50 pairs of queries share some similarity and if the similarity reaches 10%, we can say that the result of accurate. We tried six combinations, which are

1) PyTesseract + TF
2) PyTesseract + TF-IDF
3) PyTesseract + Jaccard Similarity
4) TensorFlow + TF
5) TensorFlow + TF-IDF
6) TensorFlow + Jaccard Similarity

And the results are as follows.

As shown in Figure 13, it is clear that the running time of using PyTesseract for extracting texts from the images in combination of either of the three NLP techniques are noticeably shorter than using TensorFlow. For each paired test, the use of PyTessearct can save around twenty seconds in

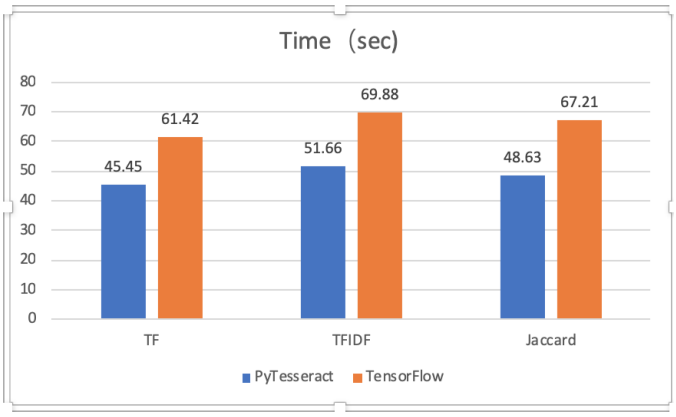

Fig. 13. The running time of the combination on testing OCR and NLP techniques

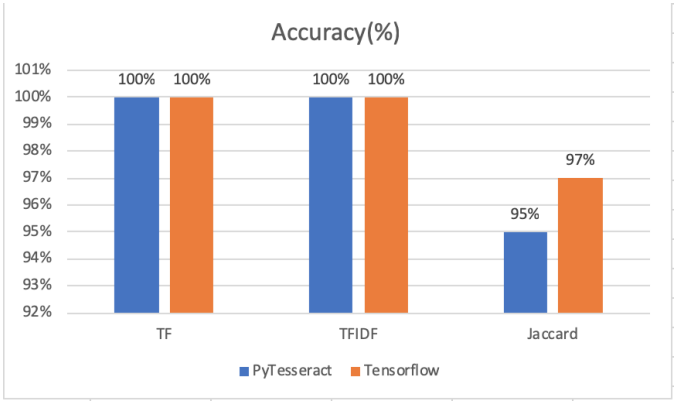

Fig. 14. The similarities of the combination on testing OCR and NLP techniques

the total running time. As a result, our team chooses using PyTesseract for our final program technique on OCR.

Figure 14 shows the results of the accuracy of the three NLP techniques. As shown in the graph, Jaccard similarity reaches a lower accuracy compared to tf and tf-idf techniques. With the combination of considering the running time, our team decides to choose the combination of PyTesseract and tf technique since tf techniques have the same level of accuracy as tf-idf, and saves five seconds compared to the use of tf-idf.

We believe that the reason Jaccard similarity returns results with a lower accuracy when compared to tf or tf-idf is because of the simplicity of this algorithm. Jaccard similarity only counts the number of occurrences of a term or word in a document without any method of normalization or processing the raw data. A common problem that might occur when applying Jaccard similarity is its bias against long documents, where meaningless words such as 'a' or 'the' can appear many times and be counted as important queries. On the other hand, we choose to use a normalized tf method which helps to reduce this bias against long documents. Furthermore, tf-idf not only uses the tf algorithm with fewer bias, but also utilizes inverse document frequency method which again reduces the importance of meaningless words in a long document. Compared to tf, tf-idf takes longer time to process the queries because tf-idf takes one more step in producing the inverse document frequency.

V. Methodology

A. Video type

There has been a lot of movies with illicit advertisements on scam such as luxury goods shopping on YouTube that are inserted onto the top of the frame of the videos. However, the gambling advertising which is specifically sponsored by one casino brand has only recently been discovered by our researchers. Since the casinos are located in Macau and Macau, Hong Kong and Cantong are demographically close to each other, people living in Macau mainly speak Cantonese and English [16]. Because of this shared culture, we blieve that productions from Hong Kong are likely to get inserted by Macau gambling advertisements. Thus, our team decided to take the data set including Hong Kong movies and Hong Kong TV series.

On the other hand, movies in Mandarin Chinese and English are popular in general, thus our data also includes these two kinds of videos because popularity draws attention of customers. In addition to movies and TV shows in Mandarin Chinese, English and Cantonese, we also take Japanese animations into our considerations. Furthermore, entertainment shows in Mandarin Chinese have been popular and proved to be a great approach for brands to put advertisements on. Because of the popularity of entertainment shows in China, our team also decided to take entertainment shows into account.

B. Data source of video list

This study takes the video list from iQiyi, a website that provides with online streaming services in China. It is one of the largest platforms with online streaming services. In 2018, the company of iQiyi, Baidu, raised $2.25 billion after the initial public offering of iQiyi [17]. iQiyi provides millions of movies in different languages from various countries, such as Chinese movies, English movies, French movies, Cantonese movies, etc. In addition, they also provides with entertainment shows or animations from all over the world. For example, the movies can be searched by language, country or region, genre of the movie and other criteria.

Our team decided to take the data set of video list from iQiyi by different requirements. Our video list includes movies in English and Mandarin Chinese, Japanese animation series or movies, entertainment shows in Mandarin Chinese, and TV series in English, Mandarin Chinese and Cantonese. We get a list of 110 kinds of videos in total.

list.png

	Mandarin Chinese	Japanese	Cantonese	English	Total Number
Movies	30	10	20	20	80
Other videos	20	10	20	20	70

Fig. 15. Video types and original languages of the videos

C. Automatic Search

This study uses an automatic searching snippet in Python that is used specifically on Google Chrome. For implementing the automatic searching snippet, we use the package from selenium, which is a set of tools for web browser automation [18]. More specifically, we make use of selenium webdriver that can be used on a variety of web browsers or under different environments.

Our original design involves the automatic searching process for possible videos which (1) fits the description of the video source list, (2) are long enough to contain the content of the full movies or the whole TV shows. This process can be achieved by using selenium webdriver, which can manipulate web browser activities. The selenium webdriver can find where the search bar of YouTube page is in Google Chrome web browser, and types in the input, which is the video name into the search bar and hit the search button. The selenium webdriver simulates human activities in web browsers. Then, selenium package can be used to look for videos that are over the full length of the video, for example, over an hour long for a movie and forty to fifty minutes for a TV episode.

However, due to the limit of our time, our research process is heavily depended on the researchers manually checking if any of the search results would be suitable for our research project. Our project first searches on YouTube by video name, and our researchers would look through the top twenty to fifty search results in order to find the videos with correct hour. There is no guarantee that there are the actual videos instead of blank videos.

Furthermore, our program takes an input link from the video list and reads the list line by line. Our program opens the web page with a YouTube video, and then our program starts to take screen-shots while the video is being played full-screen. The speed is set as one screen-shot per second because this is the right amount of time for a string of words to roll over on the screen. It is not too much time that many words in the frame, if there are any, stay almost the same, and it is also not too little time that the context of the potential advertisements on screen completely changes.

D. Implementation of screen-shot taking and OCR

For screen-shot taking, our program makes use of PyAutoGUI, a Python package tool for controlling keyboard and mouse activities [19]. By using PyAutoGUI, our program can decide on what portion of the screen should it take a screen-shot. Our study takes screen-shots at each second during the first three minutes of the videos, for three minutes again every thirty minutes also at the speed of one screen-shot per second for movies and every twenty minutes for shorter videos, such as TV drama and animation series. We take screen-shots at each second because it is the most suitable for acquiring information from possible illicit advertisements but not to wait for too long that there are missing words on the screen. We take screen-shots for continuous three minutes because it is not too long but it leaves enough time for a string of words to pass by on the screen if there are such advertisements. The reason that time gaps are set as every twenty minutes and every thirty minutes is based on our previous observation in our manual checking process. We believe that parties who put up the advertisements do not want the advertisements to interrupt the videos themselves so under most circumstances the advertisements only show up after a period of time. Twenty-minute and thirty-minute gaps are most common according to our observation.

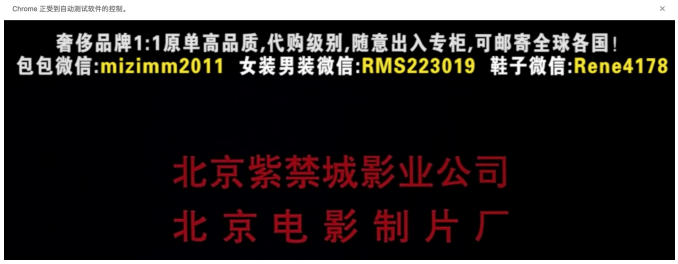

Fig. 16. Screen-shot sample

After taking each screen-shot, our program saves the file and immediately runs the OCR technique with PyTesseract to extract text information from the frame. Extracted texts from the same video are stored into the same file. PyTesseract provides with packages for identifying languages other than English, and in this case we utilize the packages for simplified and traditional Chinese. The default language for PyTesseract is English and even though the program sets the language as Simplified Chinese, PyTesseract can still automatically detect English characters in a paragraph of Chinese characters. This way, if there are external links that are in English showing up on the screen in the middle of the advertising, our program does not have to misclassify them as Chinese characters.

E. NLP

After extracting texts from a video, our program runs the extracted texts against a list of key words that are closely associated with three categories of illicit advertisements:

- Gambling
- Luxury goods shopping

- Money loaning

The list of key words include 'WeChat', 'luxury goods', 'gambling', 'Macau', etc. We set the threshold as 5% of the similarity. If a file of extracted text reaches similarity of 5% with the list of key words, it means that this video contains illicit advertising content. Since this is a preliminary project, our team double checks the text files manually to make sure that there is no mistakes when identifying them as illicit advertising content.

VI. Results

In the end, we find out that out of 150 videos, there are 68 of them contains illicit advertisements, and most of them contain advertising on both luxury goods shopping and gambling.

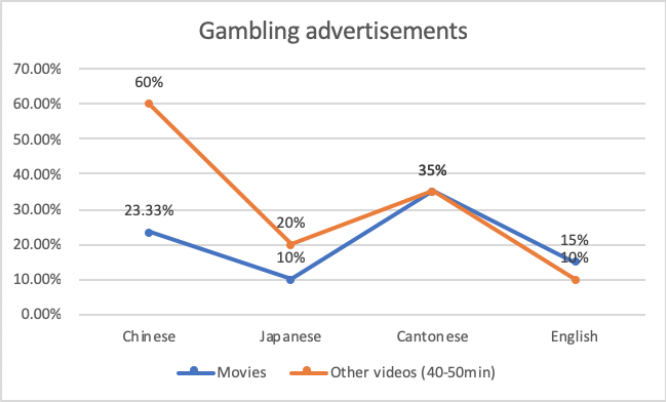

Fig. 17. Percentage of videos that contain gambling advertisements among the total number of videos

Figure 17 shows the percentage for videos with advertising on gambling in their content of the total number of videos in Mandarin Chinese, Japanese, Cantonese and English separated by the type of videos. For example, within 30 videos that are movies in Mandarin Chinese, 7 of them contain advertisements on gambling. Thus, the percentage of Chinese movies that contain gambling advertisements is 23.33%.

Figure 18 shows the percentage for videos with advertising on luxury goods shopping of the total number of videos in the same four languages. Our study uses the same technique when calculating the percentage for both of the two categories on advertisements.

We originally run our NLP technique on three categories of advertisements, gambling, money loaning and luxury goods shopping. However, in the past month of our research process, our research has not found advertisements on YouTube about money loaning yet. As a consequence, we are not showing this category in our results. In future studies, researchers can conduct a more detailed research on looking for advertisements on money loaning.

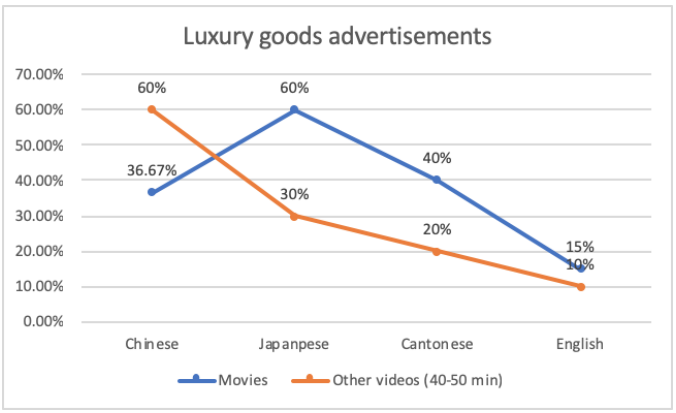

Fig. 18. Percentage of videos that contain luxury goods shopping advertisements among the total number of videos

VII. Conclusion

This study aims to establish a basic knowledge base on illicit advertising content in YouTube videos, not in comment section or description section. We focus on videos that have advertisements inserted to the frames and the videos themselves violate YouTube policies on copyrights and Federal law on copyrights since all the videos we find are illegal versions of movie productions, TV series productions and entertainment shows. In addition, those videos which contain advertisements about gambling and luxury goods shopping violate YouTube policies on scam. The list of videos we study include movies in Mandarin Chinese, Japanese, Cantonese and English, TV series in Mandarin Chinese, Cantonese and English, Japanese animation movies and animation series, and entertainment shows in Mandarin Chinese. As for the content of the advertisements, our research studies three categories of illicit advertising content: gambling, money loaning and luxury goods shopping.

For the methodologies of our research, we use OCR technique to extract texts from the screen-shots of the frames in the videos. More specifically, we utilize the tool PyTesseract in comparison to the use of TensorFlow since PyTesseract saves running time and our study has to deal with a huge amount of data. We use NLP techniques to find the similarities between the key words of each category of the illicit advertisements with the extracted texts from the videos in order to establish a connection between the two. After comparing the three NLP techniques, tf, tf-idf and Jaccard similarity, our research decides to use tf for our NLP methodology since this is time-efficient when compared to the use of tf-idf, and reaches a high accuracy when compared to Jaccard similarity.

The process of our study involves manual checking which requires humans to double check the accuracy to reduce the mistakes our program might make. The reason is that within such a short time and lack of experience, our research team wants to make sure that our results are trustworthy. However, this being said, it is executable to develop a fully automated

program to detect these advertisements on YouTube.

As for the results, we find out that luxury goods shopping from videos of the four languages has a higher ratio compared to the gambling advertisements. The reason could be that luxury goods shopping advertisements have occurred in illegal free online streaming videos for a long time, and advertisements on gambling has just appeared recently. Thus, when looking at the total number of these two different kinds of advertisements, it is reasonable that there are more luxury goods shopping advertisements.

We also notice that Mandarin Chinese have the highest percentages in both gambling and luxury goods shopping advertisements. We believe the reason behind this is that the advertisements are in Chinese which means the advertising is targeted at the group that speak Chinese. Moreover, luxury goods shopping advertisements in Japanese animation movies and series is as high as 60% and we believe the reason is that there are not many such videos on YouTube. Since the total number of the videos is small, the percentage would be higher even though there are not many videos that contain the advertisements.

One of the other findings is that videos in English are harder to find among all the videos in the four languages. We suggest that YouTube is based in America and therefore does a better job on filtering out English videos that violate copyright policies. Because of this, there are not many recent uploaded videos in English, and the videos we can find are uploaded a few years ago. We believe that they were uploaded before or during the beginning of inserted advertisements on luxury goods or gambling, resulting in small numbers of videos in total and videos with illicit advertising content.

As for videos in Cantonese, the occurrences of both gambling and luxury goods shopping advertisements are not the highest but also not the lowest. We believe the reason is the moderate popularity of videos in Cantonese in general.

The result of looking for illicit advertisements suggest that YouTube is more efficient on detecting videos in English that violate YouTube policies, more particularly, Mandarin Chinese. The result of our study suggests that it is possible that there are other situations where the video content violates YouTube policies by containing illicit advertisements but it is harder for YouTube to realize this because they are in a foreign language to English. However, that being said, this is a preliminary research and just the very beginning of understanding illicit advertisement contents on YouTube.

Future researches can be conducted on a more thorough and detailed analysis of the videos in various languages and of other video types. Future studies should look for a larger data set on YouTube which will make the results more convincing and more clear. Another possible approach to improve our study is to work on a more detailed analysis on the content of the advertisements, for example, figuring out what specifically the advertisements point to. We can check out the external links and research on the casino brands, or contact the luxury goods sellers in order to gain a further understanding. Methodologies can also be improved by implementing a fully automated program. The techniques in this study on OCR and NLP are applicable but maybe there are more efficient approaches than PyTesseract and term frequency method.

REFERENCES

[1] ”Copyright on YouTube” https://www.youtube.com/yt/about/copyright/#support-and-troubleshooting

[2] ”YouTube: What is Copyright?” https://support.google.com/youtube/answer/2797466?hl=en

[3] ”YouTube: Spam, deceptive practices and scams policies.” https://support.google.com/youtube/answer/2801973?hl=en

[4] ”ACMA: Do you think a gambling ad has crossed the line?” https://www.acma.gov.au/theACMA/alcohol-and-gambling-ads

[5] Julia Braverman, Howard J. Sharffer. How do gamblers start gambling: identifying behavioural markers for high-risk internet gambling. European Journal of Public Health, Volume 22, Issue 2, 1 April 2012, Pages 273278.

[6] Jeffrey Derevensky, Alissa Sklar, Rina Gupta, Carmen Messerlian. An Empirical Study Examining the Impact of Gambling Advertisements on Adolescent Gambling Attitudes and Behaviors.

[7] M. Zubair Rafique, Tom Van Goethem, Wouter Joosen, Christophe Huygens, Nick Nikiforakis. Its Free for a Reason: Exploring the Ecosystem of Free Live Streaming Services. Published in NDSS 2016.

[8] ”Percentage of U.S. internet users who use YouTube as of January 2018, by age group. ” https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/

[9] Milad Dehghani, Mojtaba Khorram Niaki, Iman Ramezani, Rasoul Sali. Evaluating the influence of YouTube advertising for attraction of young customers.

[10] Boris Epshtein, Eyal Ofek, Yonatan Wexler. Detecting text in natural scenes with stroke width transform.

[11] Xiaojing Liao, Kan Yuan, XiaoFeng Wang, Zhongyu Pei, Hao Yang, Jianjun Chen, Haixin Duan, Kun Du, Eihal Alowaisheq, Sumayah Alrwais, Luyi Xing, and Raheem Beyah. Seeking Nonsense, Looking for Trouble: Efficient Promotional-Infection Detection through Semantic Inconsistency Search.

[12] ”Wikipedia: term frequency-inverse document frequency”. https://en.wikipedia.org/wiki/Tfidf

[13] ”Wikipedia: Jaccard Index.” https://en.wikipedia.org/wiki/Jaccard_index

[14] ”Basic Statistical NLP Part 1 - Jaccard Similarity and TF-IDF.” http://billchambers.me/tutorials/2014/12/21/tf-idf-explained-in-python.html

[15] ”Overview of Text Similarity Metrics in Python. ” https://towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50

[16] Sylvia Sao Leng Ieong. Reflections on the Language Issues in Macau: Policies, Realities, and Prospects.

[17] Bloomberg News. ”Baidu’s iQiyi Drops in Debut After IPO Raising $2.3 Billion. ” https://www.bloomberg.com/news/articles/2018-03-29/baidu-s-iqiyi-drops-in-trading-debut-after-raising-2-3-billion

[18] SeleniumHQ. https://www.seleniumhq.org

[19] ”Welcome to PyAutoGUIs documentation! ”https://pyautogui.readthedocs.io/en/latest/