Youtube Safety

Rangfu Hu

November 26, 2019

Abstract

With the growth in the popularity of information technology, social media's presence can be seen in day-to-day life of people of different ages and races. One of the benefits of this popularity is the increase in flow of information and the decentralization of media outlates. Video-sharing companies such as Youtube have provided a platform for users to produce/share videos. However, not everything is nice with regards to these platforms. The number of videos that are being uploaded on a daily basis is huge and sometimes some inappropiate videos do not get tagged and become available to the entire world. In this paper, we look at the problem of youtube safety and present a direction for identifying and reporting/removing harmful videos for children and adults. To acheive this goal, we introduce four categories of videos that should be taken care of and suggest methods for finding videos related to the four categories. We illustrate the results of our experiments using our proposed method.

1 Introduction

Social media, is an inseparable concept in our today life. Due to the amount of daily interaction, it has the potential to affect us in many ways. One of the most popular platforms is Youtube. Almost every one of us has used it for different reasons, from watching educational videos related to our field of study to watching entertainment videos such as music videos or funny video clips. Based on a report [3], Youtube has over 1 billion active users. They say: "If YouTube were a country, we'd be the third largest in the world after China and India" [3]. This huge amount of active users' is due to the freshness and amount of content that is being provided. About 300 hours of video gets uploaded per minute. [2] As a result, guaranteeing the quality and the content of this huge dataset of public videos is a real challenge. Everyone has access to it in order to upload and/or view videos with minimial restrictions.

The newly uploaded videos on Youtube do not necessarily have the minimum required quality as videos that have already been published. Considering the quality of the subject, some of them might be inappropriate for certain users or even all of them. This is a threat because, a user does not necessarily need to search for inappropriate videos. Some times, a user ends up at an inappropriate video by unintentionally. By having a smart-phone or a simple browser on a computer that you can find everywhere, you can browse unlimited hours of Youtube videos. You might start with one video that you where looking for, go to the related videos, from there, watch another related video and so on, untill you end up with a set of videos that are completely unrelated to the first video you started from. This raises this concern for many users: what if we end up seeing something really inappropriate?

1.1 Importance of Human Judgment

Youtube is hiring many people in order to review the videos for them. We believe that this way is not efficient and sustainable. First of all, youtube needs to pay many people in order to watch those videos and then decide whether something is inappropriate or not. This take so many employees to review all videos that has been uploaded to Youtube so far, so it is not feasible to check all videos. Hence, youtube only reviews the videos that already has been reported as inappropriate. We believe that this was is inefficient as well, because when a video gets reported, if it is really inappropriate, then it means that people have already watched it. The fact that all people do not report a video if it is inappropriate makes it worse, in other words, when a video gets reported, it might mean that many people have already watched it. In other words, a human judgement while accurate is more of a re-active approach. We want to provide a more pro-active approach. Our

goal is to help youtube recognize these videos in a faster manner. We want to automate the part the role of people reporting the videos, but not the role of people who review those videos, because we believe that reviewing these videos and judging them by human being is really important, since the judgment of automated programs might not be accurate enough.

Reviewing the videos, even those who have been reported is really an important and inevitable part of the process. Unfortunately, there is an unhappy incident which happened April 2018, [1] when a Youtuber was angry of reviewers, since her videos where reported so she could not make any money from the advertisements on her channel, so she went to youtube headquarters and killed herself and shot people. Not that everyone acts like her, but imagine the big picture of so many people getting angry, because a fully automated algorithm had detected their videos as inappropriate. Because of the importance of this matter, in this paper we never try to remove the human being as the source of judgment. Instead, we try to help them find the candidates for inappropriate videos faster than just waiting for the users to detect them.

In the next chapter, first we explore the different categories we suggest as good categories as major inappropriate videos.

2 Categories

Lots of students use Youtube for educational purposes. Without any restrictions, watching an inappropriate video can endanger our children's feelings for a long time. For instance, assume that an innocent 9 years old girl is going through the videos related to her biology class. She may go to related videos and go on and on to finally find an adult video. Where she had no idea what she could find out there in the first place, she will end up with a lot of questions in mind, true questions of course but at a wrong time.

Furthermore, some videos can be inappropriate in general for all ages. For instance, a normal human being regardless of religion, race, place to live, etc, does not want to see someone murdering another one. Some videos can leave a mental scar forever in the mind of any healthy person. Later we will represent more details on how many violence videos we found on youtube.

There are other types of videos that can be inappropriate for people. The videos that might insult people with certain race, sex or religion. The users who upload these videos come from different cultures. In some of them a certain thing can be an insult while in the others it can be interpreted as offensive. Moreover, some people might want to insult others on purpose. We should not allow those people to satisfy their malicious intent.

Last but not least, we can claim that no parents want their children to be involved with drugs. All of us have heard stories about someone who were involved with drugs, all of them had a hard time quitting it and we believe most of these stories start in their age of youth. In our research, it did not took so long for us to find many videos about people harming each other or themselves when they were on drugs, or videos about how to use a certain sort of drugs, or about people talking how amazing it is to use drugs.

As a conclusion, we suggested that there are four categories that must be considered in order to filter out the inappropriate videos: Sexual, Violent, Racism and Drug-Related. Note that here Racism, stands for everything that insults people based on sex, age, race, nationality, religion, etc. or in other word, everything that people might get offended by them.

2.1 The Dilemma

The hardest part of our research is how to answer the question of what is appropriate or what is inappropriate. Different people have different opinions and different backgrounds, and this effects their judgment. One person can think of something as appropriate while others might disagree, because these two groups of people have different cultures. This might seem not important as it is in the first place, however, there are many examples which eludes us in practice when we want to categorize videos.

The first clue for us to some extent to categorize something as sexual is definitely is nudity. However, we cannot categorize all videos with nudity as sexual. There are a vast variety of examples which contain nudity, but they are not sexual. For instance, there are a numerous videos with someone totally nude in the middle of a class, when the rest of the class are painting her or his body. There are so many music videos with nude dancers. There are so many sports that take place with nude participants known as "naked yoga". There are so many events that at nude beaches

or national events in different countries when all the people on street get naked. Definitely, to some extent this videos contain adult related content, however, the paining class is more related to art as opposed to pornographic content. The music videos are definitely there for the sake of entertainment and the events can be mainly related to the news. This would be hard challenge for each of them to decide whether to remove them or not. From one side, we do not want our underage children to see those videos, and we do not want to remove them from youtube, since art students, people interested in sports, people interested in news, etc, might want to have them on youtube.

This dilemma does not end in the first category, for instance there are huge number of videos, basically news that contain contents about people from different races claiming their rights, sometimes this videos can contain violence of the police against those people, sometimes people who got hurt in those videos have no than uploading those videos when they want the people all around the world to hear their sounds. Unfortunately, sometimes this does not end with a on street violent, sometimes there are people getting shot by the police in different countries around the world, sometimes the videos are just about news reporters telling truth about those unhappy incidents. Should we remove those videos, since they contain violence? Should we remove them since they contain some contents that might offend people with different races or not?

Last but not least, there are many videos that has been made for the purpose of education, however those videos are telling facts about drugs. They are not encouraging people to use those drugs, but they are telling the advantages and disadvantages of using a certain type of drugs. Although the fact that using some drugs might for instance make you feel really happy, it does not mean that they are encouraging people to use those drugs, they are just scientific researches about the drugs. Moreover, in some places, some drugs might be illegal while in some other places they are legal, for instance in some states of the united states, the recreational use of cannabis is legal, while in the rest of them it is not, so some people from the legalized states might want to see so many videos about weed on youtube, while in other places it is a big taboo for the people. From which point of view should we look at the problem?

As a result, we suggest that since there is no accurate definition of what is appropriate and what is inappropriate, there is no deterministic algorithm or method to detect inappropriate from appropriate, hence, we suggest training a machine learning algorithm to help us go further.

3 Architecture

Our architecture consists of three main parts:

- Reviewer (Human being): As explained in section 1.1, this part of the system is inevitable. This part of the system is responsible to finalize the judgment of the whole system. If this part of the system recognizes a video as inappropriate, it will be labeled as inappropriate and added to the database of inappropriate videos, otherwise, it is a good example for exploring what might look like inappropriate at first, but indeed it is an appropriate video to enhance the machine learning algorithm.

- Database: This database contains both appropriate videos as well as inappropriate videos. The result of human judgment will be store here, and the machine learning code will learn from the labeled items in this database.

- Machine Learning Algorithm: This part of the system is responsible for finding candidates for inappropriate videos. It is trained based the data in the database. The most challenging part of the experiments is how to train this algorithm, in order enhance the accuracy of the candidates. This algorithm has access to the dataset and the output of this part of the system will be passed to the reviewers.

As you can see, this part of our architecture consists of 3 part, which perform a cycle of length 3, where the output of each part of the system is the input of the next part of the system. Without considering the challenges, the architecture seems to work perfect. In the next section, we are going to explore the challenges of this system.

4 Challenges

The main challenge of our work is finding the initial dataset to expand the system. We had several suggestions about how to provide the dataset to our system. Namely, the Youtube 8M dataset and Youtube API V3 are the main ideas. In the next section, we will explore each of them in details.

4.1 Youtube 8M

Youtube provides a set about 8 million videos known as Youtube 8M. Youtube has provided this set of videos based on searching for certain keywords. Then they have narrowed down the result of the searches by removing the videos of length less than 120 seconds or more than 500 seconds and by removing the videos with less than 1000 views. Then they have removed the sensitive videos from the dataset and finally they have published the remaining videos as Youtube 8M. We believe that this dataset is not a good choice for our system, based on the following reasons:

- They have removed the sensitive videos. Hence, there is nothing to be labeled as inappropriate videos in order to train our machine learning algorithm.
- There dataset is biased, since they have searched for certain keywords, the result appearing in this dataset is totally biased and does not represent a sampling of what is really on youtube. Thus, if we use this data as appropriate videos, the machine learning algorithm would be biased. For instance, there might be an appropriate video which was not related to the keywords they used to select Youtube 8M, then there is a chance that our algorithm detects it as inappropriate videos. Moreover, the dataset needs to have appropriate videos close to the inappropriate ones, in order improve the accuracy. For instance, we need to have educational videos about weed for our algorithm to detect appropriate videos and inappropriate videos about weed as opposed to detecting everything about the weed as inappropriate.

Consequently, we could not use this dataset as our dataset, however, we figured out that even youtube its self uses searching keywords in order to sample the videos. This was a great observation for us for the rest of the work. We will discuss this in details later. We realized that crawling youtube, the way that google uses to crawl all web-pages is not a feasible solution. Since the youtube its self did not implement such crawler to collect Youtube 8M. Moreover, there are over 7 billion videos on youtube (i.e. the same as the population of human being on earth) and more than 300 hours of videos are being uploaded to youtube per minute on a daily basis. Consequently, we have to forget the idea of crawling a real sampling of youtube.

We believe that the idea of crawling using certain keywords is a reasonable idea. Although for the same reason as Youtube 8M, the result of the crawling would be biased, we can use this to make our dataset more sensitive to the inappropriate videos and make our machine learning more accurate. In the next section, we will discuss our approach in more details.

4.2 Our Crawler

Our crawler uses the YouTube Data API v3 which is the API provided by youtube and contains almost all functionalities that youtube provides to its users. This API supports several languages, including pythin which we used, and provided the detailed information about each type of resource that is on the youtube. For example, one can fetch the title, description, comments, and subtitles from a video by providing its id. On the other hand, one can uses this API to explore videos by two main high-level methods: searching a keywords, and fetching the related videos.

A single video is represented by the Video object type, which contains the title, description, category Id, and another common information of the video in its snippet part. There are a numerous range of other information that comes with a single video that the user could select among them. For example, recording details, location, video details, and statistics are extra information of a video. The API lets the user select which information he or she needs because there a quota limit of 1000000 units for each user of this API, and each piece of information regarding each video has a specific cost in quota units. Therefore, users might need to select a susbet of those information pieces to minimize the quota usage. It can get more complicated when the comments, for example, are represented by two different resources types, CommentThreads and Comments. A video resource has at least one CommentThread, and Comments are availabe in paginated format for each CommentThread. Furthermore, the replies of each comment are partly included in the

comment resource, but needs additional queries to retrieve a full list of them. Each of these queries require a certain amount of quota units, i.e., fetching the CommentThread, fetching each page of comments, fetching the replies of each comments. (which could be recursively nested) Therefore, we need to take care when using this API such that we do not pass the quota limit.

We implemented a high-level abstraction of the API features, which seamlessly fetches the features we need for each video, especially comments and subtitles (we are referred as Caption in the API reference) which need complicated queries. However, we weren't able to fetch the subtitles at the end because of unexpected errors. Since we registered our API client in Google Cloud Console as a API driver which does not need any authetnication, maybe there is a need to authenicate to fetch subtitles. However, we had to use the youtube without authentication for two reasosn: 1- We need to have unbiased recommendations, i.e., in related videos. 2- We have a terminal API, and without authentication we had to login each time we ran the crawler.

In addition to fetching the textual features of each video, that is the features we need to learn, we implemented the ability of searching for a keyword, and fetching the related videos in our library. We especially need these functionalities to find safe videoes to include in our dataset, because the videoes retrieved by searching a suspicious keywords are generally safe videoes. By running our machine learning algorithm on a collection of videos retrieved by a keywords, we will find some candidates to include in our dataset as bad videoes. We can use these to trick to facilitate the process of finding bad videos, as well as using a number of the other videoes as safe videoes at the same time. (after checking weather it is really safe) Actually, using these videoes as the safe videoes in our machine learning algorithm would boost our performance, as these videos are closer to the border of safe videos, and help the algorithm to separate videos better.

The process of finding a collection of videos based on a keyword is worth mentioning too. We can observe that the results of a search query would diverge the main concept going down the search results. In addition, the chance of a video being safe will decreases when we go further in the search results. On the other hand, exclusively using the related video of a source video would have the same problems too. Thus, we applied a combination of these two functionality to gain a collection of videos based on a keyword. We search for a keyword, and then fetch a limited number of search results. For each of those videos, we go to a limited number of its top related videos and continue this process recursively.

5 Our Experiments

So far we our main challenge was to find the initial database of both appropriate and inappropriate videos.

In order to collect the set of appropriate videos, we need a set of sensitive keywords to start our crawling. There is a dataset made available by Microsoft, consisting of 500 adult words.

In order to provide the set of inappropriate videos, we believe that the probability of finding one in the result of searching those adult words is higher than finding them in ordinary keywords, hence, we started by looking at them one by one. This was a time consuming work, however, we believe there was no other way to do it.

The final step is implementing a machine learning algorithm that trains the pre-labeled videos and test the search results in order to find more candidates that we can review later. Since our dataset was smaller, we narrowed down the choices to Naive Bayesian methods. On the other hand, we need to detect multiple categories, thus we used Gaussian Naive Bayesian method in order to detect new algorithms. In the next section we discuss the performance of our system and the problems and the solutions we think they would work.

5.1 Our Performance

We started by a few inappropriate videos. Running the algorithm, the algorithm was able to detect a few more inappropriate videos. We reviewed the results and added them to our database as inappropriate. With the new dataset of inappropriate and appropriate videos available, we repeated the step in order to find more videos. Unfortunately, after a while the algorithm was not able to detect all inappropriate videos, so we had to find new videos by reviewing the dataset randomly. After each time we found a new set of inappropriate videos, the algorithm was able to detect more and more. This did not prove that the algorithm could clean the whole dataset,

however, it gave us the idea that what would happen if we had access to all the inappropriate videos which already have been removed from youtube?

Another challenge we were facing was the limited number of request we could make to the server. This means that we could never crawl more than 5000 videos each day. We believe the more videos we have, a better sampling of youtube videos would be possible, therefore the algorithm would be more accurate.

In the next section, we will discuss the suggestions that could come over the two mentioned issues we were facing.

5.2 Suggestions

We recall that we mainly had two problems:

- Limited number of inappropriate dataset available to us: Our experiments showed that the more inappropriate videos we provided to our algorithm, the more accuracy was achievable. We believe that youtube has a great chance of overcoming this issue, because there are many videos being reported to them on a daily basis and also they have access to all the videos that have been already removed from youtube. We believe that this dataset would work perfectly with this mechanism.

- Limited number of queries to youtube API V3: This limitation kept us back from achieving a more realistic dataset of videos on youtube. However, youtube itself has no limitation and has unlimited access to its own API. Hence we believe that youtube must be able to perform our mechanism in order to achieve the highest possible accuracy.

6 Conclusion

First of all, we discussed the reasons why the presence of a human being part is essential for any architecture detecting inappropriate videos. Then we discussed that it is impossible to define what is an inappropriate video. Furthermore we suggested an interactive architecture, which we believe it would work efficiently on a real dataset. We discussed the challenges and the problems in implementing such architecture and by our experiments, we provided clues that support the idea that youtube can use this architecture to help the automation of detecting inappropriate videos.

References

[1] Youtube shooting: at least three injured and female suspect dead in apparent suicide. https://www.theguardian.com/technology/2018/apr/03/youtube-headquarters-report-shooting-police-san-bruno.

[2] Youtube statistics – 2018. https://merchdope.com/youtube-statistics/.

[3] Youtube stats: Site has 1 billion active users each month. https://www.huffingtonpost.com/2013/03/21/youtube-stats_n_2922543.html.