Detecting child unsafe videos using transcripts

Shritishma Reddy & Pratik Kamble

February, 17th 2019

1 Abstract

In this project, we propose an end-to-end pipeline for the detection of child unsafe videos using video transcripts. Given the spread of digital media and increase in media consumption by kids (ages 3-12) over the past decade, there is a great need for the filtering of unsafe content found within reach of said age group. While traditional parental systems do exist, this would be an online scalable system, that would only benefit from the increasing video consumption.

2 Introduction

This pipeline has three phases:

- Data acquisition: Collection of a holistic video-transcript dataset representative of children viewing habits and trends.
- Unsafe pattern detection: Analyze speech patterns and clustering hotspot zones to form a timeline representation.
- Video Classification: Classification of videos into unsafe types, if unsafe, viz. religious, abusive, drug use, etc.

2.1 Data Acquisition

The data acquisition process depends a lot on the type of genre distribution. A study [3] has found that children consume just over three hours of media, as of 2015; the numbers having since then gone up further.

The time spent with on screen media dramatically increases from the toddler to preschool to school-age years. Children under two have a screen time average of 53 minutes per day. This increases to almost two and a half hours per day among two to four year old and almost three hours for kids in the five to eight year old range. By age eight, 96% of children have watched TV, 90% have used a computer, 81% have played console video games, and 60% have played games or used apps on a portable device. Thus by order of data-source, we should

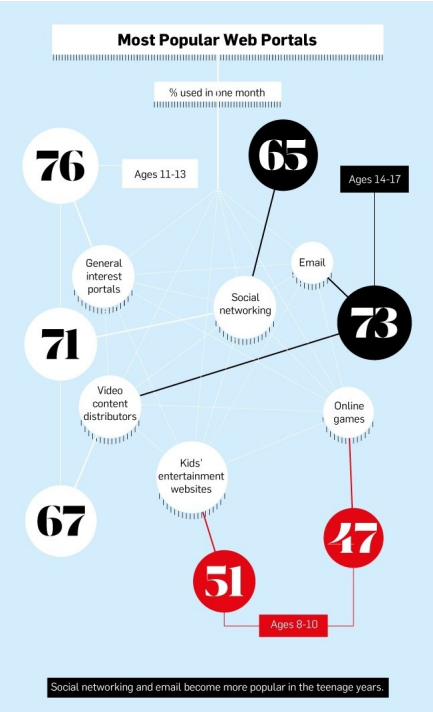

Figure 1: Major content sources and distribution

have a dataset representative mainly of TV show content, followed by YouTube and lastly, game video data.

Luckily, all video footage from various sources can be obtained from YouTube, along with transcript data. TV show snippets, actual online vlog and other consumable media, as well as video game footage can be found on the streaming platform.

2.2 Unsafe Pattern Detection

Once we have the annotated data, we have to run through it and classify according to unsafe-ness. Once criterion that is of importance is categories that may influence minds may be quite different from ones for adolescents or adults.

As seen from the figure , the main categories of media streaming are entertainment websites, omline games, email, social networking and general interest portals. Breaking down the type of content found on these sites, the main type of genres are as follows:

- Religion-based hate speech: This occurs in chatrooms and online forums,

both high target kid influence

- Color-based hate speech: Depending on race and ethnicity, such sort of speech may be hidden in meaning and harder to detect.
- Foul Language: Kids usually pick up on foul language and the age of such exposure is lower this decade than it ever was.
- Other offensive terminology: Dealing with kids in the lower age bracket, systems need not worry too much about false positives, because of the ease of influence of media on early youth.

2.3 Video Classification

Once the data collection for the transcripts is done, a simple classical model is used in cascade, to first analyze and mark scenes and zones in the time frame with dangerous content.

This meta information is then passed to a zonal mapping pipeline, that, depending on how many zones of what clusters there are, estimates a score for each category mentioned prior. Finally, depending on the thresholding for each genre distribution across ages, a weighted average will judge the videos overall scores.

3 Annotation and labeling

The first step in the pipeline is the data collection and annotation. It is important to keep in mind the divisons and genre distribution amidst children and have a dataset representative of that. We are collecting 5 videos in each positive category, approximately of 5 minutes each, unless the genre specifies (eg, vlogs are usually longer). The videos should be at such an age genre balance that younger children usually are pressured into by peers, and boundary videos are easy gateways to explicit content.

This gives us a total of 4 categories of positive videos, for each age group bracket within kids, targetting most common videos in the given set.

$$N = n_{videos} \times n_{categories} \times n_{agegroups} \times m \times \delta \times \alpha$$

Here n_{videos} is 5 per category, $n_{agegroups}$ is 3, m being the average length per video of selection. Now the average rate of speech is δ. This may vary across various demographics, but we choose the global average, 110 words per minute. Let's assume, since we are targetting positive samples to seed the dataset, there is a 20% conversion rate, α. This means that 1 minute in a positive truth, 5 minute video contains positive samples of the content we're trying to flag. This is a fair assumption, and goes in line with average statistics [2].

Once this data is labeled, we're ready to pass it to the speech tagging and sentence marking - both statistics and performance with training as well as well as pre-trained weights has been discussed. This will tag the basic parts of the speech and mark the transcript with labels.

4 Speech Classification

The network that has been used to analyze and tag setence and n-grams as potentially unsafe is discussed below. Because of the relatively simpler nature of detection of unsafeness for children (false positives are beneficial if anything), it is based on majorly classical methods and does not do anything fancy [1] Bi-LSTM based methods. However, a comparison to such methods as the core model has also been added for brevity.

The speech classification module can be broken down into basic components:

- Processing and tokenization
- Feature extraction and supplementing
- Regression & Other feature extraction

5 Zonal clustering and overall classification

Once we have the features extracted and a model trained on the existing dataset, we can start to map out a heatmap of the video timeline so indicate zones of high unsafe-ness clustered on the basis of categories. Once these zones are established in the video, we can supplement it with additional release information and setup an unsupervised secondary pipeline, one that now on the basis of the various genre clusters within a video, an overall unsafeness score and probability is given on the basis of the intravideo analysis.

6 Results

7 Conclusion & further work

References

[1] Bi-lstm based tweet hate speech detection. `https://github.com/sebastiandziadzio/hate-tweet/`.

[2] Biteable. Video statistics 2018. `https://biteable.com/blog/tips/video-marketing-statistics/`, 2018.

[3] Dr. Brent Conrad. Viewing statistics by time. `http://www.techaddiction.ca/media-statistics.html`, 2015.