Slide 1

Detecting Child Unsafe Videos Using Transcripts

Shritishma Reddy & Pratik Kamble

Apr 6, 2019

Slide 2

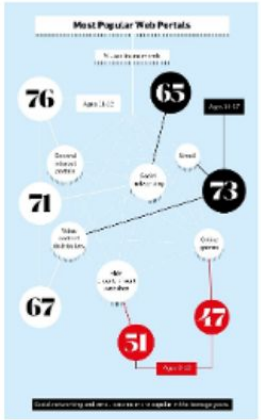
Basic Introduction

This project proposes an end to end pipeline for the detection of child unsafe videos using video transcripts

The pipeline has three phases:
1. Data acquisition: Collection of a holistic video-transcript dataset representative of children viewing habits and trends.
2. Unsafe pattern detection: Analyze speech patterns and clustering hot spot zones to form a timeline representation
3. Video classification: Classification of videos into unsafe types, religious, abusive, drug use, etc.

Slide 3

Data Acquisition



The main categories of media streaming are entertainment websites, online games, email, social networking and general interest portals.

The main type of genres are as follows:
- Religion-based hate speech: found in chat rooms and online forums
- Color-based hate speech: hidden in meaning and harder to detect
- Foul language

Slide 4

Video (Media Content) Classification

- After data collection of videos (media) is identified, a simple classical model is used in cascade to analyze and mark scenes and zones in the time frame with dangerous content
- Above meta-data is passed through a zonal mapping pipeline, where a score is assigned for each category to each media file. Finally depending on the thresholding for each genre distribution across ages, a weighted average is taken to judge the video's overall score.

Slide 5

Video Annotation and Labeling

- Dataset should be representative of divisions and genre distribution amidst children
- We are collecting 5 videos in each positive category, approximately of 5 mins each
- The videos must be representative of age genre balance. Younger children are usually pressured into by peers and boundary videos are easy gateways to explicit content

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

nvideos = 5 per category, nagegroup = 3 per category, m = average length per video of selection, delta = average rate of speech, alpha = conversion rate (we are only looking at positive dataset to seed the dataset)

Slide 6

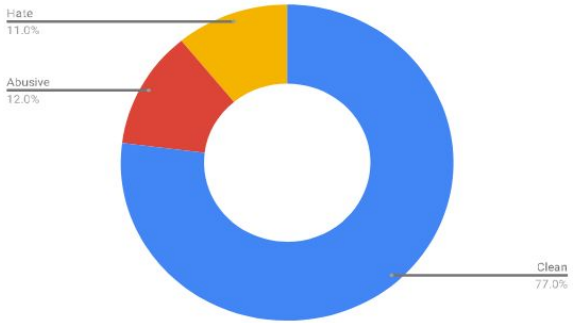
Raw Transcript Collection

- The collected files were formatted using the srt standard and were converted into a *dataset* for annotation.

	Vedio_Id	Genre	start	end	text
0	3tgZ...	0	0.00	3.71	mr. mine you say babe ...
1	3tgZ...	0	3.72	22.79	mr. mine babe what's d...
2	3tgZ...	0	22.79	22.80	of a dynamic punch mr. line
3	3tgZ...	0	22.80	28.40	of a dynamic punch you'll...
4	3tgZ...	0	28.40	28.41	you'll get stronger your

Slide 7

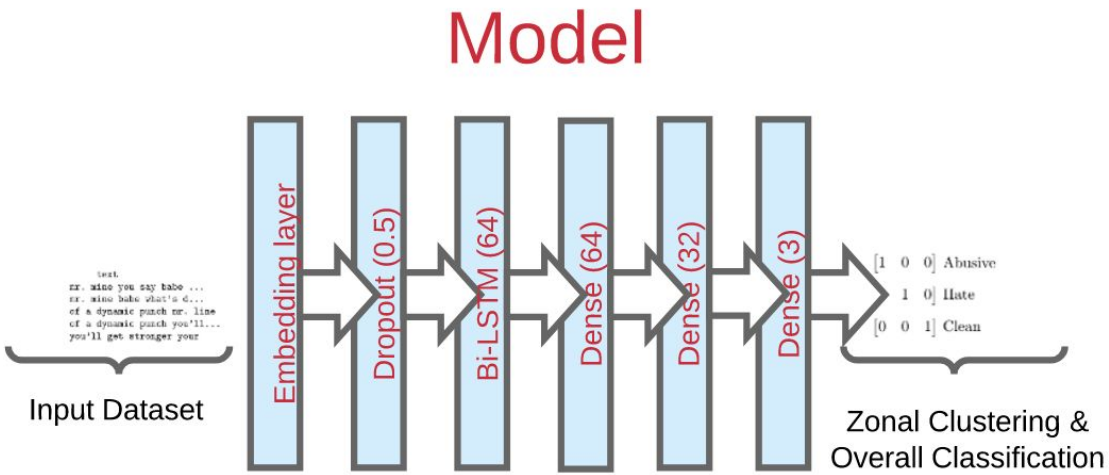
Data Insights



77% of our dataset is "Clean"
12% of our dataset is "Abusive"
11% of our dataset is "Hate"
Average running time of video: 4.5 mins

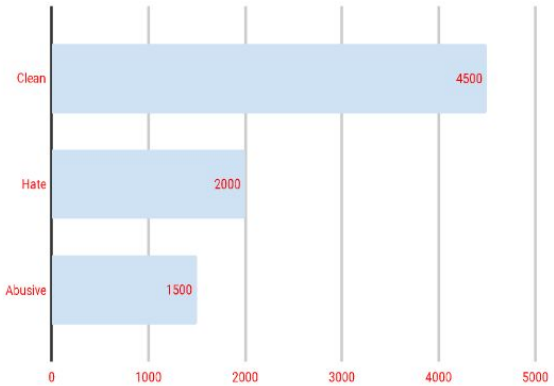
Additionally, some inference can be drawn as to the average "unsafe" duration of a video and the subsequent variation with genres.

Slide 8


Model
text
mr. mine you say babe ...
mr. mine babe what's c...
of a dynamic punch mr. line
of a dynamic punch you'll...
you'll get stronger your
Input Dataset
Embedding layer
Dropout (0.5)
Bi-LSTM (64)
Dense (64)
Dense (32)
Dense (3)
[1 0 0] Abusive
1 0] Hate
[0 0 1] Clean
Zonal Clustering &
Overall Classification

Slide 9

Dataset Preparation



Training Dataset - Distribution across categories

For training: used Davidson dataset and 20% of data corresponding to each label of our Youtube dataset for training

To keep our dataset bias-free, a dataset spanning the categorical distribution shown in the figure on the left are used

Slide 9

Dataset Preparation

Dataset cleaning:
- Lowercased each sentence and stemmed it using the Porter Stemmer.
- Created bigram, unigram and trigram features - each weighted by its tf-idf
- NLTK

Slide 9

Results