NLA PROJECT OUTLINE

Title:

“Detecting Child Unsafe Videos on YouTube Using Transcripts”

Introduction:

According to the statistics released by YouTube, the user base of YouTube is over a billion users and 400 hrs of viewable content is getting uploaded in 60 seconds. YouTube provides ease of publishing and is highly leveraged by content producers to host their video content on YouTube.

Our work mainly deals with detecting unsafe content on Youtube from the transcripts of videos. We focuses on cartoon videos which are typically watched by children. Various cartoon production houses have uploaded their trademark cartoon series on YouTube for wider publicity and brand building. In lieu of our focus on content being watched by kids, we restricted ourselves to collection of cartoon videos only.

Two types of child unsafe content are studied, one, which is sexually explicit and second, which contains violence. By violence we include Abusive Speech too, as our main analysis focuses on text. Put together, and we have three unsafe classes namely sexual, violent and both.

DataSet Collection:

We plan on collecting transcripts from cartoon videos and shonen anime which are mostly targeted for kids under 15 years. Youtube-dl is used to get the transcripts from the youtube video links.

We plan on manually annotating the collecting data and use it for testing. We are looking at possibilities of finding annotated large datasets for inappropriate content with sentence structure similar to the youtube transcripts, so that we can use it for training our networks. In a situation otherwise, we’ll try to collect data and annotate it our self according to classes stated above and use the data set for training our network. We will then test the network using the data we collected from youtube transcripts.

Methodology:

We plan on using sequence model architectures such as LSTM and BLSTM with character grams as inputs for training. BLSTM and LSTM models usually consists of three sequential layers—(a) Embedding Layer (b) Bidirectional LSTM (BLSTM) Layer or LSTM Layer and (c) Fully Connected (FC) layer.

1. To deal with complex structured conversation data, we plan to use character level trigrams to represent each conversation. Character n grams are for words out of the vocabulary. We also use word embeddings like glove fastext etc.
2. Sequence of character trigrams sent through an embedding layer which takes one-hot vector of each character trigram in the sentence and learns the lower dimension representation for it.

3. Bi-Directional LSTMs have the ability to capture much richer sequential patterns from both directions of a sequence.
4. Each character trigram from the previous embedding layer is fed as a sequence of trigrams, to the BLSTM layer.
5. The output of the BLSTM layer is given as input to a Fully Connected (FC) layer which models the interactions between these features.
6. The final softmax node in the FC layer outputs the probability of the sentence belonging to the inappropriate class.

Analysis:

1. We take each input video and identify the timestamp where child unsafe content occurs.
2. We take all such instances where unsafe content occurs in a video.
3. We then cluster videos based on number of occurrences of unsafe content in it and rate it on a scale of 0-10, indicating how unsafe the video is.

Conclusion:

At the end of the project, hopefully we will be able to rate videos from youtube as child unsafe from a scale of 0-10 using text from transcripts of videos as the only input.

Team Members:

D. Shritishma Reddy (20161165)
Pratik Kamble (20161135)