Tasks:

* Read the data collection of the papers
* Milestone due Dec 8
* Gather 2 Datasets: Image frames and Captions
* Gather videos, put in excel
* 50 video each
* ~~Find research papers/literacy material for either idea~~
* ~~Starting proposal~~
* ~~Meet up Sunday~~
* ~~Meeting with Huang on MondayNovember 13 3:15 PM~~

Idea 1: Binary Classification for Youtube Video Captions through Distubred Youtube Classifer for Kids LibraryYoutube caption log: NLP Binary Classification (Safe for kids - 0, Not safe for kids -1).

Input: desired: captions for video, secondary: title for video

Potential Dataset

<https://www.kaggle.com/datasets/rsrishav/youtube-trending-video-dataset>

<https://research.google.com/youtube8m/index.html>

Potential Libraries:

[spaCy · Industrial-strength Natural Language Processing in Python](https://spacy.io/)

<https://www.upgrad.com/blog/python-nlp-libraries-and-applications/> << more NLP models

Youtube method of Tags: <https://www.youtube.com/watch?v=a_6O27R02Ng>

Idea 2: Image Captioning through LLMs with Text to Speech for Biology Students

Image Captioning ( to give image a caption)

Input: Image

Output: Caption + Audio for text to speech

Accessibility This could benefit People who have disabilities such vision impairment

Text to speech

Potential Dataset:

https://www.kaggle.com/datasets/andrewmvd/convid19-x-rays

Ultrasounds

X-rays

If not then general purpose, pose it as a tool for the visually impaired, blind, text to speech.

Potential Library

Or LLM model

[Image captioning (huggingface.co)](https://huggingface.co/docs/transformers/main/en/tasks/image_captioning)

[List of Open Sourced Fine-Tuned Large Language Models (LLM) | by Sung Kim | Medium](https://sungkim11.medium.com/list-of-open-sourced-fine-tuned-large-language-models-llm-8d95a2e0dc76)

Potential Research Papers References: <https://about.biodiversitylibrary.org>

## 

## Proposal Rough Draft

A. Introduction:

1. Which dataset will you use?

2. Why do you select this dataset? Why are you interested in this dataset?

3. Final goals.

B. Dataset: briefly introduce the dataset you select. Is there any data preprocessing or augmentation?

C. Literature review of existing works (More than 2 existing works).

The objective of this literature review is to analyze the paper, “Disturbed YouTube for Kids: Characterizing and Detecting Inappropriate Videos Targeting Young Children” by Papadamou, K., et al, for its methodology and implementation of machine learning to identify and detect patterns on Youtube Videos. The paper’s purpose is to create a deep learning model that improves Youtube’s video filtering process by identifying and detecting inappropriate videos targeting toddlers. This paper supports our project as we share the same goal of using machine learning as a tool to improve Youtube’s filter quality and offer techniques on how to acquire Youtube video data. Their paper uses Youtube Data API, supporting our initial plan to use the tool to extract data needed for our project as features. They look based on title, thumbnail, tags. Also they gathered intel from reddit and got the video seeds through Youtube API.They used a large dataset. They used LSTM and other models and resulted in 84% accuracy. Using recommended videos mentioned on /r/elsaGate/ subreddit and /r/fullcartoonsonline subreddit, they were able to create 2 major categories disturbed, suitable, disturbing, restricted and irrelevant.

The objective of this literature review is to analyze the paper, “Customized video filtering on Youtube” by Vishal Anand et al, for its methodology and implementation of machine learning to identify and detect patterns on Youtube Videos. Their goal was to identify and detect inappropriate videos on Youtube. Similar to (Papadamou K.), they used Youtube Data API to extract features for their model. New approach they did, was using the Youtube videos transcript using youtube-transcript-api.

D. Any existing methods that you plan to use (at least two).

Using Youtube API

Paper 1)’s model

<https://github.com/kwstantinos-papadamou/disturbed-youtube-for-kids-classifier>

Youtube-Transcript-API

<https://pypi.org/project/youtube-transcript-api/>

E. Any novelty or difference from the existing works (Required)?

<https://dl.acm.org/doi/abs/10.1145/3341161.3342913>

<https://link.springer.com/chapter/10.1007/978-3-030-01054-6_21>

F. Experiment/Simulation plans. Evaluation methods

### This project involves using NLP models to determine if a YouTube video is restricted or not, and evaluating its performance is crucial. The model's accuracy, precision, recall, and F1 score will be key metrics for assessment. The confusion matrix will provide a detailed breakdown of true positives, false positives, false negatives, and true negatives. To visually represent the evaluation, ROC and precision-recall curves, along with a confusion matrix heatmap, will be generated. The plan for improvement includes addressing data quality through cleaning and balancing, optimizing model parameters and architecture, exploring different text representations, incorporating additional features like metadata, enhancing interpretability, ensuring robustness through adversarial and out-of-distribution testing, implementing continuous monitoring for performance updates, and considering user feedback for iterative improvements. This comprehensive approach aims to not only assess the current model but also establish a framework for ongoing refinement and optimization to ensure accurate and reliable

G. Reference (Required)

<https://doi.org/10.1609/icwsm.v14i1.7320>

<https://arxiv.org/abs/1911.04013>

11/12/2023 Meeting Notes

Title of paper: Binary Classification for Youtube Video Captions through Disturbed Youtube Classifer for Kids Library

Bianca

* ~~F~~
* Take a look at datasets

Andrea

* ~~E~~
* Datasets
* <https://dl.acm.org/doi/abs/10.1145/3341161.3342913>
* <https://link.springer.com/chapter/10.1007/978-3-030-01054-6_21>

Anya

* ~~C and D for ( 1 ) paper~~
* ~~Find another method/paper~~
* [~~https://arxiv.org/pdf/1911.04013.pdf~~](https://arxiv.org/pdf/1911.04013.pdf)
* https://arxiv.org/abs/1911.04013
* Test this <https://pypi.org/project/youtube-transcript-api/>

Feedback:

Combine 2-3 models ( Language processing, image/ video classification, labeling )

Try to do 2 and combine them

Using existing methods for the 2 taks

Try to find dataset

Download project, send to Huang

11/15/2023

First steps after proposal.

Anya - Figuring out how to do caption scrapping for one video

Andrea - Figuring out how to image slicing for one video

Bianca - Automating both methods that way we have our initial datasets for the two *individual models*

Blurb for Metholody maybe:

The paper’s purpose is to create a deep learning model that improves Youtube’s video filtering process by identifying and detecting inappropriate videos targeting toddlers. This paper supports our project as we share the same goal of using machine learning as a tool to improve Youtube’s filter quality and offer techniques on how to acquire Youtube video data. They uses Youtube Data API, supporting our initial plan to use the tool to extract data needed for our project as features. They look based on title, thumbnail, tags. Also they gathered intel from reddit and got the video seeds through Youtube API. They used a large dataset. They used LSTM and other models and resulted in 84\% accuracy. Using recommended videos mentioned on /r/elsaGate/ subreddit and /r/fullcartoonsonline subreddit, they were able to create 5 major categories: disturbed, suitable, disturbing, restricted and irrelevant. We agree with the

Both K. Papadamou et al (2020) and V. Anand et al (2019) input a collection of inappropriate phrases or phrases linked to inapproriate toddler-targetted video pattern into Youtube Data API to gather video seeds. These seeds were used to extract videos and their meta data. They used metadata such as the thumbnail, title, captions, views, likes and dislikes, etc.

### Youtube Videos Download Tool Link

<https://yt2k.com/en4>

### Useful Youtube Related Links

<https://towardsdatascience.com/making-an-image-dataset-from-youtube-videos-5116252d20a3>

### Video Links

<https://www.youtube.com/watch?v=qi2m4V21bw4>

<https://www.youtube.com/shorts/VzKamrUPbzo>

Youtube Data API Links:

<https://developers.google.com/youtube/v3/docs/videos/list?apix_params=%7B%22part%22%3A%5B%22snippet%2CcontentDetails%2Cstatistics%22%5D%2C%22chart%22%3A%22mostPopular%22%2C%22regionCode%22%3A%22US%22%7D>

### 11/22/23:

Next Steps:

Bianca: Take a look at word embeddings, see if you can get something running for the two videos.

Project Milestone Report: Dec 8

• Midterm report about your final project progress.

• Every group should schedule a meeting with me (15 minutes) to discuss the proposal before the deadline. If you don’t meet with

me, you will get 0 points in the proposal!! All the team members should show up

• Before the meeting, prepare a draft of your Milestone Report!

• Submission:

1. One team should submit one Milestone report of at least 1 page of A4/letter size paper. (IEEE Conference Template for graduate teams).

2. Each member submits one team member evaluation report: describe the percentage of the effort of each team member

in preparing the Milestone report. Describe the contribution of each member.

**• You should finish at least dataset loading, implement at least 1 existing method, and have results to show me.**

**• Here are some examples**

**1. Dataset:**

**1) Data visualization**

**2) Data preprocessing**

**3) Date loading**

**2. Implement at least 1 planned method**

**1) Explain the algorithm you used.**

**2) Show the preliminary evaluation results.**

**3) Analysis of the results and any modifications or plans before the final report deadline.**

### 

### 

### 11/29/23:

Bianca: looking into hugging face for vectorization for NLP

Andrea: looking into yolo to reduce # of frames

Anya: gathering video ids for suitable, disturbed, restricted