

TN POLICE HACKATHON

CybFlagger

**SOCIAL MISCONDUCT DETECTION IN YOUTUBE VIDEOS
WITH ADVANCED AI TECHNIQUES - COMPUTER VISION
AND NATURAL LANGUAGE PROCESSING**

ANNA UNIVERSITY – MIT CAMPUS
Department of Information Technology

TEAM
PROGTOG

Mentor : Dr. G. Rajesh
Members : Mehal Sakthi M S
Sowbarnigaa K S

INTRODUCTION

Social misconduct refers to behavior that violates social norms, values, or expectations, and can cause harm to others or society as a whole. To regulate that behavior, an AI-based tool is used that

- Classifies the metadata information available in the YouTube videos such as Genre, Language, People, etc.,
- Convert the speech-to-text
- Finds the Channel details such as the timestamp of the video posted, personal information such as emails, mobile number, and associated social media profiles of the uploader
- Provides the location details of the uploader of the video or image
- Detects and flags instances of social misconduct in YouTube videos
- Analyses video content and identify potential instances of harassment, hate speech, bullying, and other forms of social misconduct
- Help make YouTube a safer and more inclusive platform by identifying and removing harmful content in a timely and effective manner.

PROBLEM IDENTIFIED

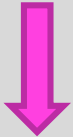
- Social media has created new opportunities for **misusing technology**, which can lead to cyberbullying, spreading false information, addiction, and other negative impacts on users.
- **Harmful content** on YouTube, such as hate speech and harassment, can have negative impacts on users, reinforcing negative stereotypes and biases, promoting intolerance and discrimination, and contributing to mental health problems.
- Develop an AI-based tool for detecting and flagging **social misconduct** in YouTube videos
- The tool is expected to **classify metadata** information, convert **speech-to-text**, find channel details, and analyze video content to identify potential instances of **harassment**, **hate speech**, bullying, and other forms of social misconduct.

Digital Evidence Copy of audio and text (Fully Featured)



Digital Evidence Copy
Date: 2023-03-28
Video URL: <https://www.youtube.com/watch?v=TQM9K5707B4>
Audio Link: /kaggle/working/ytaudio.mp3
Transcript in Original Language: பாசிஸ்டுகள் சாதனைகள் தீவிரவாதிகள் என்றெல்லாம் பேசி பணி சுமத்தினீங்க காவிரியில் தண்ணீர் கேட்கும் போது எங்க பேர் வந்து பிடித்து வைத்துஎங்க படங்கள் ஒன்னு திரையரங்கில் நொறுக்குறது எங்க மக்களுக்கு உயிர் பயத்தை காட்டி ஒரு அச்சுறுத்தலை கொடுப்பது அடிப்பது எங்க பேருந்துகளை கல் எடுத்து வைத்துநொறுக்குவது அதிலே தண்ணீர் அதற்கு சிறுநீர் தரம் குடிங்க என்று எழுதி அனுப்புவது எல்லா பக்கமும் எங்களை அடிக்கிறாங்க ஆனா இந்த நிலத்தில் ஏதாவது நடக்குதுன்னு பாருங்கஎவ்வளவு நேரம் ஆயிடும் இங்க இருக்குற கன்னடர்களை நாங்கள் அடிக்கிறதுக்கு விரட்டறதுக்கு என் பிள்ளைகளை அடிக்கிறான் பேருந்தை உடைக்கிறான் அந்த அரசு யாரையாவது கை செஞ்சிருக்காணு பாருங்க தொடர்ச்சியா இதை சகித்துக் கொண்டே இருப்போம்எதிர்பார்க்கிறது மிகப்பெரிய தவறு எந்த நேரம் திடீர்னு கோவம் வரலாம் அடிக்கலாம்
Transcript in English: When we ask for water in the Cauvery, our people come and hold us while asking for water in the Cauvery. How long will it take to see if something happens?He beats my children to drive them away, he smashes the bus, see if the government has done anything to anyone.
Evaluator Signature:

Inappropriate video (flagged by YouTube)



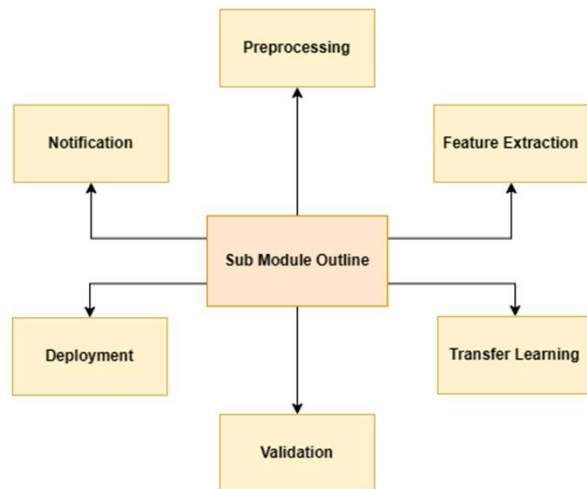
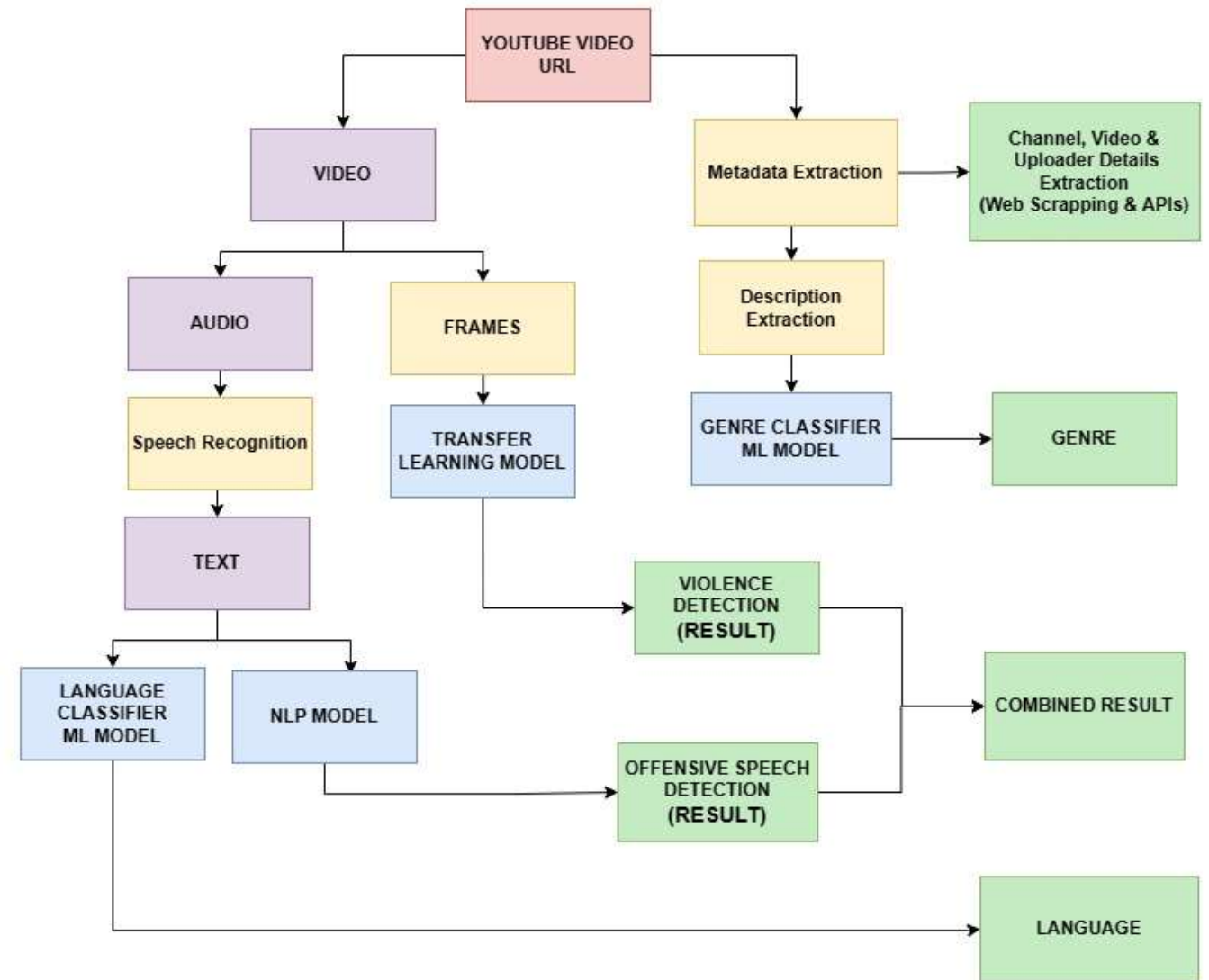
Digital Evidence Copy
Date: 2023-03-28
Video URL: <https://www.youtube.com/watch?v=dLctXWrkFK4>
Audio Link: The Video content has Access Restrictions and has been flagged as inappropriate by Youtube Itself
Transcript in Original Language: The Video content has Access Restrictions and has been flagged as inappropriate by Youtube Itself
Transcript in English: The Video content has Access Restrictions and has been flagged as inappropriate by Youtube Itself
Evaluator Signature:

No recognizable audio



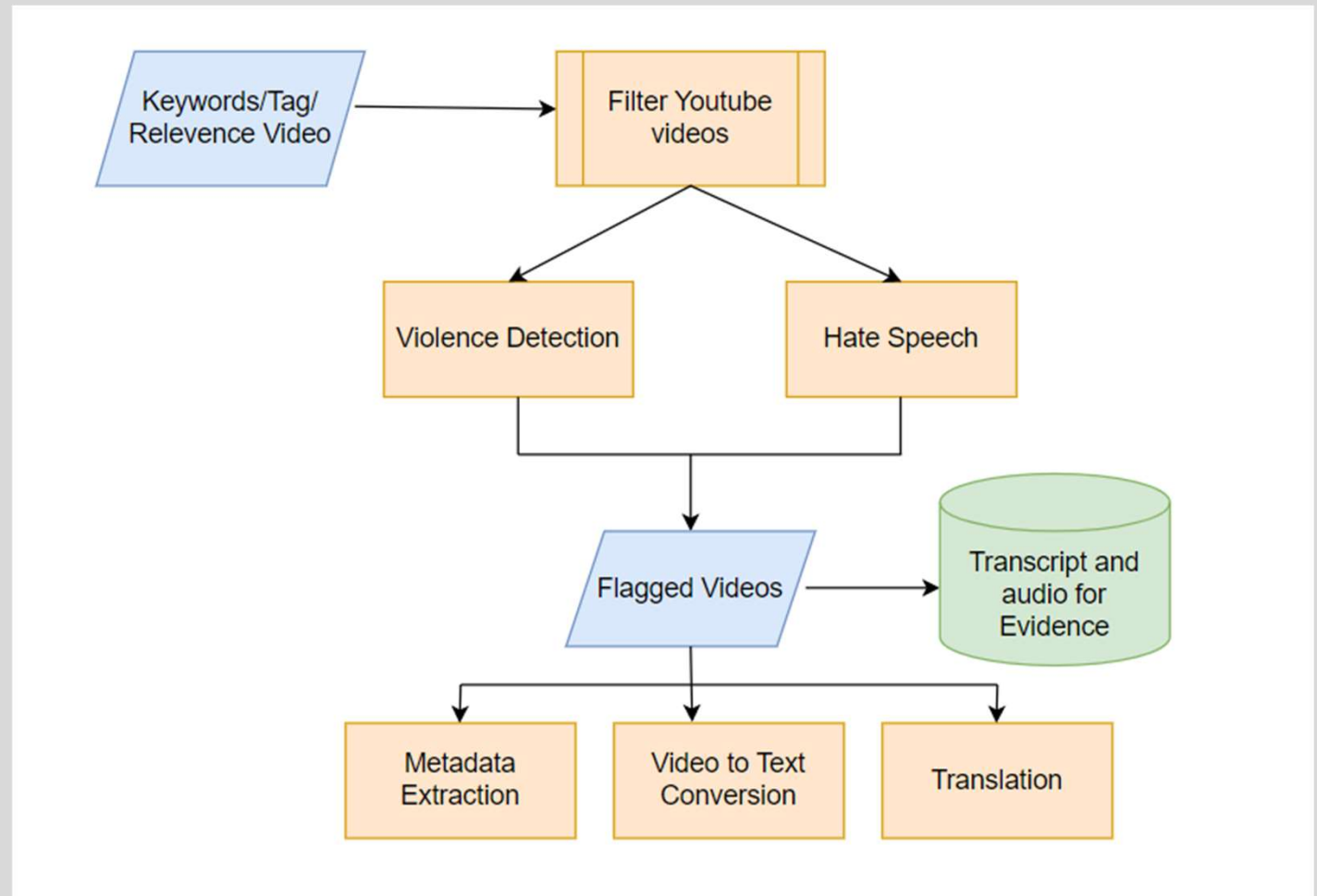
Digital Evidence Copy
Date: 2023-03-28
Video URL: <https://www.youtube.com/watch?v=fCW5vMCovs8>
Audio Link: /kaggle/working/ytaudio.mp3
Transcript in Original Language: No Content
Transcript in English: No Content
Evaluator Signature:

ARCHITECTURE

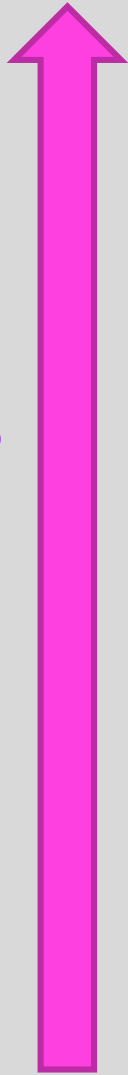


Worked on Multiple videos and extraction of evidence

MODIFIED ARCHITECTURE



OUR
PROGRESS
FROM
PREVIOUS
REVIEW



- Evidence Extraction
- Frontend display
- Integration of all modules
- Support multiple videos
- Compatible to all devices
- Custom dataset
- Fit for Real-time usage by Department Officers



MULTIPLE VIDEOS AS INPUT

Keywords of issue

Relevance video

url_list

```
[ 'https://www.youtube.com/watch?v=fCW5vMCovs8',  
  'https://www.youtube.com/watch?v=TrjbvwJ4dVE',  
  'https://www.youtube.com/watch?v=dHNABMaQsXc',  
  'https://www.youtube.com/watch?v=dLctXWrkFK4',  
  'https://www.youtube.com/watch?v=TQM9K5707B4' ]
```



```
test_pred[0]
```

```
: array([0.29261315, 0.64962023, 0.0577666 ], dtype=float32)
```

+ Code

+ Markdown

VIOLENCE DETECTION

HUGGINGFACE PROTOTYPE

PROGTOG VIOLENCE DETECTION

Model:

microsoft/xclip-base-patch16-zero-shot

Youtube URL

Local File

Youtube URL

Youtube URL:

<https://www.youtube.com/watch?v=s5pNnJCa27Y>

Labels Text:

violence, non violence

Predict



Predictions:

violence

violence 91%

non violence 9%

```
Epoch 1/3
496/496 [=====] - 528s 1s/step - loss: 0.4020 - accuracy: 0.8649 - val_loss: 0.2501 - val_accuracy: 0.9170

Epoch 00001: val_accuracy improved from -inf to 0.91704, saving model to model.h5
Epoch 2/3
496/496 [=====] - 514s 1s/step - loss: 0.2714 - accuracy: 0.9108 - val_loss: 0.2373 - val_accuracy: 0.9143

Epoch 00002: val_accuracy did not improve from 0.91704
Epoch 3/3
496/496 [=====] - 514s 1s/step - loss: 0.2318 - accuracy: 0.9206 - val_loss: 0.2445 - val_accuracy: 0.9183

Epoch 00003: val_accuracy improved from 0.91704 to 0.91831, saving model to model.h5
```

```
x=["How mean you are. I will kill you if these persists. Its a very long day after hearing brutal comments"]

train_input = bert_encode(x, tokenizer, max_len=250)

test_pred = model.predict(train_input)

1/1 [=====] - 0s 27ms/step

test_pred[0]

array([0.19789992, 0.60688466, 0.1952154 ], dtype=float32)
```

HATE SPEECH BERT IMPLEMENTATION

```
x=["kill you idiot. how stupid you are?"]

train_input = bert_encode(x, tokenizer, max_len=250)

test_pred = model.predict(train_input)

1/1 [=====] - 0s 28ms/step

test_pred[0]

array([0.29261315, 0.64962023, 0.0577666 ], dtype=float32)
```

VIDEO TO TEXT MODULE

Output (9.9MB / 19.5GB)

- ▼ /kaggle/working
 - ▶ output
 - ▶ digital_evidences
 - ▶ chunks
 - ytaudio.mp3
 - ytaudiowav.wav
 - violvideo.mp4
 - digital_evidences.zip
 - ytvideo.3gpp

Download Video & Convert to...

Whisper → Language Detection

Audio to Text

Translate

பாசிஸ்டுகள் சாதனைகள் தீவிரவாதிகள் என்றெல்லாம் பேசி பணி சுமத்தினீங்க காவிரியில் தண்ணீர் கேட்கும் போது எங்க பேர் வந்து பிடித்து வைத்துஎங்க படங்கள் ஒன்னு திரையரங்கில் நொறுக்குறது எங்க மக்களுக்கு உயிர் பயத்தை காட்டி ஒரு அச்சுறுத்தலை கொடுப்பது அடிப்பது எங்க பேருந்துகளை கல் எடுத்து வைத்துநொறுக்குவது அதிலே தண்ணீர் அதற்கு சிறுநீர் தரம் குடிங்க என்று எழுதி அனுப்புவது எல்லா பக்கமும் எங்களை அடிக்கிறாங்க ஆனா இந்த நிலத்தில் ஏதாவது நடக்குதுன்னு பாருங்கஎவ்வளவு நேரம் ஆயிடும் இங்க இருக்குற கன்னடர்களை நாங்கள் அடிக்கிறதுக்கு விரட்டறதுக்கு என் பிள்ளைகளை அடிக்கிறான் பேருந்தை உடைக்கிறான் அந்த அரசு யாரையாவது கை செஞ்சிருக்கான்னு பாருங்க தொடர்ச்சியா இதை சகித்துக் கொண்டே இருப்போம்எதிர்பார்க்கிறது மிகப் பெரிய தவறு எந்த நேரம் திடீர்னு கோவம் வரலாம் அடிக்கலாம்

▶ 0:00 / 12:03 — 🔊 ⋮

METADATA EXTRACTION

[CybFlagger](#) [Home](#) [API Model Prototype](#) [Kaggle Notebook](#) [GitHub Repo](#)

URL


Load Video

Return Home

Andhra Pradesh Toll Plaza Attacked ...

Watch Later

Share

Watch on  YouTube

Video details

Title Andhra Pradesh Toll Plaza Attacked By Tamil Nadu Students | Tirupati | English News | #Shorts

Video id fCW5vMCovs8

Description Students of a law college from Tamil Nadu went on a rampage over payment-related issue at a toll plaza in Andhra Pradesh's Tirupati on Monday and attacked the booth staff. #andhrapradesh #tamilnadu #tirupati #clash #fight #tollbooth

Video URL
<https://www.youtube.com/embed/fCW5vMCovs8>

No of Views 647241

No of Likes 11028

No of Comments 423

Published on 2022-10-24T15:49:13Z

Channel details

Channel Name CNN-News18

Channel Id UCef1-8eOpJgud7szVPIZQAQ

Channel Url
<https://www.youtube.com/channel/UCef1-8eOpJgud7szVPIZQAQ>


No of Videos fCW5vMCovs8

No of Subscribers 3600000

No of Views 1204864110

Location India

Uploader Image



INNOVATION

MULTIMODALITY OFFENSIVE SPEECH DETECTION VIOLENCE DETECTION

- The main innovation in the proposed tool is the main 2 modules involved namely, the violence detection and the offensive detection modules that help the society by detecting violence and hate speech from the videos posted in YouTube.
- Such innovation is achieved in addition to satisfying the outcomes of the Given Problem Statement of social media video analytics.
- Another technical in this field is the usage of multimodal NLP
- It improves NLP models by leveraging multiple modes of data and captures a more complete picture of the language - improved accuracy in tasks such as sentiment analysis, named entity recognition, and machine translation
- Another innovative idea is to integrate all modules into a single tool to save time and improve user accessibility
- Existing modules are separate applications built on different data sets, but the proposed tool pre-processes models and intermediate processing states for further computations and analysis

CONCLUSION

- The proposed tool based on multimodal ML DL can improve social misconduct detection in YouTube videos
- Multiple modes of data can be leveraged to capture a more complete picture of the language being analyzed
- Improved accuracy in detecting instances of hate speech and violence can be achieved
- The use of these techniques can help create a safer and more inclusive online community
- The proposed tool overcomes implementation challenges and has high feasibility, contributing significantly to the promotion of responsible online behavior
- With appropriate resources and safeguards in place, the proposed tool can help prevent social misconduct

The background features a light gray central area with abstract, flowing, wavy lines in shades of blue and purple framing the text. The lines are dense and create a sense of movement and depth.

THANK YOU