

Social Media Detoxifier

~ a step towards cybersecurity / against cybercrimes.

Submitted in the partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

IN

B.E – CSE (Internet of Things)

Submitted by:

PRABAL MANHAS

20BCS4513

Under the Supervision of:

SUPERVISORS NAME :

Rajat Tiwari Mam

Department of AIT-CSE

DISCOVER . LEARN . EMPOWER

Outline

- Introduction to Project
- Problem Formulation
- Objectives of the work
- Methodology used
- Results and Outputs
- Conclusion
- Future Scope
- References

Introduction to Project



In order to express oneself, getting insights about what's happening around the world and adaption to new trend almost everyone is excessively using social media platform.

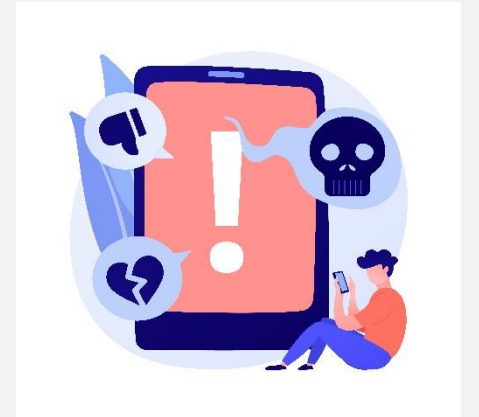
But it comes at cost of mental peace, with increase in number of users, the graph of cybercrimes is seeing an upward trend bullying, harassing and personal threats due to which lot of innocent people have either lost their lives or have suffered mentally.

Introduction to Project (cont.)

So we will be uploading some datasets containing some toxic audio/video messages to classify the toxic elements present in it.

Also we are performing the analysis on real-time twitter data set to fetch out all the toxic elements present in it.

Not only the classification but also it will generate reports about its category i.e. whether the posted comment is toxic, severe toxic, obscene, harassment or personal identity hate.



Introduction to Project (contd.)

The other part of the project is about tracing the scammers/cybercriminals who are performing all these illegal activities using the social media platforms.

The IP tracker part of our proposed project helps us to fetch real time precise location coordinates of the target IP.

Also for the ease of access, we have designed the program in such a way that the user needs not to use google maps or any other GPS service to plot coordinates on maps.

All the entered IP logs will be saved in a HTML file in the project directory which can be downloaded thus increasing the portability.



Introduction to Project (contd.)

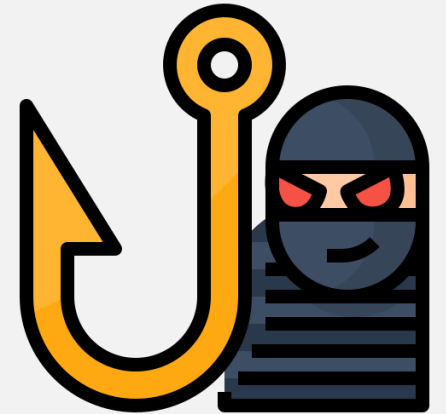
Fetching Phone Number Details –

Using phone numbers python library for making it easy for the users to catch the scammers in real time as well maintaining security of their social media and financial accounts.

The user needs to entered the phone number from which he/she is receiving scam messages or calls, so the program will generate a report containing the required details such as

location, time zone, country, and validity reports of that particular phone number

Based on which we will decide whether it's a authenticated mobile number or a scam mobile number being used for illegal activities.



Problem Formulation

- So our project aim to build a social media detoxifier using which we can analyze toxic audio as well as video message, CSV files or text string containing toxic comments.
- As there is no limit for datasets we can track a bulk amount of data at once and the trained model will automatically perform the toxicity analysis.
- So ultimately this proposed model will be beneficial for organizations where data is generated in bulk, thus saving there time, money and resources.



Problem Formulation (contd.)

- Also with the advent of technology and better internet facilities there is a rapid increase in cybercrime and cyberbullying cases such as scammers, spam callers, bank frauds etc.
- So there is immediate need of the hour to curb these negative elements of social media and technology. So as we described earlier some modules such as IP Tracker and Phone Number details fetcher will be helping us to accomplish the desired goals.



Objectives of the Work



- **Classifying Toxic Audio Files** - To perform toxicity analysis for the uploaded **toxic audio messages** of any format such as .wav,.ogg,.mp3 etc.
- **Classifying Toxic Video Files** - Performing toxicity analysis of **toxic video messages** and extracting its features such as fetching the audio from it then converting into text strings to predict its toxicity value.
- **Fetching Mobile Number Details** - Fetching real time scam mobile numbers details such as **country, timezone, operator service name, validity reports**.
- **Fetching IP Details** - Fetching target IP address, location coordinates.
- **Toxicity Analysis using Detox** - Using **detox** library to first predict the toxic value of a given input data and then further dividing them accordingly into their **respective toxicity category** and finally using matplotlib to generate a **graph** for the same for **better visualization**.

Methodology used

The following methodology will be followed to achieve the objectives defined for proposed research work:

- **Detecting Toxicity from Audio Files** - Making use of some libraries such as **transformers, librosa, Wav2Vec2Tokenizer** for the analyzing the toxic audio files since the data will be converted into numerical format or vector format which can be used by model and also this format can be used to find semantic relationships between the associated words by calculating the difference between those two respective vectors.

Methodology used

- **Detecting Toxicity from Video Files** - Installing the **moviepy** library to performing analysis on **large video files containing toxic messages** or visuals, since moviepy offers various features for video processing such as generating the text string from the uploaded video data set, extracting the audios from video files, etc.
- **Fetching IP Address Location** - Using two popular python libraries “**geocoder**” and “**folium**” helping us to extract the **precise location coordinates** of the entered IP Address of scammers also via folium we will save all the traced **IP logs in a HTML file**, having all the location data, maps data in a vector graphics format containing lanes, route names, location marker etc,

Methodology used

- **Toxicity Comments Analyzer** – Importing libraries being used for Machine Learning process such as **Pandas, Numpy and Matplotlib**. These will help us to train our model to performing toxicity analysis and categorical data analysis i.e. dividing the toxic elements into their respective categories further such as **toxic, severe toxic, personal identity hate, harassment, obscene**.
- **Generating Bar Plots, Histograms** - Also after getting the fetched data we will make use of **matplotlib** to generate **histograms** and **bar plots** of the toxic comments present in our dataset which will the user to analyze and understand the result data in an effective manner.

Methodology used

- **Data Cleaning and Merging** – As we imported the twitter data set containing all the comments text strings with their **id's, usernames, and comment text** etc. so with the increase in data there is also an increase in discrepancies, errors etc. which reduces the accuracy of our designed model.
- So in order to fix this issue we will perform **data cleaning and merging** process using the natural language processing toolkit using which we will create a list of stop words such as me, you, your, about etc. as these words will neither have a positive nor negative result on our data.
- Our aim is to **fetch only the toxic element** present in the posted tweets, or comments. Also we will **remove unnecessary punctuations and symbols**, thereby making our model as much precise and accurate as possible.

Results and Outputs

Audio Analysis - First a **toxic audio message** with a .wav format (*can be of any format .mp3, .ogg etc.*) was created as a part of our project which was containing a message from hacker who is harassing a user and warning him to perform cyberattacks on his family members. So we are making use of **transformers library** by hugging face which helps us to recognize speech audio and hence generate the text strings from it which we will put in the toxicity classifier model to predict the output.

```
print("THE TOXIC 🤬 COMMENTS 💬 FOUND IN YOUR AUDIO SET ARE AS FOLLOWS:\n")  
print(transcriptions)
```

Python

THE TOXIC 🤬 COMMENTS 💬 FOUND IN YOUR AUDIO SET ARE AS FOLLOWS:

HEY YOU IDIOT SUCH A LUSER YOU ARE SHAME ON YOU I AM GOING TO HACK INTO YOUR ACCOUNTS STAY AWAY FROM US
OTHER WISE YOU WILL FACE ADVERSE CONSEQUENCES

Results and Outputs

Video Analysis - In video analysis we used **moviepy** library for extracting toxic elements from uploaded video, as it offers various features for video processing such as **generating text strings**, **extracting audios** etc. Also made use of **ipython.display()** to play the uploaded toxic video in integrated terminal/console so that the user get better insights about the data set on which we are going to perform operations.

```
toxic_video = mp.VideoFileClip(r"toxic_video.mp4")

toxic_video.audio.write_audiofile(r"extracted_audio_message.wav")
display.Audio("extracted_audio_message.wav", autoplay=True)
```

[MoviePy] Writing audio in extracted_audio_message.wav

100%|██████████| 210/210 [00:00<00:00, 1096.49it/s]

[MoviePy] Done.

▶ 0:00 / 0:00 🔊 ⋮

```
In [ ]: from IPython.display import HTML
        from base64 import b64encode
        mp4 = open('toxic_video.mp4','rb').read()
        data_url = "data:video/mp4;base64," + b64encode(mp4).decode()
        HTML("""
        <video width=400 controls>
          <source src="%s" type="video/mp4">
        </video>
        """) % data_url)
```

Out[58]:

Toxic Video Sample

[TOXIC]

*Hey Idiot, this is a warning
for you, you and your family
can be a victim of cyber
attacks, We are going to
hack into your systems this
is a warning to you losers*

Results and Outputs

- **IP Tracking** - We will enter the obtained IP address of the scammer and the moment we will run the program it will generate all the required information which will be displayed to user on output console consisting of precise location coordinates, followed by saving the **traced logs in an HTML file named as my_map.html** which we will be located in the project directory.

<<< 🌐 IP TRACKER 🔍 - MINOR PROJECT - PRABAL MANHAS >>>

+++++

> 🔍 TRACING YOUR ENTERED IP ADDRESS ... PLEASE WAIT ⌚

> FETCHING IP ADDRESS LOCATION COORDINATES 📖

YOUR LONGITUDE & LATITUDE VALUES ARE AS FOLLOWS: 📄

[32.7353, 74.8617]

TRACED IP DETAILS SUCCESFULLY ... STORED IN THE HTML FILE 🌐

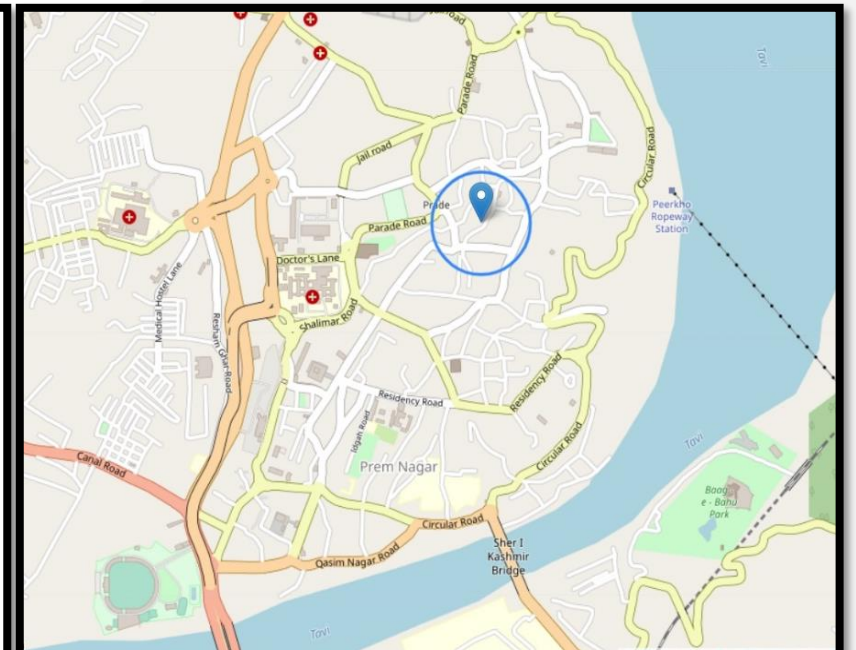
OPEN HTML FILE TO TRACE ON MAP 📖

+++++

PRABAL MANHAS 20BCS4513

ANURAG KUMAR 20BCS4567

GIRJANAND TIWARY 20BCS4506



Results and Outputs


- **Phone Number Tracking** - The user needs to enter the phone number from which he/she is receiving scam messages or calls, so the program will generate a report containing all the required details such as **location, time zone, country, and validity reports** of that particular phone number based on which they can decide whether it's an authenticated mobile number or a spam mobile number being used for illegal activities.

 PLEASE ENTER THE PHONE NUMBER YOU WANT TO TRACE (WITH COUNTRY CODE) ---> +9118001800257

SUCCESSFULLY FETCHED THE DETAILS ... 

 TIMEZONE --> ('Asia/Calcutta',)

 OPERATOR NAME -->

 LOCATION --> India

 CHECKING AUTHENTICITY 

 VALIDITY REPORTS ---> True

Results and Outputs

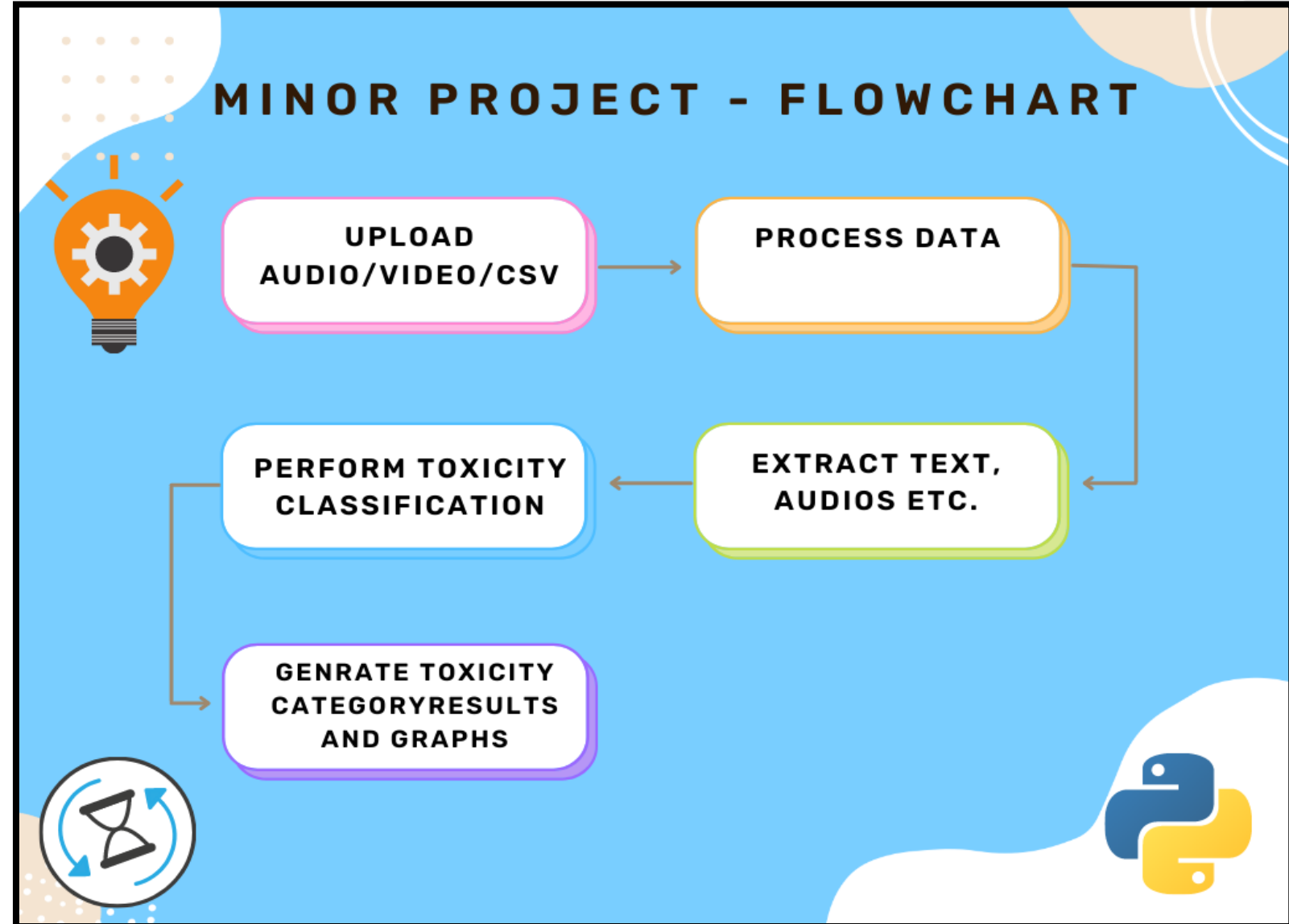
- **Detox / Comments Classification** – This is the final part of our project. Here first of all we will upload a real time dataset of twitter platform consisting of all the tweets saved in a csv file. By using natural language processing toolkits and detox library we carried out desired operations.
- The NLTK will help us in data pre-processing and cleaning/merging part while the detox library will help us to precisely classifying the toxic string/elements and further display its respective category such as **toxic, severe toxic, obscene, personal identity hate, harassment, etc.**

```
label = df[['toxic', 'severe_toxic', 'obscene', 'threat', 'insult', 'identity_hate']]  
print(label.head())  
label = label.to_numpy()
```

	toxic	severe_toxic	obscene	threat	insult	identity_hate
149981	0	0	0	0	0	0
9912	1	0	1	0	1	0
68559	1	0	0	0	0	0
97244	0	0	0	0	0	0
63923	0	0	0	0	0	0

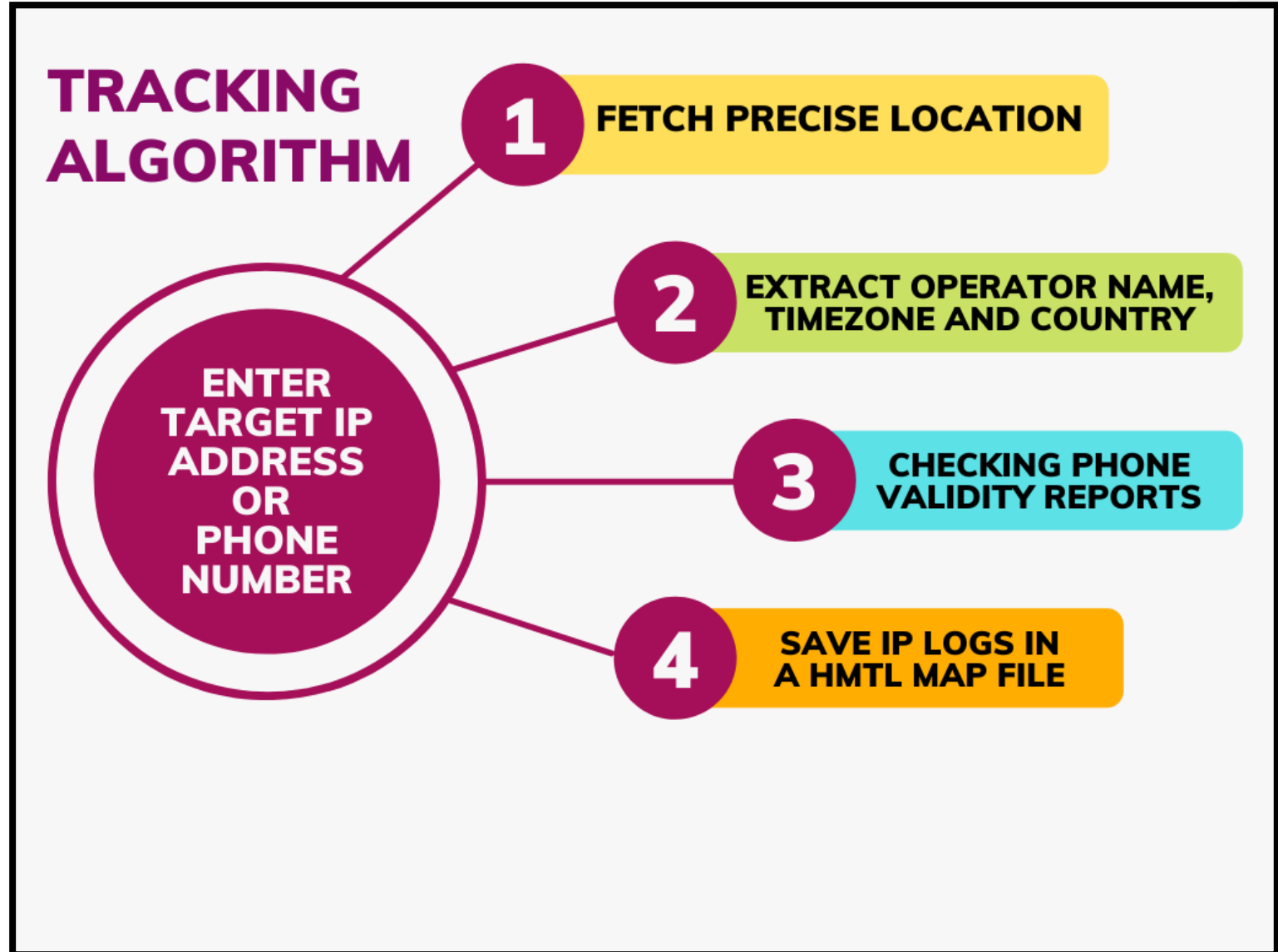
Results and Outputs

Program Flowchart:-



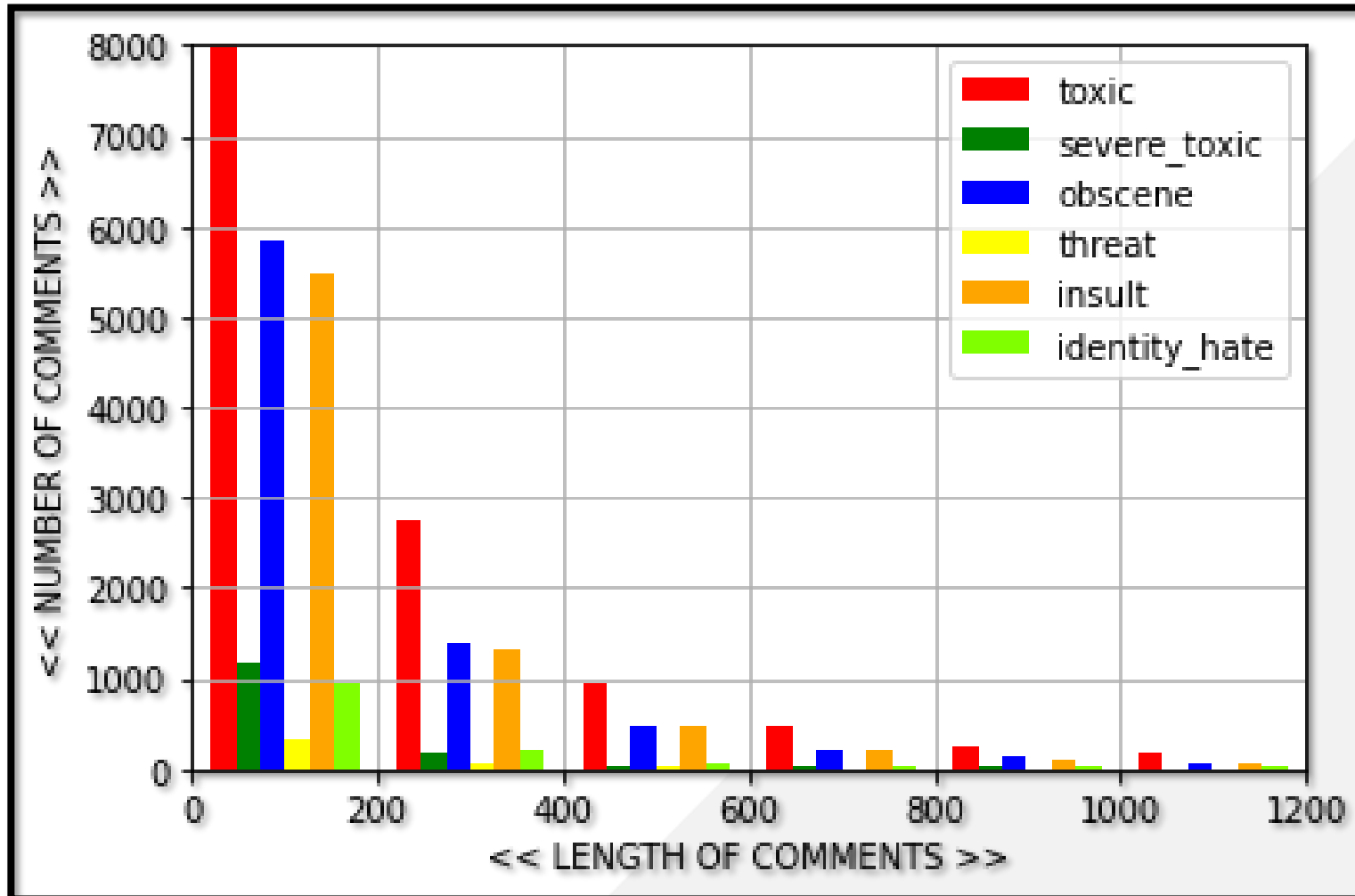
Results and Outputs

Program Algorithm:-



Results and Outputs

Graph for the toxicity analysis :-



Results and Outputs

Comment Toxicity Analysis Report:-

```
sentences = [  
    'You are a liar, shut up',  
    'Hello how are you, good morning',  
    'What the hell you are upto',  
    'Thanks mate, see you soon',  
    'I will hurt you'  
    'Hey tais toi menteur perds toi va en enfer'  
]  
for sentence in sentences:  
    results = predictor.predict(sentence)  
    print (results)
```

```
{'toxicity': 0.9969229, 'severe_toxicity': 0.0042666155, 'obscene': 0.21738318, 'identity_attack': 0.0021789984, 'insult': 0.97  
928965, 'threat': 0.0016220572, 'sexual_explicit': 0.004927032}  
{'toxicity': 0.001624595, 'severe_toxicity': 0.000122038364, 'obscene': 0.0013280034, 'identity_attack': 0.0002733198, 'insul  
t': 0.00097220205, 'threat': 8.832936e-05, 'sexual_explicit': 7.825082e-05}  
{'toxicity': 0.8285392, 'severe_toxicity': 0.0026884545, 'obscene': 0.2458874, 'identity_attack': 0.0019076148, 'insult': 0.073  
71691, 'threat': 0.003725234, 'sexual_explicit': 0.0018394766}  
{'toxicity': 0.00041514577, 'severe_toxicity': 3.7835165e-05, 'obscene': 0.00022453474, 'identity_attack': 7.537777e-05, 'insul  
t': 0.00026801814, 'threat': 4.859589e-05, 'sexual_explicit': 3.2056956e-05}  
{'toxicity': 0.9962993, 'severe_toxicity': 0.03647479, 'obscene': 0.11043426, 'identity_attack': 0.016410543, 'insult': 0.14900  
208, 'threat': 0.07958029, 'sexual_explicit': 0.11128409}
```


Conclusion

- So our proposed project aims at easing the toxic comments, audio, video analysis procedure by automating each and every part.
- Since classification process will be automatically performed by using python and training the model accordingly, thereby save the time and efforts of the companies which will ultimately make the moderation process quite easy and fast.
- Our project is not only limited for text string but it successfully performed toxicity analysis on large toxic audio/video files as well as the CSV files containing the real time comments data.
- Furthermore it is also focused on maintaining the decorum of social media and technology by promoting cybersecurity program such as tracking scammers, IP addresses and fake phone numbers.
- So ultimately all these modules helped us in making a **Social Media Detoxifier**.

Future Scope

- Making this project platform independent and developing an android application for the same to sync the data regularly.
- Getting the latest dataset .csv file of all the social media platforms such as Twitter, Instagram, Facebook etc. at regular intervals of time automatically.
- Building our own server to store the large dataset files in order to increase the processing speed and remove the data storage limit problems.
- Increasing the security features, so allowing only the authorized users to access the data stored on our server, and keeping the suspicious third parties out of our designed model.

References

Kaggle Toxic Comment Classification - <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

IJERT.org - <https://www.ijert.org/research/audio-and-video-toxic-comments-detection-and-classification-IJERTV9IS120099.pdf>

Graphics Elements and Vectorized Templates designed by **Prabal Manhas** using Canva
<https://www.canva.com/templates/>

Wav2Vec2 - https://huggingface.co/docs/transformers/model_doc/wav2vec2

PhoneNumbers Library Documentation - <https://pypi.org/project/phonenumbers/>

References

MoviePy - <https://zulko.github.io/moviepy/>

Detox - <https://pypi.org/project/detox/>

Matplotlib - https://www.w3schools.com/python/matplotlib_pyplot.asp

PNGs in this PPT =

<https://www.pngegg.com/>

<https://www.flaticon.com/>

Platforms Used in the Project-

<https://research.google.com/colaboratory/>

<https://jupyter.org/>

<https://code.visualstudio.com/>