*Annexure-1*

# SOCIAL MEDIA DETOXIFIER

**~ a step towards Cybersecurity / against Cybercrimes**

**A Project Work**

*Submitted in the partial fulfillment for the award of the degree of*

**BACHELOR OF ENGINEERING**

**IN**

**CSE – Internet of Things**


**Submitted by:**

**PRABAL MANHAS**

**20BCS4513**

**Under the Supervision of:**

**Rajat Tiwari Mam**




**DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING**
**APEX INSTITUE OF TECHNOLOGY**

**CHANDIGARH UNIVERSITY, GHARUAN, MOHALI - 140413,**
**PUNJAB**


**Feb-May - 2022**

*Annexure-2*

# DECLARATION

I, **'Prabal Manhas'**, student of **'Bachelor of Engineering in CSE-Internet of Things'**, **session: Feb-May - 2022**, Department of Computer Science and Engineering, Apex Institute of Technology, Chandigarh University, Punjab, hereby declare that the work presented in this Project Work entitled '**Social Media Detoxifier – a step towards Cybersecurity / against Cybercrimes'** is the outcome of our own bona fide work and is correct to the best of our knowledge and this work has been undertaken taking care of Engineering Ethics. It contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

**Date: 16 / May / 2022**
**Place: Chandigarh University (Gharuan)**

**PRABAL MANHAS**
**Candidate UID:20BCS4513**

***Annexure-3*** **(A typical specimen of table of contents)**

# Table of Contents

# ***List of Tables***

## ***List of Figures***

# ***List of Symbols***

## *Symbol Description*

- ***.py -*** *Notation for python files.*
- ***CSV -*** *Notation for comma separated values file. (used as dataset).*
- ***.MP4/.MKV -*** *Video dataset extension.*
- ***.WAV/.OGG -*** *Audio dataset extension.*
- ***HTML -*** *Hypertext Markup Language*
- ***IP -*** *Internet Protocol*
- ***&*** *- and*
- */ - or*
- ***NLTK*** *- Natural Language Toolkit*

# 1 INTRODUCTION

**1.1** 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 when it comes at cost of mental peace, with increase in number of users the graph of cybercrimes is also seeing an upward trend inbullying, harassment and personal threats cases, due to which lot of innocent people have either lost their lives or have suffered mentally.

It is an immediate need of an hour to monitor and detoxify these social media platforms from these scammers, bullies and counter incentives.

So we have designed a project for the same consisting of several modules helping in the detoxification of social media platforms.

**1.1.1** Some of the objectives of our project are toxic audio/video messages classification, fetching precise location of IP Address, and Phone Number tracker. Also we have taken twitter data set consisting all the real time tweets containing the comments, and the goal is to detect the toxic elements present in the posted tweets and further dividing them into their respective toxicity category. So that the concerned authorities can take the required action against these people to curb the occurrence and growth of negative influences.

**1.1.2** As classification process will be automatically performed by using python and training the model accordingly, thereby saving the time and efforts of the companies which will ultimately make the moderation process quite easy and fast.

# 2 LITERATURE REVIEW

So the different methodologies that we used for precise classification of toxic elements are:-

- One of the widely used library i.e. **Transformers by Hugging Face** helps us to train the data set and state of pre trained models thereby reducing cost, time and computer resources. This model can be used for the analysis of different modalities such as text classification, extracting information, text generating in various languages, image classification segmentation, as well as for audio classification i.e. recognizing speech.

- **Wave2Vec2 -** This library is used for speech recognition, that we will use to analyze the uploaded toxic audio message and it's improved automatic speech recognition for many more languages and domains with much less annotated data but State of the art results

- **Natural Language Processing Toolkit -** This performs a major role in the toxicity classification by providing more than 50 lexical resources such as WordNet, StopWords etc. along with text classification, tokenization, stemming semantic reasoning libraries.

- **MoviePy-** This library is used for analyzing the video messages as it offers various features for video processing and helping us to generate text string from the uploaded video dataset.

- **Geocoder -** We are using this library for the IP Tracking part, it makes it easy for locating the coordinates, cities, landmarks for the entered IP address.

- **Folium -** It makes the data visualization easier for us that is manipulated in Python as well as providing us the vector/HTML visualization options which can be used for marker on the Map.

- **Phonenumbers** - This is one of the popular library in Python which will be helping us to fetch several details such as geolocation, operator name, time zone, country and the validity reports just by entering the phone number.

## 2.1 Literature Review Summary

**Table 2.1: Literature review summary**

| Year and citation | Article Title | Purpose of study | Tools / Software Used | Comparison of technique | Source (Journal / Conference) | Findings | Data set (if used) | Evaluation parameters |
|---|---|---|---|---|---|---|---|---|
| 2020 | Audio and Video Toxic Comments Detection and Classification | Creating a word embedding mechanism that can help to identify the slang or negative terms in the comment. | Convolution Neural Network (CNN)<br><br>Recurrent neural network (RNN)<br><br>Natural Language Processing & Supervised Machine Learning (ML) | The designed model technique is compared to pre build libraries and models such as Word2Vec and GLOVE model. | International Journal of Engineering Research & Technology (IJERT) | Toxic comments classification using NLTK | Kaggle Twitter Comments Data Set | CNN model to perform classification process by training the model with training data to fit and test to model to evaluate the accuracy rate. |
| 2022 | Social Media Detoxifier<br><br>~ a step towards cybersecurity | Creating an automated multilingual toxicity classifier performing analysis on any type of data in bulk i.e. audios, videos or CSV files. | Python, Transformers, moviepy, NLTK, VS Code. Jupyter Notebook, Canva | Unlike the previous models this model is applicable for any type of input fata may it be audio, videos or CSV files of any social media platform. | Kaggle Toxicity Classifier | Multilingual Toxic Audio/Video/CSV files and Text Classification.<br><br>Real Time CSV file containing all the tweets, comments etc . | Twitter Real Time Tweets Dataset. | Performing toxicity analysis and plotting the fetched results in graphical format using matplotlib, and testing the accuracy based on regression model. |

## 3 PROBLEM FORMULATION

As we know that Cybercrimes and Cyberbullying cases are taking place at a rapid pace nowadays, which leads to several hate comments, threats to someone's personal life, fake accounts, bots, all taking place with the help of social media.

So our project aims at building a Social Media Detoxifier by which we can automatically track the hate comments, fetch the hate speech audio messages, as well as the fake accounts and bots being used for illegal activities and further restricting & automatically reporting them to their concerned authorities.

Thereby maintaining the decorum of all the social media platforms such as FB, Insta or Twitter etc. making it a safe environment for every user.

Therefore we are building a model to perform multilingual toxic comments analysis along with other features for toxicity analysis.

Unlike the previous models our trained model not also performs toxicity analysis only on the text data but also on large audio/video files consisting of toxic elements. Also the automation process will take this project to next level because it will be quite beneficial for organisation where data is generated in bulk and therefore saving their time and resources.

## 4 RESEARCH OBJECTIVES

The 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, therefore it will save the time and efforts of the companies which will ultimately make the moderation process quite easy and fast.

**The proposed aim will be achieved by dividing the work into following objectives:**

**4.1** To read the data in bulk amount which will be further analyzed by our automated toxicity classifier model.

**4.2** To not only perform analysis on text strings, but also on large audio/video dataset files and CSV files containing the real time monitoring data of any social media platform (Twitter) in our project example.

**4.3** Also tracking the real time location of scammers using their fetched IP addresses.

**4.4** Fetching the desired reports such as their time zone, country, operator names, and validity reports of the phone numbers which are being used by scammers to perform cyber attacks on users. These details are quite beneficial for the tracking process.

# 5 METHODOLOGY

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

**5.1 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.

**5.2 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.

**5.3 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 vectorized graphics format containing lanes, route names, location marker etc.

**5.4 Fetching Phone Number Details –** Using phonenumbers 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 spam mobile number being used for illegal activities.

**5.5 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 for 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.

**5.6 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 help the user to analyze and understand the resultant data in an effective manner.

**5.7 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 (NLTK) 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 impact on our model.

Our aim is to fetch only the toxic elements 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.

# MODEL WORKING – FLOWCHARTS & ALGORITHM

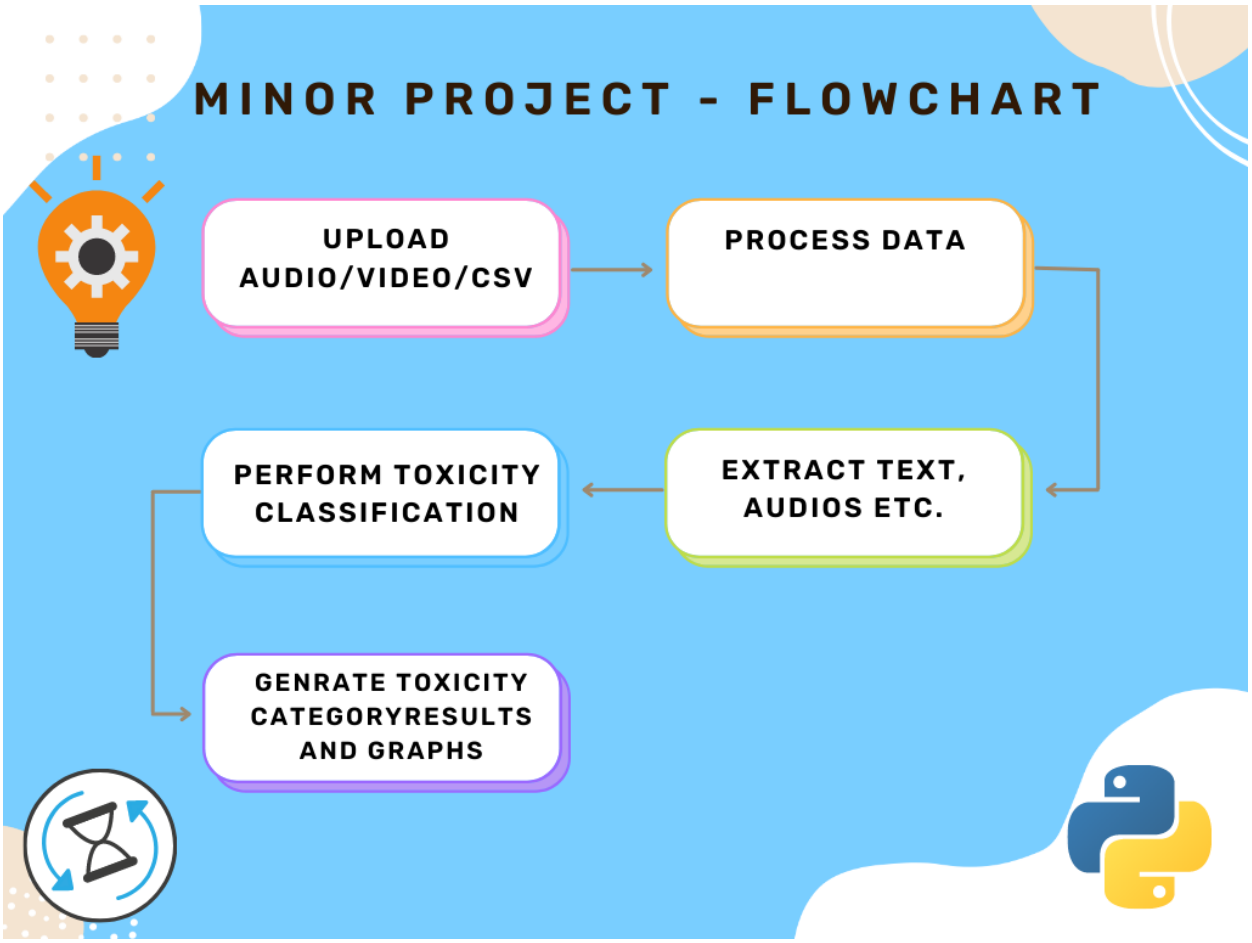



Figure 1 – Flowchart

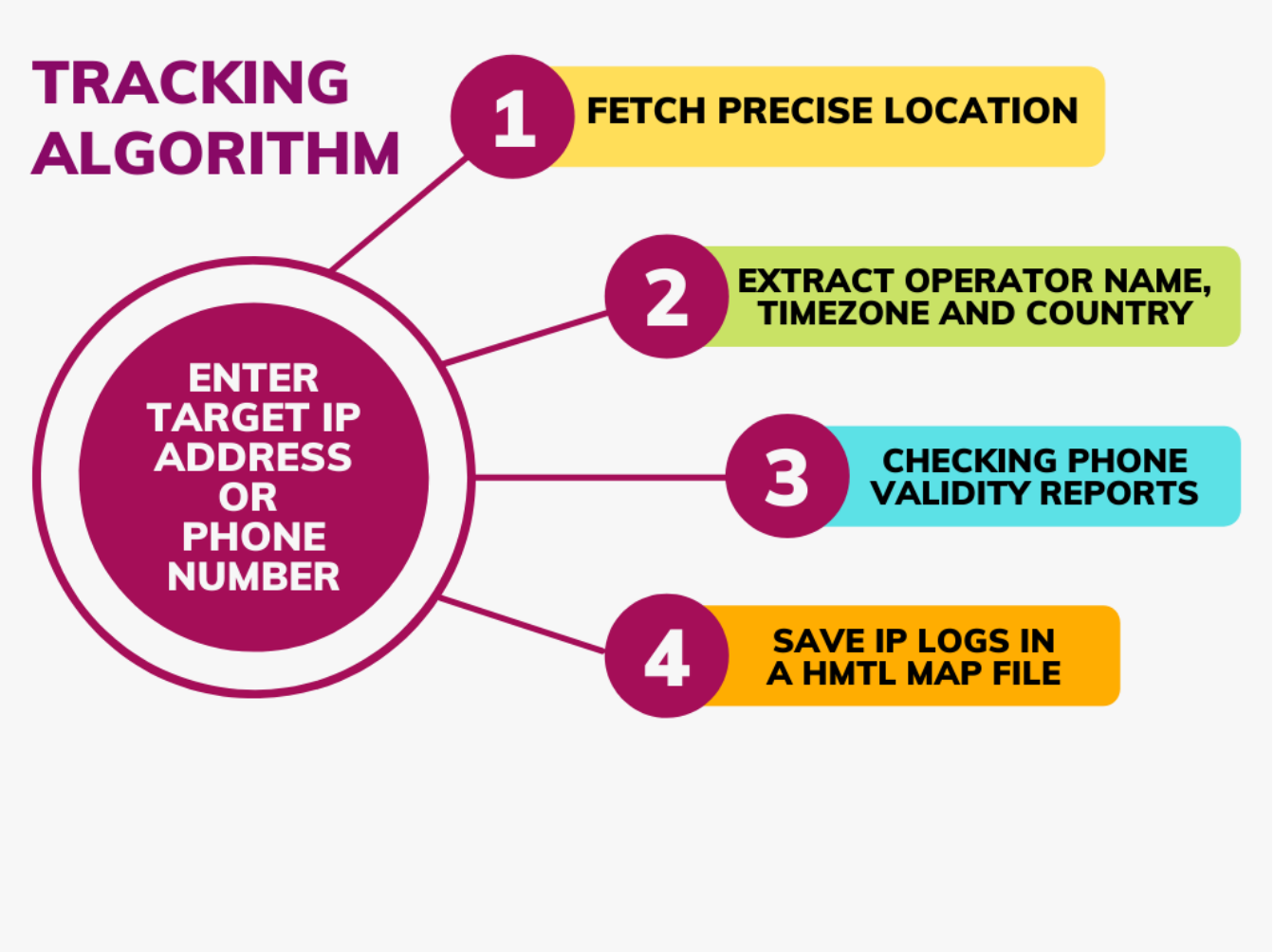



Figure 2 – Tracking Algorithm

## 6 RESULTS AND DISCUSSION

So our designed social media detoxifier model was designed with an aim to perform analysis of a bulk amount of data and also automate the tasks of each and every module i.e. audio toxicity analysis, video toxicity analysis, toxic comments classification, performing data sets cleaning and trimming, and phone number and IP addresses tracking.

- **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 using 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

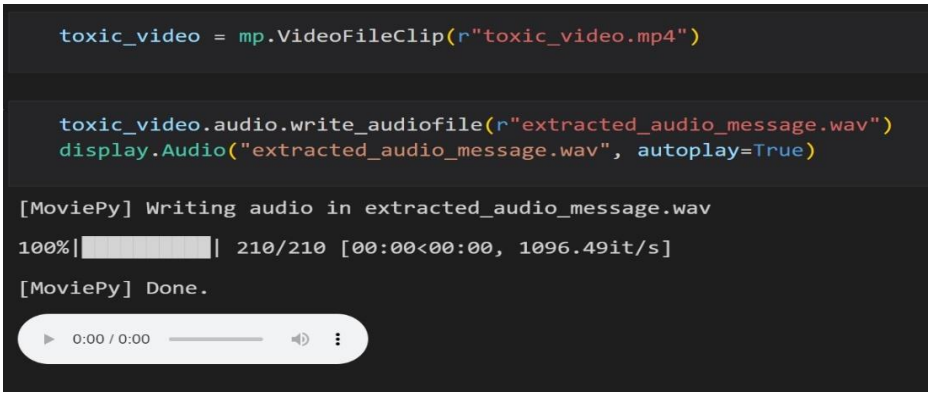
```
THE TOXIC 😡 COMMENTS 💬 FOUND IN YOUR AUDIO SET ARE AS FOLLOWS:

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
```

Figure 3 - Audio Analysis

- **Video Analysis -** Secondly, in the other part of our project we are also performing our analysis on the large videos files with any format may it be .mp4, .mkv etc. containing some toxic images, logos, and sounds.


So our model is capable to analyze the large toxic video files also, we are making use of **moviepy** library containing various video pre-processing features for this which will help us to extract the video features like visuals, text strings, generating and saving the fetched audio files also.

Also we made use of **ipython.display()** to play all the uploaded audio/video files in the integrated google colab or jupyter notebook console so the user gets better insight of the program and datasets on which all the analysis part will take place.



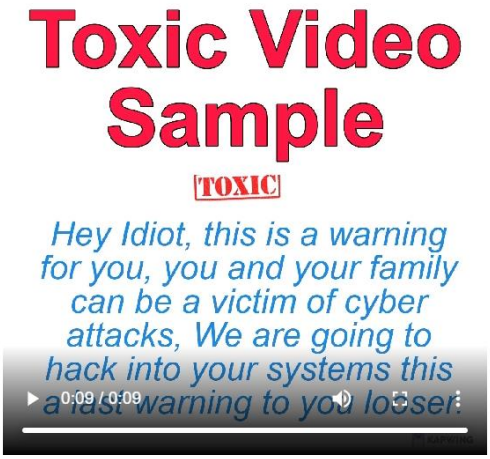




Figure 4 - Video Analysis

- **IP Tracking -** Talking about tracking the IP address location, we used the geocoder library for locating the coordinates, cities, landmarks for the entered IP address. And for the data visualization folium will make the data visualization easier for us that is manipulated in Python as well as providing us the vector/HTML visualization options which can be used for marker on the Map.

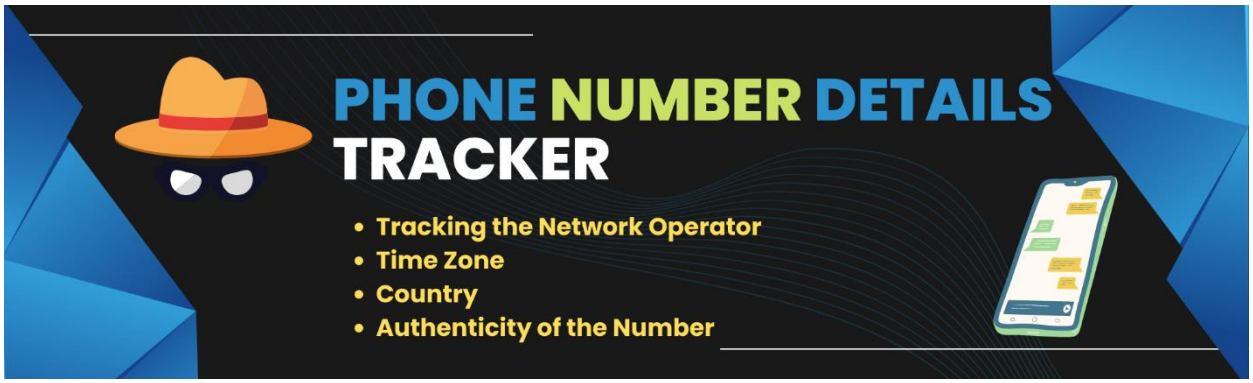
  We will enter the obtained IP address of the scammer and the moment we will run the program it will generate all the required information which will be display to user on output console consisting of precise location coordinates, followed by saving the traced logs in an HTML file named as my_map.html which we will be located in the project directory, so that the user can access it anytime anywhere thus increasing the project portability.


```
                <<< 🌐 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
```

Figure 5 – IP Address Tracker

- **Phone Number Tracking -** Using phonenumbers 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 spam 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 a 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

SUCCESFULLY FETCHED THE DETAILS ... 📱🔍

🗺️ TIMEZONE -->  ('Asia/Calcutta',)
🌐 OPERATOR NAME -->
🏠 LOCATION --> India

✅ CHECKING AUTHENTICTY .... 📞

📊 VALIDITY REPORTS --->  True
```

Figure 6 – Phone Number Tracker

- **Detox / Comments Classification –** This is the final part of our project. Here first of all we will uploaded 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
```

```
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}
```

**Figure 7 - Toxicity Analyzer**

- **Data Visualization -** Last but not the least, since picture speaks more than words so for better understanding and visualization of result we made use of matplotlib to generate a graph on which we will plot all the comments with the toxicity category – and assign heading and colours to each category.


**Figure 8 - Data Visualisation**

The length of the toxic comments posted will be plotted on the x axis whereas the number of comments posted belonging to that particular category will be plotted on the y-axis.

***The image of the graph for the same is given below:-***

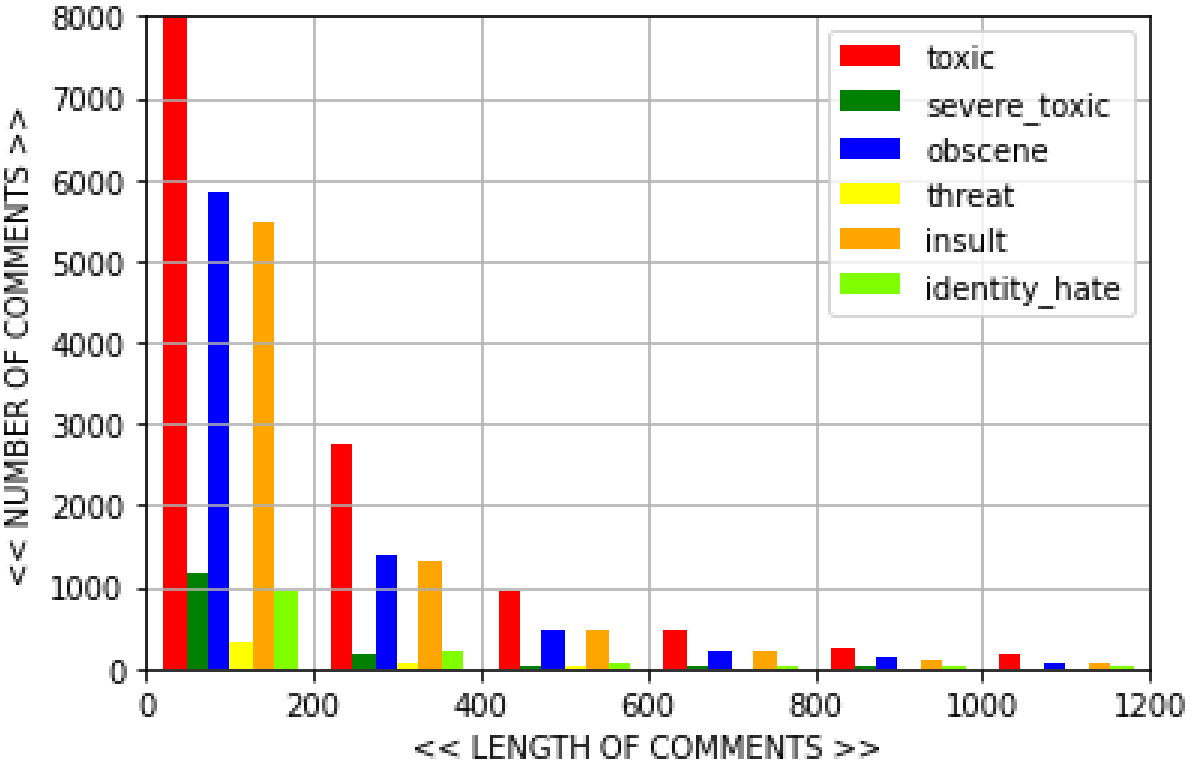



Figure 9 – Graph Analysis

## 7 REFERENCES -


[1] Kaggle Toxic Comment Classification - https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[2] IJERT.org - https://www.ijert.org/research/audio-and-video-toxic-comments-detection-and-classification-IJERTV9IS120099.pdf

[3] Graphics Elements and Vectorized Templates designed by Prabal Manhas using Canva https://www.canva.com/templates/

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

[5] PhoneNumbers Library Documentation - https://pypi.org/project/phonenumbers/

[6] MoviePy - https://zulko.github.io/moviepy/

[7] Detox - https://pypi.org/project/detox/

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

[9] Platforms Used in the Project-

[10] https://research.google.com/colaboratory/

[11] https://jupyter.org/

[12] https://code.visualstudio.com/