

# Cyber Harassment Detail Analysis and Prediction Using Machine Learning

<sup>1</sup>Bhamidipati Venkata Sai Teja, <sup>2</sup>P. Rama Krishna

<sup>1,2</sup>Dept. of CSE, KIET, Kakinada, AP, India

## Abstract

Cyberbullying or cyber harassment is a form of bullying or harassment using electronic means. Cyberbullying and cyber harassment are also known as online bullying. It has become increasingly common, especially among teenagers, as the digital sphere has expanded and technology has advanced. In this research, we have addressed the problem of cyberbullying detection on Twitter Data Set. Various inbuilt Multiclass Classification Algorithms in Python such as Naive Bayes, Decision tree, Random Forest, K Nearest Neighbor (KNN), Support Vector Machine (SVM), and Natural Language Processing (NLP) techniques are used to classify bullying and non-bullying tweets and performance of these algorithms is compared and model with highest accuracy is selected for the prediction. Bullying tweets will be reported to the nearest Cyber Crime Branch.

## Keywords

Cyberbullying, Machine Learning, Twitter, Multiclass Classification

## I. Introduction

Cyber-bullying is an aggressive act that can be carried out by a single person or a group of aggressive people, using online platforms, repeatedly against a person who is not capable of defending himself. Twitter is one of the social networking services on which users communicate using tweets. Cyber-bullying is especially present on Twitter according to the survey done by Pew Center, and Twitter users face many forms of bullying such as death threats, stalking and sexual abusive threats.

### A. What is Multi-Class Classification?

Multi-class classification refers to those classification tasks that have more than two class labels.

Examples include:

- Face classification.
- Plant species classification.
- Optical character recognition.

Unlike binary classification, multi-class classification does not have the notion of normal and abnormal outcomes. Instead, examples are classified as belonging to one among a range of known classes.

The number of class labels may be very large on some problems. For example, a model may predict a photo as belonging to one among thousands or tens of thousands of faces in a face recognition system.

Problems that involve predicting a sequence of words, such as text translation models, may also be considered a special type of multi-class classification. Each word in the sequence of words to be predicted involves a multi-class classification where the size of the vocabulary defines the number of possible classes that

may be predicted and could be tens or hundreds of thousands of words in size.

In this research, we have proposed the model for cyberbullying detection on Twitter Platform using various Multiclass

Classification Algorithms such as Naive Bayes, Decision tree, Random Forest, K Nearest Neighbor (KNN), Support Vector Machine (SVM), and Natural Language Processing (NLP) techniques to classify bullying and non-bullying tweets and report them to nearest cybercrime branch.

## II. Related Work

### A. Cyberbullying Detection: A Comparative Analysis of Twitter Data [1]:

In this paper, authors have attempted to conduct analysis on "tweets" using various inbuilt machine learning algorithms in python. Author had attempted to classify the tweets as bullying and non-bullying. Supervised machine learning techniques like Linear regression, logistic regression, Naive Bayes, SVM, Decision trees and neural networks are used.

The author has used feature extraction techniques such as Bag of Words (BOW) and TF-IDF. Then the word cloud is used to visualize the most frequent words in tweets. The dataset consists of 18,000 tweets out of which 75% were used for training the model and the rest 25% were used for testing the model. The confusion matrix and bar plot were used to visualize the performance of the various algorithms. When recall is high and precision is low, most of the bullying tweets were correctly identified as bullying tweets and few non-bullying tweets were identified as bullying tweets. Out of the all algorithms the SVM has performed better with TF-IDF and Bag of Words as a feature extraction method.

### B. Cyber-Bullying Detection using Machine Learning Algorithms [2]

Authors have selected two distinct datasets, which are recently published, related to the social network FormSpring.me and YouTube.

In this work authors have checked the density of "bad" words as a single feature. This feature is equivalent to the number of bad words that appear in a sentence, for each severity level, divided by the words in the same sentence.

Badness of the word: Authors also have added a feature to their work in order to measure the overall "badness" of a text. This feature is computed by taking a weighted average of the "bad" words (weighted by a severity assigned).

Density of upper case letters: This feature is based on Dadvar et al. [4] results. The presence of capital letters in a text message is selected as a feature, considering it as possible 'shouting' at someone behavior, as commonly treated in social networks netiquette.

This feature is given by the ratio between the number of uppercase letters and the length (number of chars) of the whole sentence.

### Exclamations and questions marks:

Just like capital letters, also exclamation points and question marks can be considered as emotional comments. Authors just stated that cyber bullying is related to an extreme case of sentiment analysis and so it can be connected to the strong (usually bad) emotions. With this premise, authors consider it helpful to introduce the number of exclamation points and question marks as a feature in their work.

### C. Comparative Study of Cyberbullying Detection using different Machine Learning Algorithms [3]

In this research, authors have focused on the detection of cyberbullying using the Instagram data set, with help of four different machine learning algorithms viz. Naives Bayes, SVM, Decision tree and logistic regression. Authors have found that comparatively Naive Bayes algorithms with trigram have the highest performance of 79% accuracy than other three algorithms.

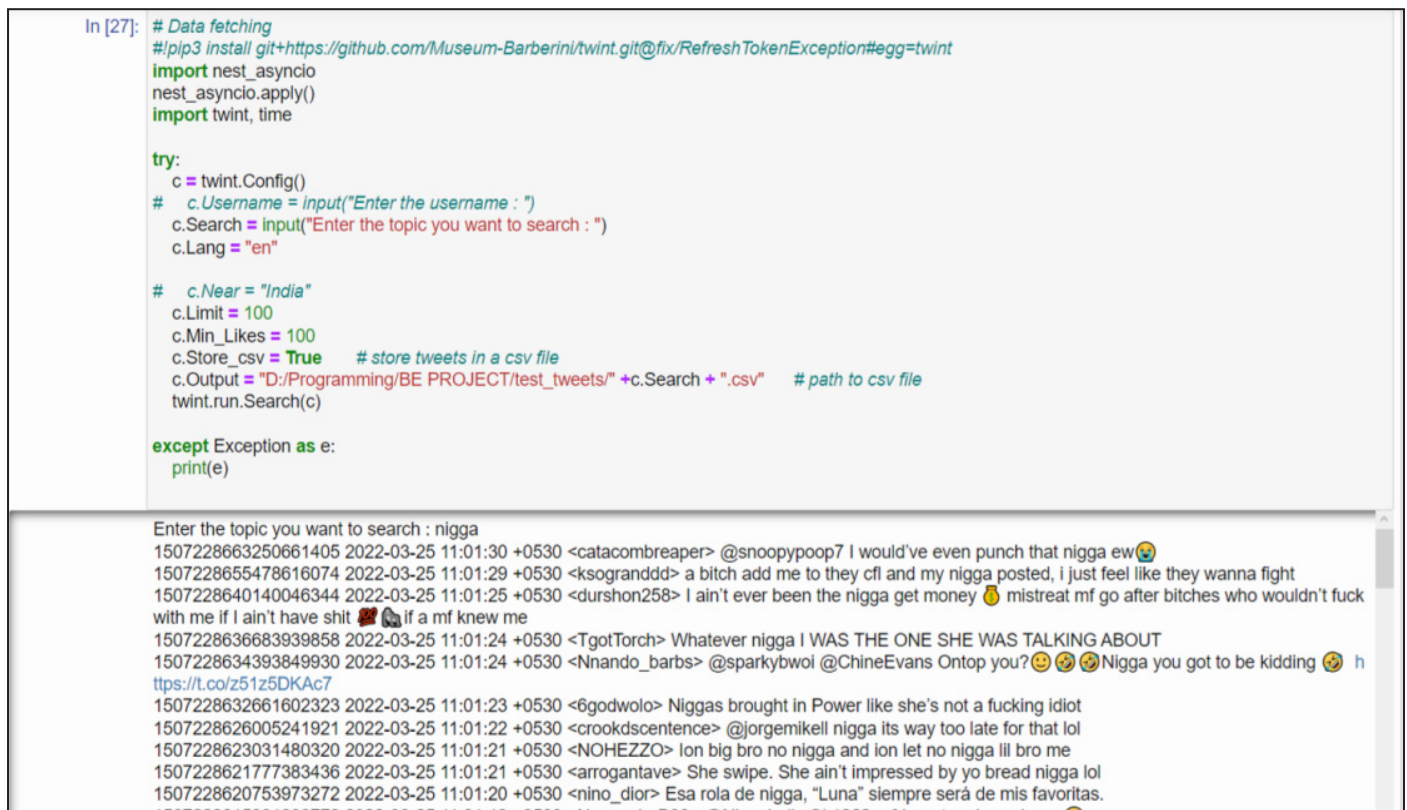
The data collected was based on a set of keywords which includes Muslim, god, tattoo, pray, Hindu, bjp, etc. Authors have chosen these keywords by analyzing different profiles and found that these are the most commonly used keywords. The extracted fields

of a profile from Instagram include the entire posts of the profile along with captions, hashtags, created time and the biography of the user and comments related with each post. The biography has an important role in ease filtering of data i.e., if bio contains words like quotes, memes, awareness, motivation etc. then we can avoid those profiles in the initial

Authors found out that these kinds of public pages will have more followers than followed by count. These points were used to filter the public pages like motivation pages or awareness pages etc. The authors took the comments of each post and labeled it manually as either bullying or non-bullying. A total of 1065 comments were considered and in which 636 were non bullying comments and the following were bullying one. Authors divided the entire data set as training as well as test data and each contains 746 and 319 respectively. During each algorithm call for every n-gram combination the authors have shuffled the dataset.

### III. Proposed Methodology

The working of the proposed system is the same as the existing system[1] except that for collection of twitter data we are using OSINT (open source intelligence) project "twint".



```
In [27]: # Data fetching
#!pip3 install git+https://github.com/Museum-Barberini/twint.git@fix/RefreshTokenException#egg=twint
import nest_asyncio
nest_asyncio.apply()
import twint, time

try:
    c = twint.Config()
    # c.Username = input("Enter the username : ")
    c.Search = input("Enter the topic you want to search : ")
    c.Lang = "en"

    # c.Near = "India"
    c.Limit = 100
    c.Min_Likes = 100
    c.Store_csv = True # store tweets in a csv file
    c.Output = "D:/Programming/BE PROJECT/test_tweets/" + c.Search + ".csv" # path to csv file
    twint.run.Search(c)

except Exception as e:
    print(e)
```

```
Enter the topic you want to search : nigga
1507228663250661405 2022-03-25 11:01:30 +0530 <catacombreaper> @snoopyoop7 I would've even punch that nigga ew 🤢
1507228655478616074 2022-03-25 11:01:29 +0530 <ksogranddd> a bitch add me to they cfl and my nigga posted, i just feel like they wanna fight
1507228640140046344 2022-03-25 11:01:25 +0530 <durshon258> I ain't ever been the nigga get money 🤑 mistreat mf go after bitches who wouldn't fuck
with me if I ain't have shit 🤢🤢 if a mf knew me
1507228636683939858 2022-03-25 11:01:24 +0530 <TgotTorch> Whatever nigga I WAS THE ONE SHE WAS TALKING ABOUT
1507228634393849930 2022-03-25 11:01:24 +0530 <Nnando_barbs> @sparkybwoi @ChineEvans Ontop you? 🤔🤔🤔 Nigga you got to be kidding 🤔 h
tps://t.co/z51z5DKAc7
1507228632661602323 2022-03-25 11:01:23 +0530 <6godwolo> Niggas brought in Power like she's not a fucking idiot
1507228626005241921 2022-03-25 11:01:22 +0530 <crookdscentence> @jorgemikell nigga its way too late for that lol
1507228623031480320 2022-03-25 11:01:21 +0530 <NOHEZZO> Ion big bro no nigga and ion let no nigga ill bro me
1507228621777383436 2022-03-25 11:01:21 +0530 <arrogantave> She swipe. She ain't impressed by yo bread nigga lol
1507228620753973272 2022-03-25 11:01:20 +0530 <nino_dior> Esa rola de nigga, "Luna" siempre será de mis favoritas.
```

Fig 3.1 Tweets scraping using twint tool

Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API.

1. Some of the benefits of using twintvs Twitter API:
2. Can fetch almost all Tweets (Twitter API limits to last 3200 Tweets only)
3. Fast initial setup
4. Can be used anonymously and without Twitter sign up
5. No rate limitations

Also in the proposed system we are adding functionality of extracting text from images

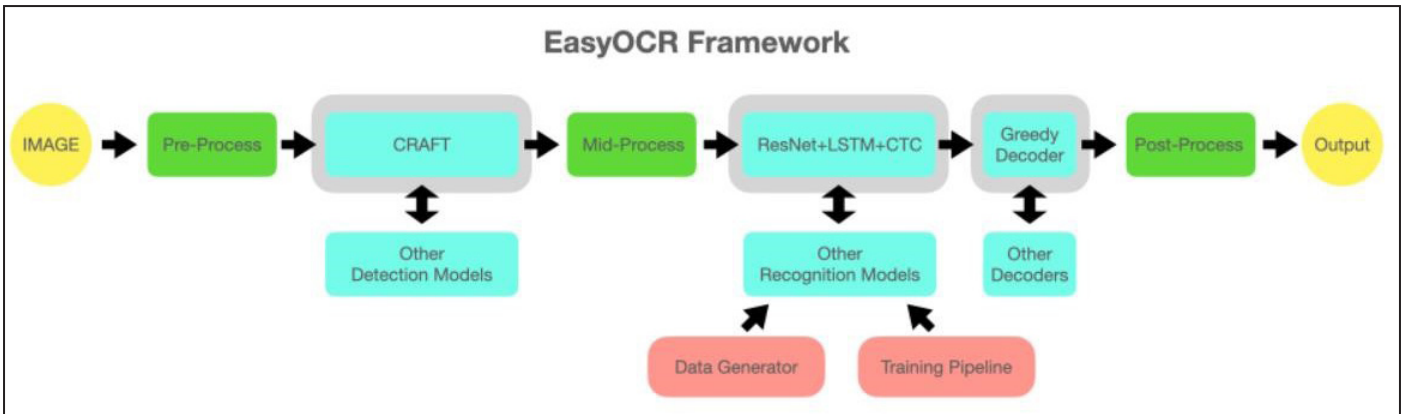


Fig. 3.2 Easy OCR workflow (Image Source: Github)

Easy OCR is actually a python package that holds PyTorch as a backend handler. EasyOCR like any other OCR(tesseract of Google or any other) detects the text from images but in my reference, while using it I found that it is the most straightforward way to detect text from images also when high end deep learning library(PyTorch) is supporting it in the backend which makes it accuracy more credible. Easy OCR supports 42+ languages for detection purposes. EasyOCR is created by the company named Jaided AI.

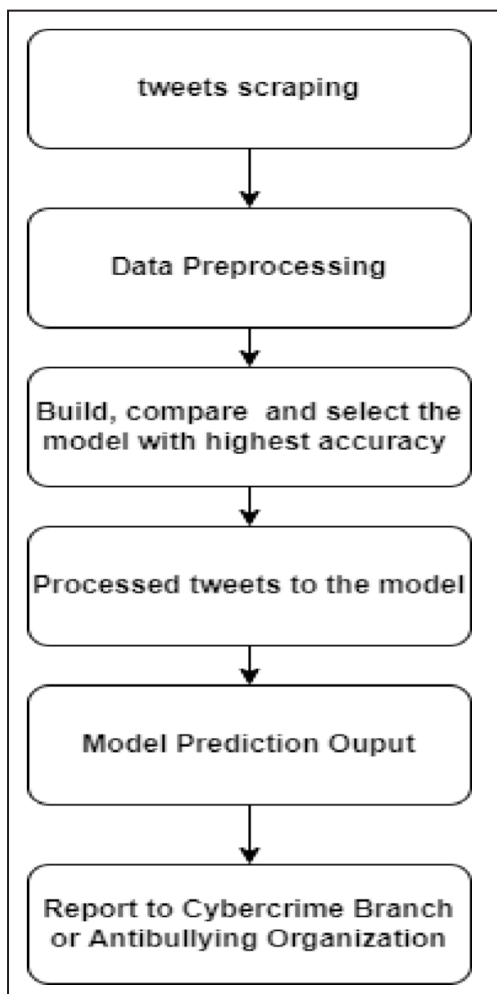


Fig 3.3 Proposed System Architecture

Dataset Description For the dataset we have scraped the tweets using Twint. Some of the keywords used to scrape the tweets

are ‘nigga’, slut’, ‘muslim’, etc. The dataset consists of three columns viz. tweet. label and category. There are four categories, racism, sexism, other and none. We have been labeling the dataset manually. Our dataset has around 17 thousands tweets out of which 10 thousands tweets are labeled manually and we are labeling the remaining tweets.

**A. Data Pre-processing:**

The data cleaning methods used in this research are: removing twitter handler, removing URLs. Since, we cannot analyze a tweet by reading Twitter handlers or URLs. Sometimes it can lead to overfitting as well. Other pre-processing methods used are: Removing punctuation. and characters, Tokenization, Removing stop words, Stemming.

**B. Feature Extraction**

To analyze the pre-processed Twitter data, it needs to be represented as features. In this research text features are constructed using Bag of Words and TF-IDF.



Fig 3.4 WordCloud for the non bullying tweets



Fig 3.5 WordCloud for the Racism related tweets





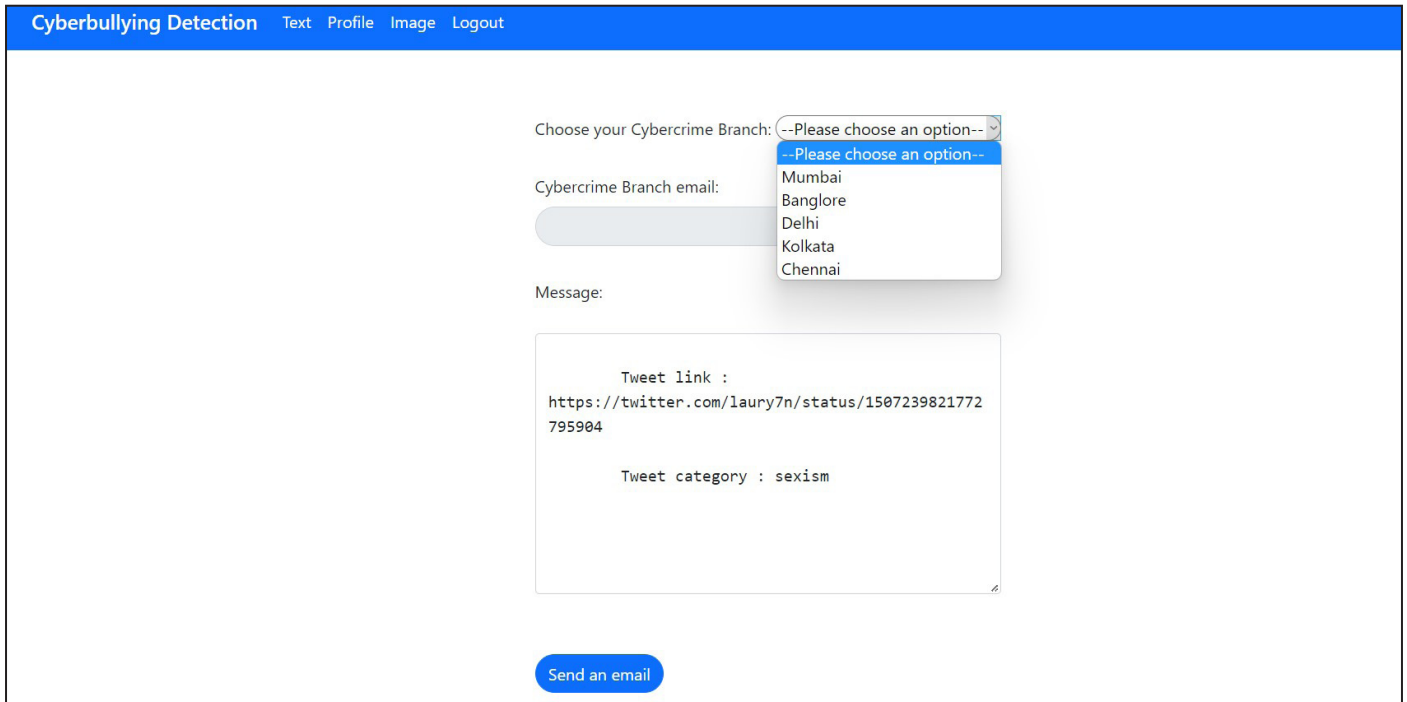


Fig 4.3 Reporting the tweet to the nearest Cyber Crime Branch



Fig 4.4 Report received at Cyber Crime Branch.

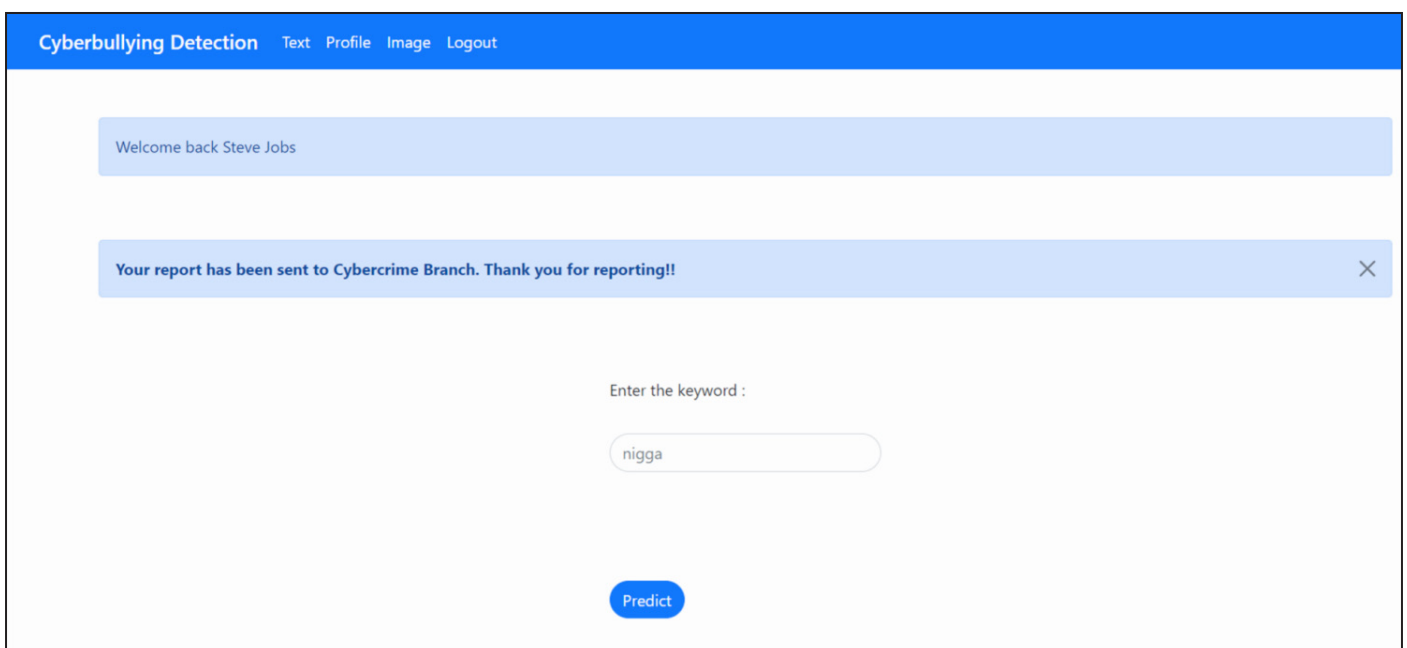


Fig. 4.5 Acknowledgement of Report to the user.

## V. Conclusion

As the dataset contains only 10 thousands tweets we have achieved only 80% accuracy. Also the algorithms are not tuned. If we increase the size of our dataset and tune the algorithm we can achieve more accuracy. The future scope of the research is to to classify and predict the videos in the tweet.

## References

- [1] Jyothi Shetty, K.N. Chaithali, Aditi M. Shetty, B. Varsha, and V.Puthran (2020). Cyber-Bullying Detection: A Comparative Analysis of Twitter Data
- [2] AnvithaKeni ,Deepa, Prof MangalaKini, Deepika K V, Divya C H (2020) Cyber-Bullying Detection using Machine Learning Algorithms
- [3] Rohini K R, Sreehari T Anil, Sreejith P M, Yedu Mohan P M (2020). Comparative Study of Cyberbullying Detection using different Machine Learning Algorithms
- [4] M. Dadvar and F.de Jong. 2012."Cyberbullying detection:a step toward a safer internet yard". In Proceedings of the 21st International Conference on World Wide Web (WWW'12 Companion). ACM, New York, NY, USA, 121-126