Social Media Toxicity Analyser

Narain C A[1] , Tirugnanam[1] ,Shrii Sudhan[1] , Mr. S. Rajesh Kumar[2]
[1]Department of Computer Science and Engineering, National Engineering College, Kovilpatti, Tamil nadu
[2] Assistant Professor Department of Computer Science and Engineering, National Engineering College, Kovilpatti, Tamil Nadu

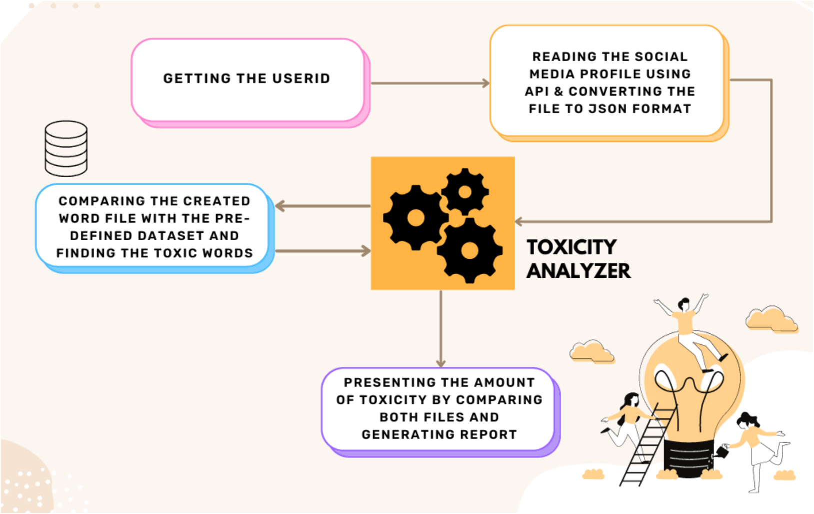
Abstract: **Cyberbullying and online abuse have been continuously increasing at an alarming rate. This is detrimental to the mental state of youngsters and one of the biggest factors leading to mental depression. Manually determining the toxicity of the comments in the avalanche of data generated daily is an impossible task. Automating the detection and censorship of such toxic comments by social media platforms can go a long way in solving this issue. But detection of toxic comments is a very daunting task because various factors such as context, perception, vocabulary, etc matters here. The motto of our project is to find the toxicity of their media profiles by the activities of the respective person involved in it. The toxicity can be identified by the captions, bio, and other wordings a person uses on social media like Twitter. The whole profile of the person is analyzed and compared with the predefined data and a report is generated containing the toxicity of the social media profile. This report can be used by the cyber-security department to find whether that person is toxic or bullying anyone on the internet. Through this, we can decently manage social media content.**

Keywords:- Cyber bullying, Online abusing, abusing. Bullying, cyber attack

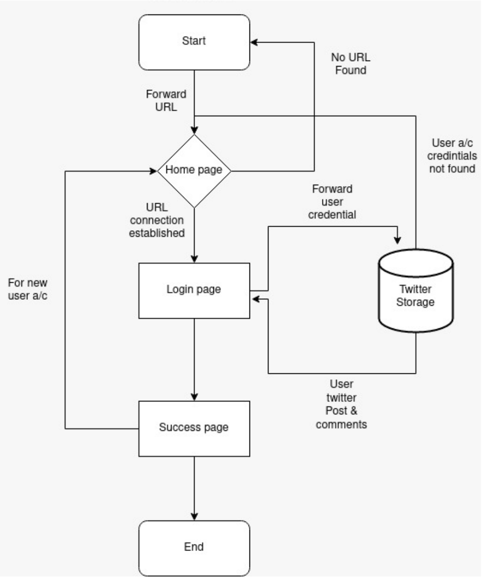
1. INTRODUCTION

Nowadays social media are becoming more toxic like fake accounts, usage of abusive words, and posting content that hurts other users. Therefore, to reduce toxicity, social media toxicity has been introduced in which the toxicity detector generates a report on the toxicity level of the given account or user id. The project's scope is to verify the social media account by scrapping all the posts, and comments posted by the user and detecting toxicity, and providing a report for the toxicity level in the account. It leads to harmless social media usage. The architecture was well-designed and organized. It makes social media more safe and more secure. Authentication works well.

2. SYSTEM DESIGN


GETTING THE USERID
READING THE SOCIAL MEDIA PROFILE USING API & CONVERTING THE FILE TO JSON FORMAT
COMPARING THE CREATED WORD FILE WITH THE PRE-DEFINED DATASET AND FINDING THE TOXIC WORDS
TOXICITY ANALYZER
PRESENTING THE AMOUNT OF TOXICITY BY COMPARING BOTH FILES AND GENERATING REPORT

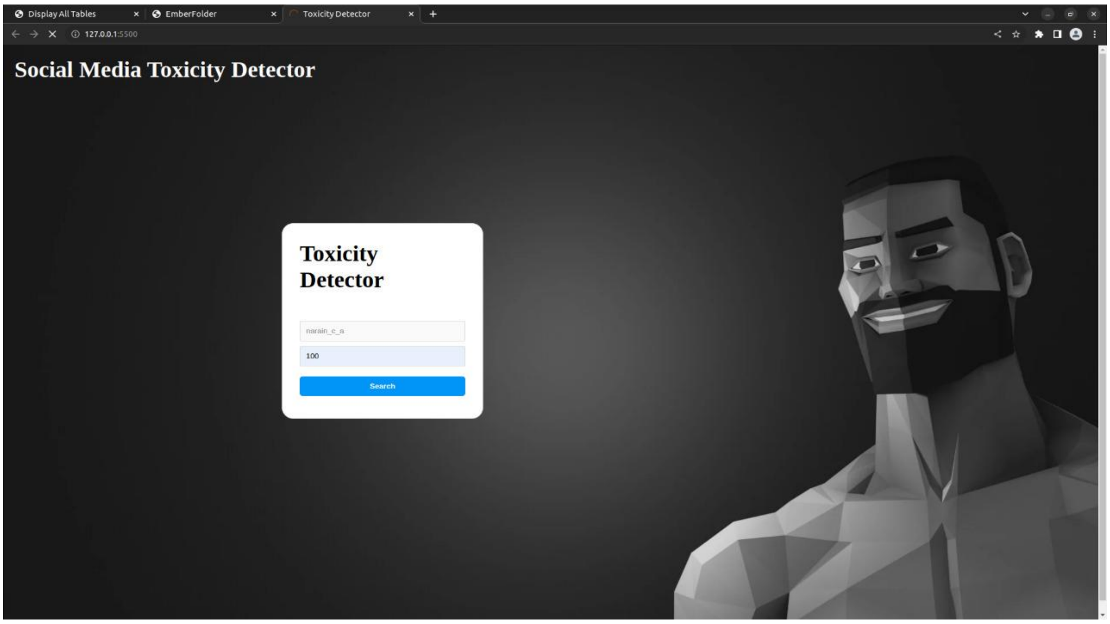
System Flowchart



2. IMPLEMENTATION

The main objective of the project is to find the toxicity level of a particular given user's social media accounts, especially for a Twitter account. For doing this, we have used python language as a platform and imported the snscrape library module, JSON library, and flask library module. The project consists of an HTML page that gets the user's Twitter account name and the total number of last tweets to be retrieved to get the contents of posts and comments by the user. These details have been sent to the python flask login page function using an HTTP POST/GET request. The details are given to the snscrape's twitterScraper.TwitterUserScraper module and tweet contents up to the specified number will be fetched and append the data into a tweets array in python. The tweets will be saved into a newly created result.json file in the current location of the code. The data will be parsed through the for loops and it checks for the swear words from the dataset.json. If it matches the swear words, then the matched word will be added to the swear category strings and counts will be taken accordingly. The count of swear words based on category as specified in the dataset file and categorized swear words will be sent to the success function of the python module and this function consists of an embedded HTML file and connect the sentence with that embedded code. Finally, display the contents in the new HTML success page for the user.

3. RESULT ANALYSIS



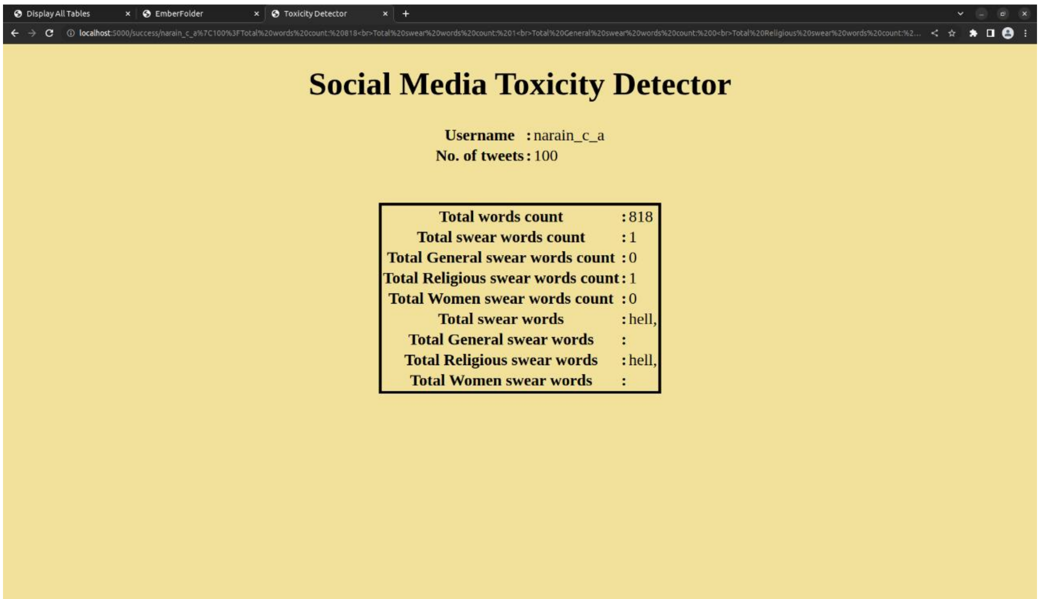
Social Media Toxicity Detector

Toxicity Detector

narain_c_a

100

Fig 3.1 Home page

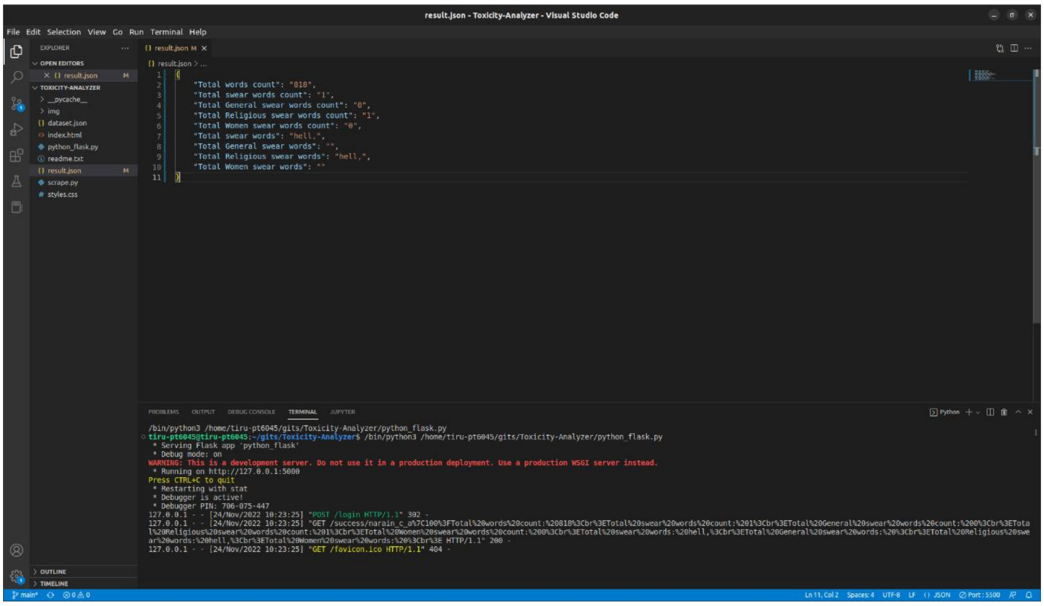


Social Media Toxicity Detector

Username : narain_c_a
No. of tweets : 100

Total words count	: 818
Total swear words count	: 1
Total General swear words count	: 0
Total Religious swear words count	: 1
Total Women swear words count	: 0
Total swear words	: hell,
Total General swear words	:
Total Religious swear words	: hell,
Total Women swear words	:

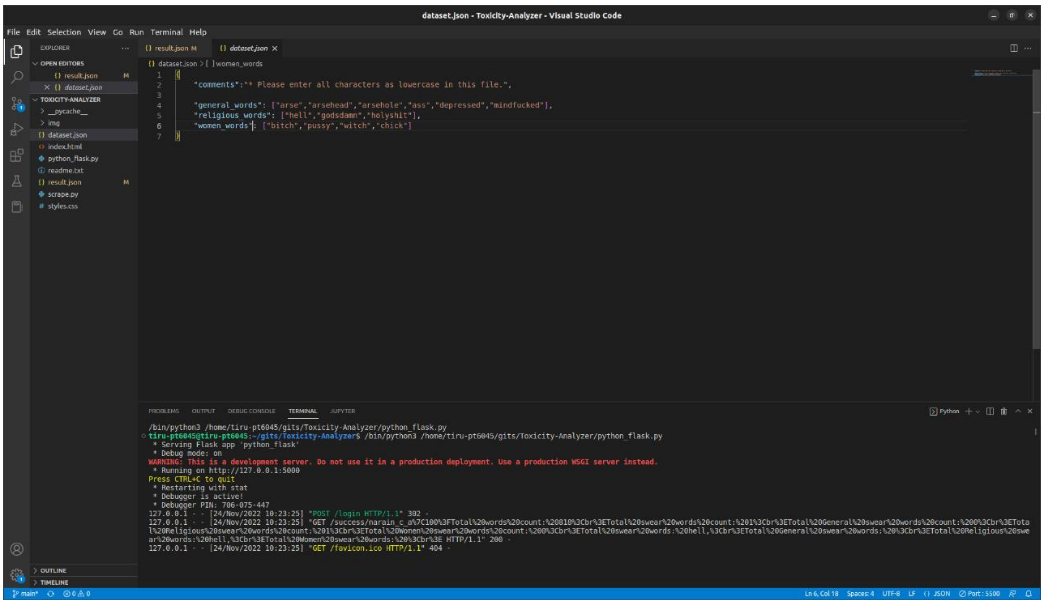
Fig 3.2 Result Page



```
{
    "Total words count": "818",
    "Total swear words count": "1",
    "Total General swear words count": "0",
    "Total Religious swear words count": "1",
    "Total Women swear words count": "0",
    "Total swear words": "hell,",
    "Total General swear words": "",
    "Total Religious swear words": "hell,",
    "Total Women swear words": ""
}
```

```
/bin/python3 /home/tiru-pt6045/gits/Toxicity-Analyzer/python_flask.py
tiru-pt6045@tiru-pt6045:~/gits/Toxicity-Analyzer$ /bin/python3 /home/tiru-pt6045/gits/Toxicity-Analyzer/python_flask.py
 * Serving Flask app 'python_flask'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 706-075-447
127.0.0.1 - - [24/Nov/2022 10:23:25] "POST /login HTTP/1.1" 302 -
127.0.0.1 - - [24/Nov/2022 10:23:25] "GET /success/narain_c_a%7C100%3FTotal%20words%20count:%20818%3Cbr%3ETotal%20swear%20words%20count:%201%3Cbr%3ETotal%20General%20swear%20words%20count:%200%3Cbr%3ETota
l%20Religious%20swear%20words%20count:%201%3Cbr%3ETotal%20Women%20swear%20words%20count:%200%3Cbr%3ETotal%20swear%20words:%20hell,%3Cbr%3ETotal%20General%20swear%20words:%20%3Cbr%3ETotal%20Religious%20swe
ar%20words:%20hell,%3Cbr%3ETotal%20Women%20swear%20words:%20%3Cbr%3E HTTP/1.1" 200 -
127.0.0.1 - - [24/Nov/2022 10:23:25] "GET /favicon.ico HTTP/1.1" 404 -
```

Fig 3.3 Saved Tweet contents in files for particular user



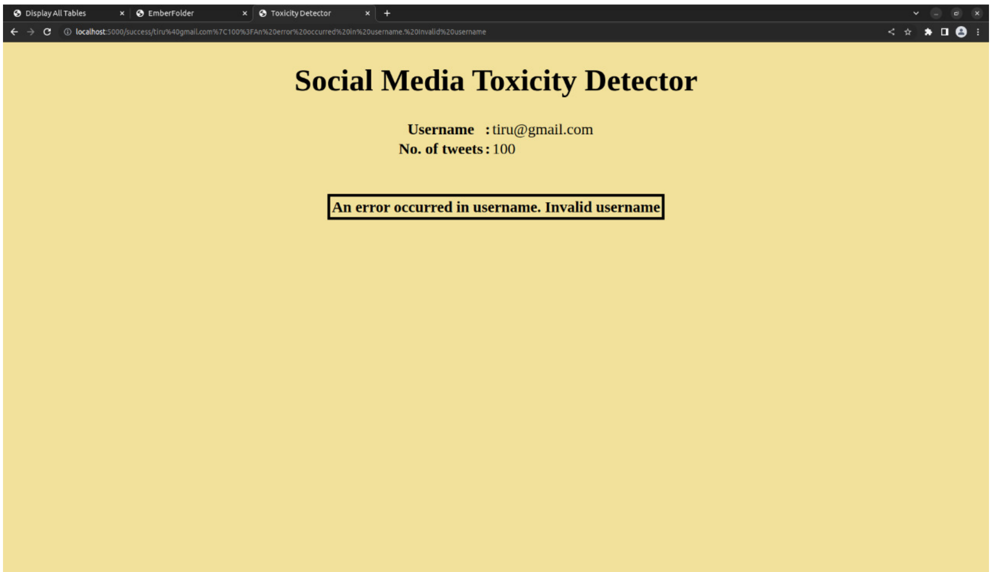
dataset.json - Toxicity-Analyzer - Visual Studio Code
File Edit Selection View Go Run Terminal Help
EXPLORER
OPEN EDITORS
{} result.json M
× {} dataset.json
TOXICITY-ANALYZER
> __pycache__
> img
{} dataset.json
<> index.html
python_flask.py
readme.txt
{} result.json M
scrape.py
# styles.css
{} dataset.json > [ ] women_words
PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL JUPYTER
Python
OUTLINE
TIMELINE
main* ⊗ 0 ⚠ 0
Ln 6, Col 18 Spaces: 4 UTF-8 LF {} JSON Port : 5500

```
{
    "comments":"* Please enter all characters as lowercase in this file.",

    "general_words": ["arse","arsehead","arsehole","ass","depressed","mindfucked"],
    "religious_words": ["hell","godsdamn","holyshit"],
    "women_words": ["bitch","pussy","witch","chick"]
}
```

```
/bin/python3 /home/tiru-pt6045/gits/Toxicity-Analyzer/python_flask.py
tiru-pt6045@tiru-pt6045:~/gits/Toxicity-Analyzer$ /bin/python3 /home/tiru-pt6045/gits/Toxicity-Analyzer/python_flask.py
 * Serving Flask app 'python_flask'
 * Debug mode: on
WARNING: This is a development server. Do not use it in a production deployment. Use a production WSGI server instead.
 * Running on http://127.0.0.1:5000
Press CTRL+C to quit
 * Restarting with stat
 * Debugger is active!
 * Debugger PIN: 706-075-447
127.0.0.1 - - [24/Nov/2022 10:23:25] "POST /login HTTP/1.1" 302 -
127.0.0.1 - - [24/Nov/2022 10:23:25] "GET /success/narain_c_a%7C100%3FTotal%20words%20count:%20818%3Cbr%3ETotal%20swear%20words%20count:%201%3Cbr%3ETotal%20General%20swear%20words%20count:%200%3Cbr%3ETota
l%20Religious%20swear%20words%20count:%201%3Cbr%3ETotal%20Women%20swear%20words%20count:%200%3Cbr%3ETotal%20swear%20words:%20hell,%3Cbr%3ETotal%20General%20swear%20words:%20%3Cbr%3ETotal%20Religious%20swe
ar%20words:%20hell,%3Cbr%3ETotal%20Women%20swear%20words:%20%3Cbr%3E HTTP/1.1" 200 -
127.0.0.1 - - [24/Nov/2022 10:23:25] "GET /favicon.ico HTTP/1.1" 404 -
```

Fig 3.4 Dataset file for data checking



Display All Tables
EmberFolder
Toxicity Detector
localhost:5000/success/tiru%40gmail.com%7C100%3FAn%20error%20occurred%20in%20username.%20Invalid%20username

Social Media Toxicity Detector

Username : tiru@gmail.com
No. of tweets : 100

An error occurred in username. Invalid username

Fig 3.4 Invalid Username Check

4. LITERATURE SURVEY

Kazi Saeed Alam; Shovan Bhowmik; Priyo Ranjan Kundu Prosun et al [1] (2019) - Proposed This research is to automate the detection process of offensive language or cyberbullying. The main aim is to build a single and double ensemble-based voting model to classify the contents into two groups: `offensive' or `non-offensive'. For this purpose, four machine learning classifiers have been chosen and three ensemble models with two different feature extraction techniques combined with various n-gram analysis on a dataset extracted from Twitter.

Divya Iyyani; Keshin Jani; Swati Mali et al [2] (2018) - outlined troll has been an issue in social media for a long time. This explores the currently existing “Anti-trolling systems”, their methodologies and technological challenges. In particular, we consider machine learning and existing expert systems that detect and prevent trolls. The paper concludes with a discussion of the issues faced by this technology, the current functioning and a few suggestions along with a modified architecture and the vision of a trolling-free internet .

Cheniki Abderrouaf; Mourad Oussalah et al [3] (2019) - presented The main content in this topic is by leveraging novel natural language processing, machine learning, and feature engineering techniques. The proposed approach advocates a classification-like technique that makes use of a special data design procedure. The latter enforces a balanced training scheme by exploring the negativity of the original dataset.

Kanwal Yousaf; Tabassam Nawaz et al [4] (2022) - expressed the exponential growth of videos on YouTube has attracted billions of viewers among which the majority belongs to a young demographic. Malicious uploaders also find this platform as an opportunity to spread upsetting visual content, such as using animated cartoon videos to share inappropriate content with children. Therefore, an automatic real-time video content filtering mechanism is highly suggested to be integrated into social media platforms.

Nadira Boudjani, Yannis Haralambous, Inna Lyubareva et al [5] (2020)- proposed supervised approach has been proposed for toxic comment classification for the French language. We choose a set of features proposed for toxic comment detection for English and use it for French toxic comment detection. Our approach is based on N-gram features, linguistic features and a dictionary of insulting words and expressions.

Chetna Sharma ,Rahul Ramakrishnan,Ayusha Pendse,Priya Chimurkar Kiran T. Talele: et al [6] (2020) presented cyberbullying is a major issues in social media. The work here is to identifying cyberbullying at is origin, meaning when it is being drafted in real-time. Identifying traces of cyberbullying before the content is uploaded on the internet can help reduce circulation of hurtful messages.

Muhammad Amien Ibrahim; Noviyanti Tri Maretta Sagala; Samsul Arifin; Rinda Nariswari; Nerru Pranuta Murnaka et al [7] (2022) presented the main aim is to classify hate speech, abusive language, and normal messages on Indonesian Twitter. Several machine learning models, such as logistic regression and BERT models, are utilized to accomplish text classification tasks.

Mahamat Saleh Adoum Sanoussi; Chen Xiaohua; George K. Agordzo; Mahamed Lamine Guindo; Abdullah MMA Al Omari et al [8] (2022) Identifying hate speech on social media has become increasingly crucial for society. It has been shown that cyberbul-lying significantly affects the social tranquillity of the Chadian population, mainly in places of conflict. This article aims to detect hate speech for texts written in “lingua franca”, a mix of the local Chadian and French languages.

Jayant Singh; Kishorjit Nongmeikapam et al [9] (2011) presented aggression and related activities such as trolling peoples, harassing online involves hate comments in various forms. There are numerous such cases coming in present time and sites respond by closing down their remark areas. After the introduction of Machine Learning and having data in massive amounts now its quite logical to build a tool which can tackle this situation.

Paraskevas Tsantarliotis; Evaggelia Pitoura; Panayiotis Tsaparas et al [10] (2014) took a novel approach to the trolling problem: our goal is to identify the targets of the trolls, so as to prevent trolling before it happens. We thus define the troll vulnerability prediction problem, where given a post we aim at predicting whether it is vulnerable to trolling. Towards this end, we define a novel troll vulnerability metric of how likely a post is to be attacked by trolls, and we construct models for predicting troll-vulnerable posts, using features from the content and the history of the post.

Lu Cheng , Ahmadreza Mosallanezhad , Yasin N. Silva , Deborah L. Hall ,Huan Liu et al [11] (2021) Increased social media use has contributed to the greater prevalence of abusive, rude, and offensive textual comments. Machine learning models have been developed to detect toxic comments online, yet these models tend to show biases against users with marginalized or minority identities.

Paula Reyero Lobo, Enrico Daga, Harith Alani et al [12] (2022) presented the use of a knowledge graph to help in better understanding such toxic speech annotation issues. Our empirical results show that 3% in a sample of 19k texts mention terms associated with frequently attacked gender and sexual orientation groups that were not correctly identified by the annotators.

B.S. Sahana; G. Sandhya; R.S. Tanuja; Sushma Ellur; A. Ajina et al [13] (2022) expressed unlike most of the work in toxic content detection where the nature of toxicity is determined, we treat the detection of toxic content as a binary classification task. Here, we have explored Support Vector Machine, Boosting and deep neural networks for classification. We have trained the model on twitter datasets. With a goal of better predictive performance, our approach uses a majority voting ensemble to aggregate the predictions of individual classifiers.

Ishan Sanjeev Upadhyay, KV Aditya Srivatsa, Radhika Mamidi et al [14] (2022) presented a dataset for toxic positivity classification from Twitter and an inspirational quote website. We then perform benchmarking experiments using various text classification models and show the suitability of these models for the task.

Yau-Shian Wang, Yingshan Chang et al [15] (2022) outlined the generative variant of zero-shot prompt-based toxicity detection with comprehensive trials on prompt engineering. We evaluate on three datasets with toxicity labels annotated on social media posts. Our analysis highlights the strengths of our generative classification approach both quantitatively and qualitatively. Interesting aspects of self-diagnosis and its ethical implications are discussed.

5. CONCLUSION

The act of bullying is a punishable offense, people on the internet are covering their faces and performing these kinds of actions. This should be condemned and thrown away from the internet and should be punished. Through this web application anyone can scrape the twitter data and find the toxicity of a particular user id and can avoid them. This helps in finding out people's behavior on the internet. The project will be improved with more features like reporting and intimating the cyber security.

6. REFERENCES

[1] Kazi Saeed Alam; Shovan Bhowmik; Priyo Ranjan Kundu Prosun (2019): - "Cyberbullying Detection: An Ensemble Based Machine Learning Approach"

[2] Divya Iyyani; Keshin Jani; Swati Mali:(2018) - "Troll-Detection Systems Limitations of Troll Detection Systems and AI/ML Anti-Trolling Solution"

[3] Cheniki Abderrouaf; Mourad Oussalah (2019) - "On Online Hate Speech Detection. Effects of Negated Data Construction"

[4] Kanwal Yousaf; Tabassam Nawaz: (2022) - "A Deep Learning-Based Approach for Inappropriate Content Detection and Classification of YouTube Videos"

[5] Nadira Boudjani, Yannis Haralambous, Inna Lyubareva: (2020) - "Toxic Comment Classification For French Online Comments"

[6] Chetna Sharma ,Rahul Ramakrishnan,Ayusha Pendse,Priya Chimurkar Kiran T. Talele: (2020) - "Cyber-Bullying Detection Via Text Mining and Machine Learning"

[7] Muhammad Amien Ibrahim; Noviyanti Tri Maretta Sagala; Samsul Arifin; Rinda Nariswari; Nerru Pranuta Murnaka: (2022) - "Separating Hate Speech from Abusive Language on Indonesian Twitter"

[8] Mahamat Saleh Adoum Sanoussi; Chen Xiaohua; George K. Agordzo; Mahamed Lamine Guindo; Abdullah MMA Al Omari: (2022) - "Detection of Hate Speech Texts Using Machine Learning Algorithm"

[9] Jayant Singh; Kishorjit Nongmeikapam: (2011) - "Negative Comments Multi-Label Classification"

[10] Paraskevas Tsantarliotis; Evaggelia Pitoura; Panayiotis Tsaparas: (2014) - "Troll vulnerability in online social networks"

[11] Lu Cheng , Ahmadreza Mosallanezhad , Yasin N. Silva , Deborah L. Hall ,Huan Liu (2021) - "Bias Mitigation for Toxicity Detection via Sequential Decisions"

[12] Paula Reyero Lobo, Enrico Daga, Harith Alani: (2022) - "Supporting Online Toxicity Detection with Knowledge Graphs"

[13] B.S. Sahana; G. Sandhya; R.S. Tanuja; Sushma Ellur; A. Ajina: (2022) - "Towards a Safer Conversation Space: Detection of Toxic Content in Social Media (Student Consortium)"

[14] Ishan Sanjeev Upadhyay, KV Aditya Srivatsa, Radhika Mamidi: (2022) - "Towards Toxic Positivity Detection"

[15] Yau-Shian Wang, Yingshan Chang: (2022) - "Toxicity Detection with Generative Prompt-based Inference"