TOXIC COMMENTS CLASSIFICATION IN SOCIAL NETWORKING.

DEVELOPED BY: MOHAMMED FAHAD

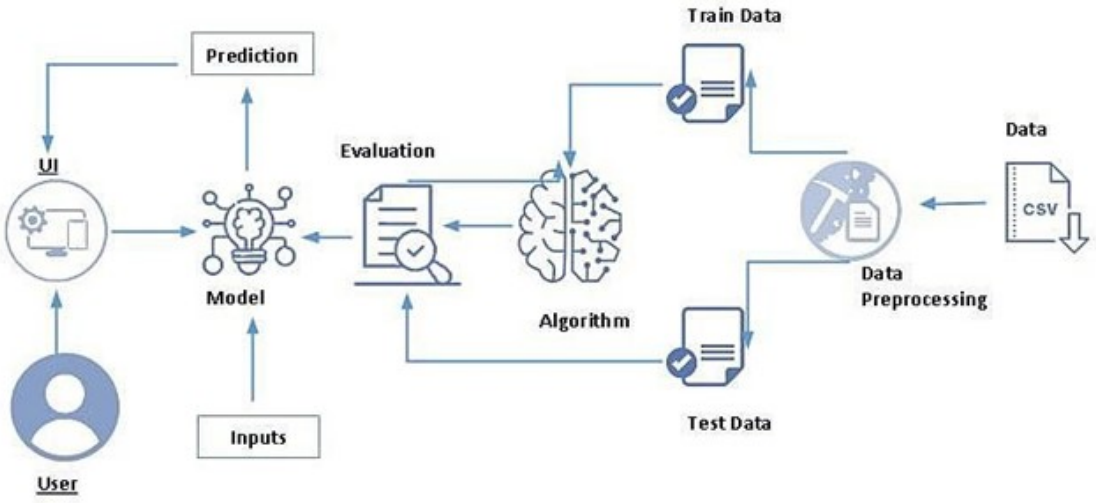
SMART BRIDGE – Major Project Report

1. INTRODUCTION

The flow of data over the internet has grown dramatically, especially with the appearance of social networking sites. Social networks sometimesbecome a place for threats,insults, and other components of cyberbullying. A huge number of people are involved in online social networks.

Toxic comments are textual comments with threats, insults, obscene, racism, etc. In recent yearsthere have been many cases in which authorities have arrested some users of social sites because of the negative (abusive) content of their personal pages. Hence, the protection of network users from anti-social behavior is an important activity.One of the major tasks of such activity is the automateddetection of toxiccomments.

Bag of words statistics and bag of symbols statistics are the typical source information for toxiccomments detection. Usually,the following statistics-based features are used: length of the comment,number of capital letters, number of exclamation marks, number of question marks, number of spelling errors, number of tokens with non-alphabet symbols, number of abusive, aggressive, and threatening wordsin the comment, etc. A neuralnetwork model is used to classify the comments.



1.1. OVERVIEW

By the end of this project:

- ➤ We'll be able to understand the problem to classify if it is a regression or a classification kindof problem.
- ➤ We will be able to know how to pre-process/clean the data using different data pre-processingtechniques.
- ➤ Applying different algorithms according to the dataset
- ➤ We will be able to know how to find the accuracy of the model.
- ➤ We will be able to build web applications using the Flask framework

1.2. PURPOSE

Social media has made it very easy for us to communicate quickly and easily with family, friends, and acquaintances, as well as share experiences and let others know of our opinions and beliefs. These opinions and beliefs may be about world events or local affairs, politics or religion, interests, affiliations, organizations, products, people, and a wide variety of other topics. Our conversations and commentscan be closely targeted or widely broadcastto the point that depending on the subject,they can go viral.

Unfortunately, social media is also widely used by abusers, for exactly the reasons listed above.Many perpetrators 'hide' behind the fact that they may not be able to be readily identified, saying things that they wouldn'tconsider saying face-to-face, which couldbe regarded as cowardly.

Social media has significantly simplified and expedited our ability to connect with loved ones, friends, and acquaintances, while also providing a platform to share our experiences and express our views on a multitude of topics such as global and local events, politics, religion, interests, affiliations, organizations, products, individuals, and more. Through social media, we can engage in targeted discussions or broadcast our messages to a wide audience, often resulting in content going viral. Online abuse takes several forms, and victims are not confined to public figures. They can do any job, be ofany age, gender, sexual orientation, or social or ethnic background, and live anywhere.
So the purpose of this project is to classify the comments based on their nature and categorize themas clean or toxic.

2. LITERATURE SURVEY

2.1. EXISTING PROBLEM

The negative impacts of cyberbullying are numerous. Consequences of cyberbullying can include poor academic performance, school dropout, physical violence, and suicide, and it is a method of bullying that is frequently hidden from adults (Willard, 2006a). According to Patchin and Hinduja, (2008), cyberbullying is linked to serious effectssuch as low self-esteem, familyproblems, academic problems, school violence, and delinquent behavior. However, the worst consequences are suicide and violence. While cyberbullying has some of the same negative impactsas traditional face-to-face bullying, it can be done without any physical contact or knowledge of the perpetrator's identity (Willard, 2006). These random acts of harassment go well beyond the scope of traditional face-to- face bullying because unlike traditional bullying, cyberbullying can occur not only at school, but in the home and any place where technology is accessible (Shariff & Hoff, 2007; Stover, 2006; Strom &Strom, 2005). Studies have suggested that although it may occur less frequently than face-to-face bullying, up to 70% of students in the United States have experienced cyberbullying (Juvonen & Gross,2008; Wang, Iannotti, & Nansel, 2009).Therefore, there is a need for furtherstudies to obtaina conceptualized view of the number of students across the United States and beyond who have experienced some form of cyberbullying.

Existing Solutions

Being online has so many benefits. However, like many things in life, it comes with risks that you need to protect against.

If you experience cyberbullying, you may want to delete certain apps or stay offline for a while to give yourself time to recover.But getting off the Internetis not a long-term solution. You did nothingwrong, so why should you be disadvantaged? It may even send the bullies the wrong signal encouraging their unacceptable behavior.

We all want cyberbullying to stop, which is one of the reasons reporting cyberbullying is

so important. But creating the Internet we want goes beyond calling out bullying. We need to be thoughtful about what we share or say that may hurt others.We need to be kind to one another onlineand in real life. It's up toall of us!

2.2. PROPOSED SOLUTION

The solution with which we as a team came up is a classification tool that classifies each comment that is being made on social media is clean and free of toxicity so that every human on this planet can make use of a broad spectrum of social networking and share their experiences withoutharmingor bullying others.

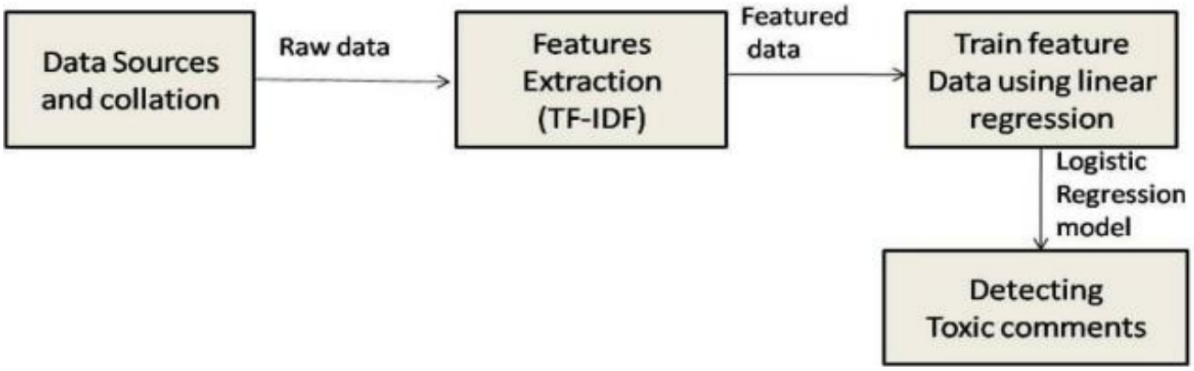
Now that we have many advanced platforms on social media but we still lag behind to make a safe and clean virtual social environment that is free from all kinds of hatred.

Toxic People still abuse others by taking advantage of social media, but this can be sorted out by installing this project, which can classify comments based on the words and special characters usedwhile posting a comment.

It is already trained with multiple datasets and is capable of classifying most of the comments that are made in social networking.

3. THEORETICAL ANALYSIS

2.1. BLOCK DIAGRAM



The above picture (Fig.2) depicts the block diagram of the process which gives clear insights into the entire process where Data sourcesand collation is done and the raw data is given as input for the processof features extraction which in turn givesthe featured data as output as shown in the figure.

In the later step, feature data is trained by using logistic regression, and a logistic regression model is built which plays the role of classifying and detecting Toxic comments in social networking.

Project Flow:

- Download the dataset.
- Preprocess the textual data.
- Classify the dataset into train and test sets.
- Add the neural networklayers.
- Load the trained data andfit the model.
- Test the model.
- Save the model and its dependencies.
- Build a Web application using flask that integrates with the model built.

2.2. HARDWARE / SOFTWARE DESIGNING

To complete this project you should have the following software and packages

Anaconda Navigator:

Anaconda Navigator is a free and open-source distribution of the Python and R programminglanguages for data science and machine learning-related applications. It can be installedon Windows, Linux, and macOS. Conda is an open-source, cross-platform, package management system. Anacondacomes with so very nice tools like JupyterLab, Jupyter Notebook, QtConsole, Spyder, Glueviz,Orange, Rstudio, Visual Studio Code. For this project, we will be usingJupiternotebook and spyder.

Sklearn: Scikit-learn, commonly referred to as sklearn, is a popular machine learning library for the Python programming language. It provides a wide range of tools and algorithms for tasks such as classification, regression, clustering, dimensionality reduction, model selection, and preprocessing of data.

NumPy: NumPy, which stands for Numerical Python, is a fundamental library in Python for scientific computing and data manipulation. It provides support for large, multi-dimensional arrays and matrices, along with a collection of mathematical functions to operate on these arrays efficient.

Pandas: Pandas is a powerful and widely-used open-source library in Python for data manipulation and analysis. It provides high-performance data structures and data analysis tools, making it a go-to library for working with structured data. The core data structure in Pandas is the DataFrame, which is a two-dimensional table-like data structure with labeled columns and rows. It allows for easy handling, cleaning, transformation, and analysis of tabular data. The DataFrame can handle heterogeneous data types and missing values efficiently.

4. EXPERIMENTAL INVESTIGATION

One should have knowledge of the following Concepts:

Supervised learning and unsupervised learning are two fundamental categories of machine learning algorithms that are used to address different types of learning tasks.

Supervised Learning:

In supervised learning, the algorithm learns from labeled training data, where the input data (features) is accompanied by corresponding target labels. The goal is to learn a mapping or a function that can predict the correct output label for new, unseen input data.

Unsupervised Learning:

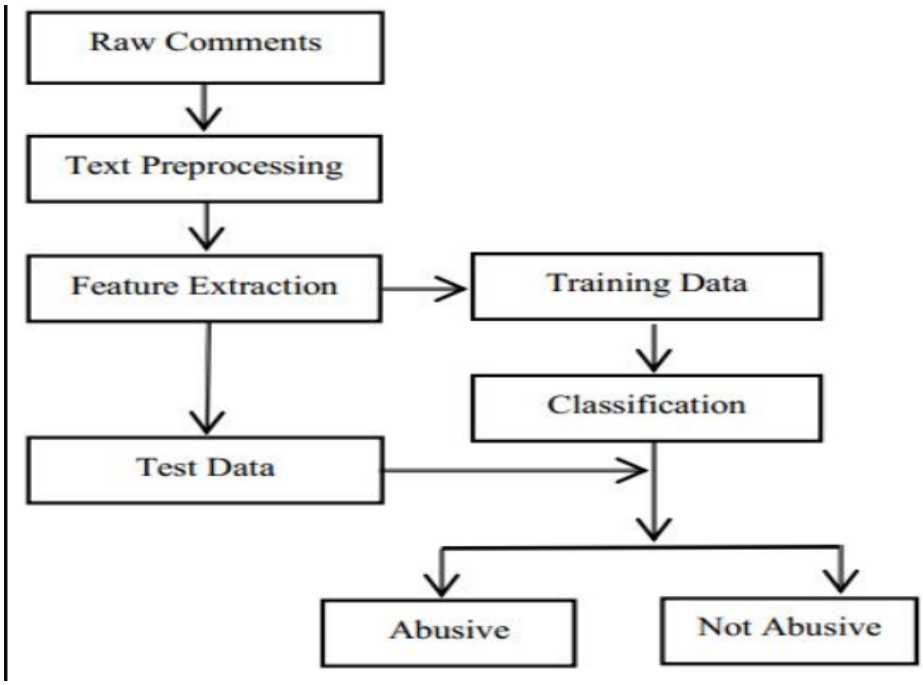
In unsupervised learning, the algorithm learns patterns, relationships, or structures within the data without any explicit target labels. The goal is to discover hidden patterns or groupings in the data or to reduce its dimensionality.

Artificial Neural Networks (ANNs) are computational models inspired by the structure and functionality of biological neural networks, particularly the brain. ANNs are a fundamental component of modern machine learning and deep learning, playing a crucial role in various applications, including image and speech recognition, natural language processing, and pattern recognition.

Convolutional Neural Networks (CNNs) are a specialized type of artificial neural network designed to process and analyze structured grid-like data, particularly in the domain of computer vision. CNNs have significantly advanced the state-of-the-art in image recognition, object detection, and other visual tasks.

Flask is a popular and lightweight web framework for building web applications using Python. It provides a simple and flexible way to create web servers and handle HTTP requests and responses. Flask's simplicity and minimalistic approach make it a popular choice for small to medium-sized web applications and APIs. It allows developers to quickly prototype and build web applications using Python without imposing a steep learning curve. Flask's flexibility also allows for easy integration with other Python libraries and frameworks, making it a versatile tool for web development.

5. FLOWCHART



We can get the following insights from the above diagram(figure 4) that,

- ☞ COLLECTION OF RAW DATA(COMMENTS).
- ☞ TEXT PREPROCESSING.
- ☞ FEATURE EXTRACTION.
- ☞ TRAINING AND TESTING THE EXTRACTED FEATURES.
- ☞ CLASSIFICATION BASED ON TRAINING.
- ☞ PREDICTION.

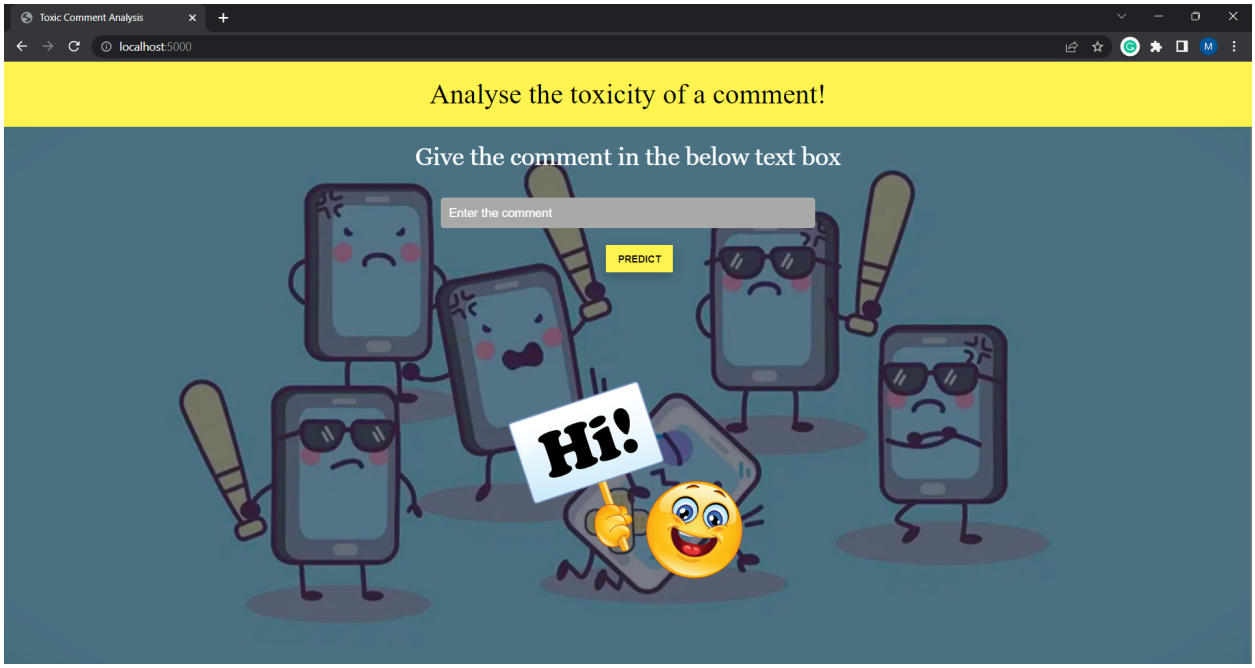
Data is given as input then the processing of the data is done by training the machine with trainingdata, here machine is trained using the appropriate algorithm and then evaluation is done to verify the outputs, and the training is continued until the machine is capable of classifying the comments fairly.

The machine is tested by providing testing data and again evaluation is done, at last, the model is built upon successful evaluation for the usage and this model is used for making the classification.

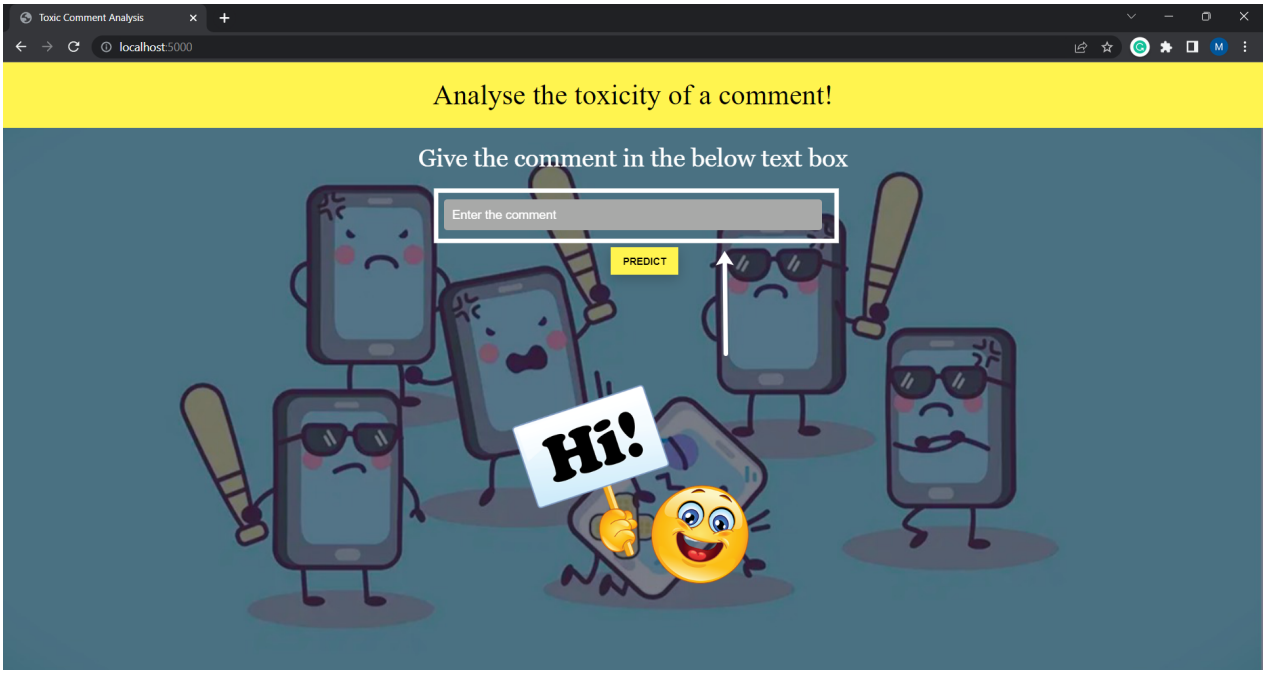
6. RESULTS

Following are the outputs which are presented in the screenshot format.

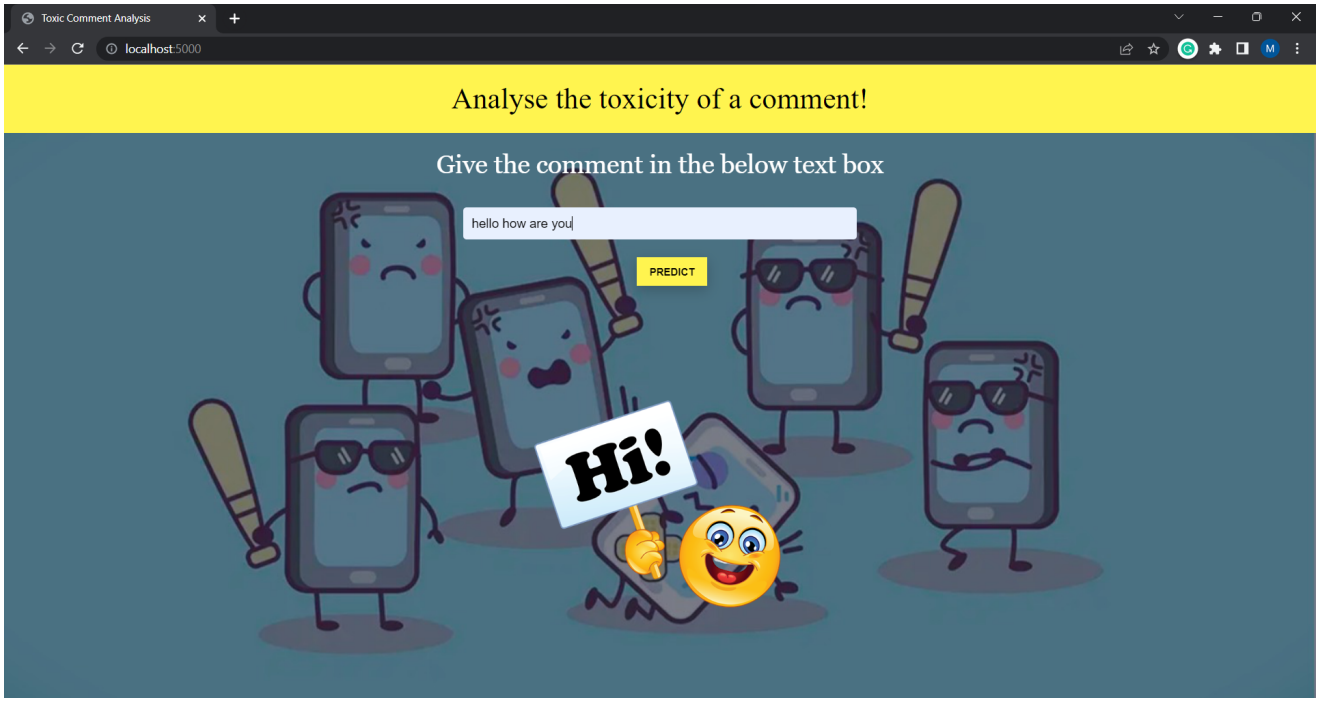
1. HOME PAGE.



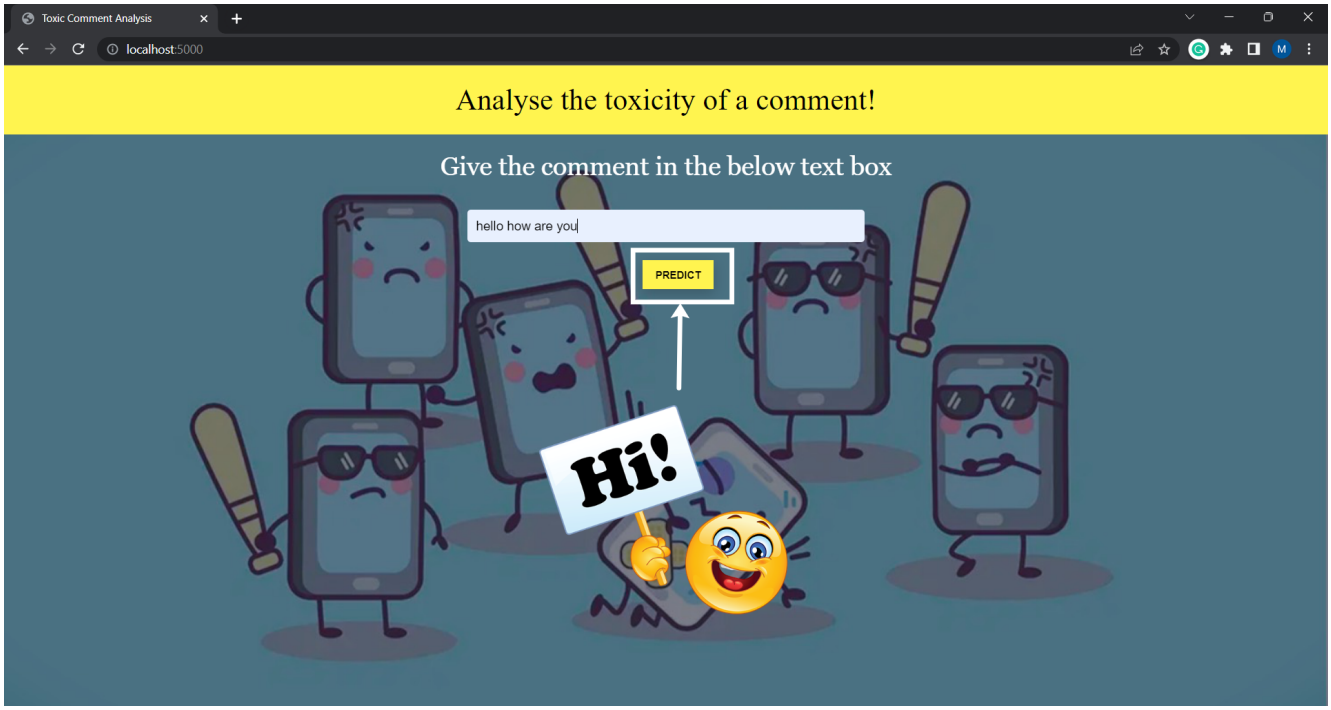
2. IT HAS A TEXT AREA THAT TAKES COMMENTS AS INPUT AS SHOWN BELOW.



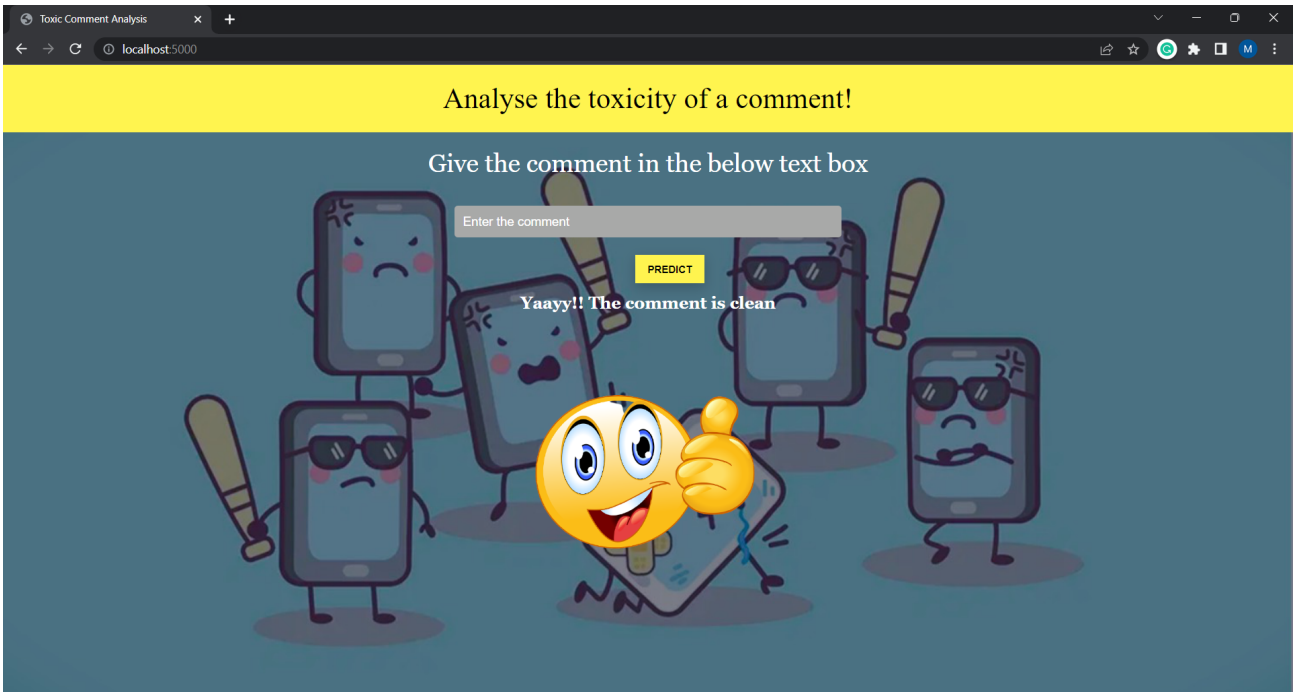
3. ENTER THE DESIRED TEXT.



4. CLICK ON PREDICT BUTTON.

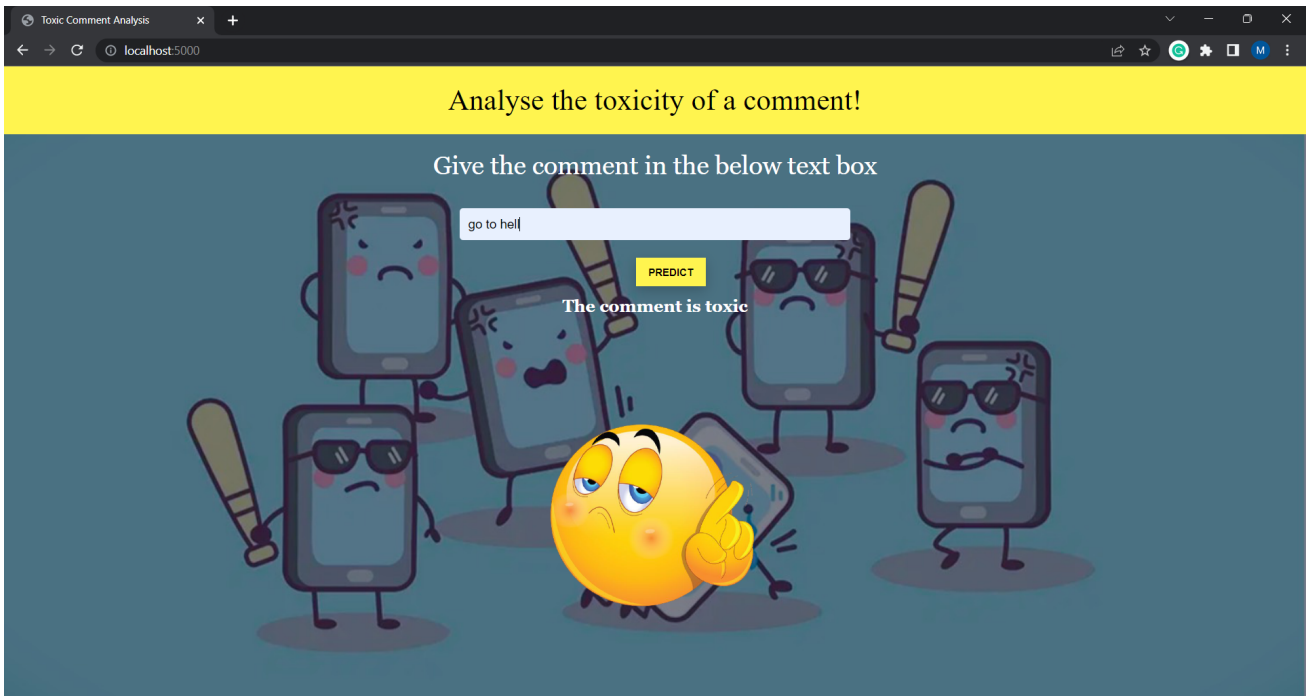


5. AS PER THE COMMENT GIVEN IT PREDICTS AS FOLLOWS.



We have seen in above figure 15 that the comments we entered were not toxic so it was predicted as "**THE COMMENT IS CLEAN**", whereas in the below figure we entered a negative and toxic comment so it predicted, "**THE COMMENT IS TOXIC**".

6. LET'S TEST WITH A NEGATIVE COMMENT.



7. ADVANTAGES

- ✓ It will not allow toxic comments on social networking.
- ✓ Time taken in classification is very quick.
- ✓ People can positively connect to the world.
- ✓ Toxic comments spoil the social virtual harmony and can be solved with this model

7. DISADVANTAGES

- ✓ Application of this model on all social media platforms is difficult.
- ✓ Error in classification may result in completely inconsistent results.
- ✓ Continuous maintenance is required.
- ✓ May be expensive for maintenance.

8. APPLICATIONS

Online forums and social media platforms have provided individuals with the means to put forward their thoughts and freely express their opinion on various issues and incidents. In some cases, these online comments contain explicit language which may hurt the readers. Comments containing explicit language can be classified into myriad categories such as Toxic, Severe Toxic, Obscene, Threat, Insult, and Identity Hate. The threat of abuse and harassment means that many people stop expressing themselves and give up on seeking different opinions.

Implementing this concept in all social media platforms will decrease toxic comments, online abuse, and cyberbullying.

By implementing this concept across all social media platforms, the prevalence of toxic comments, online abuse, and cyberbullying can be significantly reduced. This would create an environment where individuals feel more comfortable expressing their thoughts and seeking diverse opinions without the fear of encountering explicit language that can be harmful and discouraging.

Enforcing this principle across all social media platforms will reduce the prevalence of offensive comments, online abuse, and cyberbullying, thereby encouraging individuals to freely express their opinions without fear of harassment or intimidation.

9. CONCLUSIONS

By employing a neural network model that incorporates statistical features such as comment length, capital letters, exclamation marks, question marks, spelling errors, non-alphabet symbols, and the presence of abusive, aggressive, and threatening words, social media platforms can promote a cleaner and more positive online environment. This approach enhances the opportunity for individuals to freely share their positive ideas and knowledge with one another, fostering a more constructive and inclusive social community.Overall, the goal of using this model is to enhance the social media experience by creating a space where individuals feel comfortable expressing themselves and engaging in positive interactions while discouraging the spread of harmful or offensive content.

By employing a neural network model that incorporates statistical features such as comment length, capital letters, exclamation marks, question marks, spelling errors, non-alphabet symbols, and the presence of abusive, aggressive, and threatening words, social media platforms can promote a cleaner and more positive online environment. This approach enhances the opportunity for individuals to freely share their positive ideas and knowledge with one another, fostering a more constructive and inclusive social community.

10. FUTURE SCOPE

The exponential growth of social media users worldwide has facilitated the effortless sharing of knowledge, experiences, ideas, and innovations with a single click. However, alongside this positive development, there is a significant presence of individuals who deliberately aim to create a toxic and negative environment, causing harm to others. While some may perceive their actions as humorous, the impact can result in considerable mental stress for consumers of social media content.To address this issue, the implementation of the aforementioned model becomes increasingly crucial in the near future. It is essential for social media platforms to adopt this concept and take decisive measures to prevent users from posting abusive and toxic comments. By doing so, they can actively contribute to safeguarding the social well-being of their users.

As social media continues to evolve, it is imperative that platforms prioritize the implementation of such models to curb the negative impact of toxic behavior. By doing so, they can uphold their responsibility in cultivating a positive social environment where users can connect, learn, and engage in a healthy and uplifting manner.

11. BIBLIOGRAPHY

[1] The datasets are collected from kaggle.com, data.gov, the UCI machine learning repository, etc.

[2] Image Data Generator Reference

https://keras.io/api/preprocessing/image/

[3] You can download the dataset used in this project using the GitHub link:

https://github.com/Guided-Projects/Commet-Toxicity-Multi-Class-Classification

[4] Flask app reference

https://www.youtube.com/watch?v=lj4I_CvBnt0

[5] HTML reference

https://www.w3schools.com/html/

[6]. Deng, A., Yu. D., (2014). Deep Learning. Methods and Applications. Retrieved from http://research.microsoft.com/pubs/209355/DeepLearning-NowPublishing-Vol7-SIG39.pdf

[7]. Yoshua, B., (2009). Learning Deep Architectures for AI(PDF). Foundations and Trends in Machine Learning.

[8]. Nobata, C., Tetreault, J., Thomas, A. Mehdad, Y.,Chang,Y., (2016). Abusive Language Detection in Online User Content. International Conference on World Wide web, pp. 145–153.

[10]. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.,(2017). Enriching Word Vectors with Subword Information, TACL, vol 5, pp.135–146.

12. APPENDIX

A. SOURCE CODE

(commentApp.py)

```
import numpy as np
import pandas as pd
import re
from sklearn.feature_extraction.text import CountVectorizer
import pickle
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import accuracy_score
from flask import Flask, request, jsonify, render_template, url_for
loaded=CountVectorizer(decode_error='replace',vocabulary=pickle.load(open('word_feats.pkl','r
b')))
app = Flask(__name__)
def clean_text(text):
    text = text.lower()
    text = re.sub(r"what's", "what is ", text) text = re.sub(r"\'s", " ", text)
    text = re.sub(r"\'ve", " have ", text)  text = re.sub(r"can't", "cannot ", text)
    text = re.sub(r"n't", " not ", text) text = re.sub(r"i'm", "i am ", text)
    text = re.sub(r"\'re", " are ", text) text = re.sub(r"\'d", " would ", text)
    text = re.sub(r"\'ll", " will ", text)
    text = re.sub(r"\'scuse", " excuse ", text) text = re.sub('\W', ' ', text)
    text = re.sub('\s+', ' ', text)  text = text.strip(' ')
    return text
@app.route('/')
def landingpage():
    img_url = url_for('static',filename = 'images/hello.png')
    print(img_url) flag=0
    return render_template('toxic.html',flag=flag)
@app.route('/predict')
@app.route('/', methods = ['GET','POST'])
def predict():
    if request.method == 'GET':
        img_url = url_for('static',filename = 'images/hello.png')
        return render_template('toxic.html',url=img_url)
    if request.method == 'POST':
        comment = request.form['comment']
```

```
        new_row = {'comment_text':comment}
        user_df = pd.DataFrame(columns = ['comment_text'])
        user_df = user_df.append(new_row,ignore_index = True)
        user_df['comment_text'] = user_df['comment_text'].map(lambda com : clean_text(com))
        user_text = user_df['comment_text']
        user_features = loaded.transform(user_text)
        cols_target = ['obscene','insult','toxic','severe_toxic','identity_hate','threat']
        lst= []
        mapper = {}
        for label in cols_target:
            filename = str(label+'_model.sav')
            filename
            model = pickle.load(open(filename, 'rb'))
            print('... Processing {}'.format(label))
            user_y_prob = model.predict_proba(user_features)[:,1]
            print(label,":",user_y_prob[0])
            lst.append([label,user_y_prob])
        print(lst)
        final=[]
        flag=0
        for i in lst:
            if i[1]>0.5:
                final.append(i[0])
                flag=2
        if not len(final): text = "Yaayy!! The comment is clean"
            img_url = url_for('static',filename = 'images/happy.png')
            flag=1
            print(img_url)
        else: text="The comment is "
            for i in final:
                text = text+i+" "
            img_url = url_for('static',filename = 'images/toxic.png')
        print(text)
        return render_template('toxic.html',ypred = text,url= img_url,flag=flag)
if __name__ == '__main__':
    app.run( host = 'localhost', debug = True, threaded = False)
```