CYBER BULLYING DETECTION USING MACHINE LEARNING

Ancy Alexander
Department of Computer Applications
Amal Jyothi College of Engineering
kanjirappally,india
ancyalxndr143@gmail.com

Grace Joseph
Department of Computer Applications
Amal Jyothi College Of Engineering
kanjirappally,India
gracejoseph@amaljyothi.ac.in

Abstract—**As the number of people using social media has been increasing exponentially, cyberbullying has evolved as a kind of bullying that occurs through electronic messages. Bullying has been a part of society for as long as anybody can remember. Machine learning can be used to recognize the linguistic characteristics of bullies and so construct a model that can detect cyberbullying automatically. We study strategies to detect offensive words on social sentences, paragraphs while distinguishing it from general profanity in this paper, which includes a thorough assessment of some previous research on cyberbullying detection tools. Using supervised classification methods and a manually annotated open-source dataset, we want to develop lexical baselines for this task. This study presents a complete and structured overview of automatic offensive words detection, as well as a systematic comparison of a few of its existing methodologies, as well as an insightful evaluation of some published research on cyberbullying detection approaches.**

Keywords—cyberbullying,profanity,Bullying

I. INTRODUCTION

Cyberbullying is defined as the intentional use of technological means to perpetrate harmful actions against a target victim. As more people use social media, cyberbullying has emerged as a form of bullying that occurs over email. Cyberbullying can include sending threatening emails, texts, or instant messaging. Messages that are neutral to the point of harassment are frequently received. Making disrespectful remarks about another person on social media. Spreading rumours is a sort of rumour promotion, and persons with learning disabilities, as well as those who are physically challenged, are more vulnerable to cyberbullying. Online bullying has several distinct qualities, such as the ability to remain anonymous. It is possible to reach a wide audience without exposing your identity. It was only a matter of time after the internet's introduction that bullies took use of this new and lucrative medium. Bullies have been able to carry out their evil acts anonymously and with a large distance between themselves and their targets thanks to technologies like email and instant messaging, and given the negative effects of cyberbullying on victims, finding effective ways to detect and prevent it is critical.

As a result of the tremendous expansion in user-generated digital material, particularly on social networks, the volume of inflammatory words is continuously expanding. In recent years, research into detecting cyberbullying has grown in response to the spread of cyberbullying throughout social media and its detrimental influence on the younger generation. There is a growing amount of research on how to detect cyberbullying using automated methods. These methods compare textual data with chosen attributes and apply machine learning and natural language processing techniques to identify cyberbullying exchange features and automatically detect cyberbullying.

A system's software and application architecture that allows it to detect potentially harmful communications transmitted by a user. The system will examine each connection and determine if it is malicious or not. If the message is considered to be malicious, the system will use a machine learning technique to transform the contents into a "good" message. Furthermore, if the word is offensive, the system will convert it to censord format. The user receives a modified message that contains very little of the original hazardous content. To detect and assess cyberbullying, we apply Natural Processing Language and Emotional Analysis in this work.

II. LITERATURE REVIEW

[1] J. Yadav, et al. It proposes a new method for identifying cyberbullying on social media platforms that is based on the BERT model and a single linear neural network layer as a classifier on top. The form is trained and tested in the Spring Form forum and the Wikipedia dataset. The proposed model has a performance accuracy of 98 percent for the quartile model data set and 96 percent for the Wikipedia data set, which is greater than the previously utilized techniques.

[2] Yao et al ,identifies the pattern of cyberbullying on social media, which is defined as a sequence of abusive communications delivered from the bully to the victim with the goal to cause harm. Reduce the number of features used in classification while retaining high accuracy by using a sequential hypothesis test procedure. This strategy places a premium on precision, timeliness, and scalability.

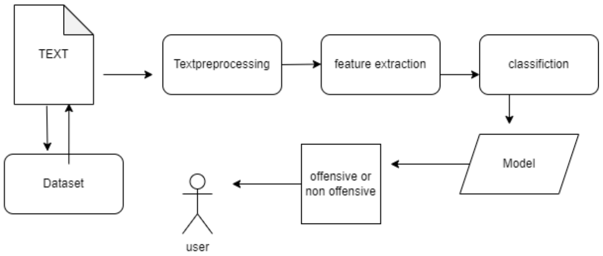
[3] Silva et al. It outlines the concept for an app named Bully Blocker and establishes a system for identifying cyberbullying based on psychological research. The software will tell the user's parents if cyberbullying is detected. It is aimed for teenagers, and it uses obsolete detecting methods. By operating as a platform, Facebook, on the other hand, has the opportunity to grow. a data collection software that employs machine learning to classify data.

[4]. G. A. León-Paredes et al. have shown how they used Natural Language Processing (NLP) and Machine Learning to construct a cyberbullying detection model (ML). The machine learning techniques Nave Bayes, Support Vector Machine, and Logistic Regression were used to create a Spanish cyberbullying prevention system (SPC). The dataset for this study was obtained from Twitter..

[5] Russell and Rasul al-Islam and their colleagues focus on removing the notes posted on social media, as well as examining whether these notes carry objectionable meaning. Offensive, hate speech, and none of the above are the three categories of reactions. The suggested approach categorizes observations on species with a 93 percent accuracy. Latent semantic analysis (LSA) was utilized as a feature selection method to decrease the amount of input data. In addition to regular feature extraction approaches, usual feature extraction methods such as encoding, N-gram, and TF-IDF were utilized to expose the primary findings. We used three different machine learning models to accomplish the calculations, analysis, forecasting, and forecasting: Random Forest, Logistic Regression, and Support Vector Machines (SVMs).

III. METHODOLOGY

Using the data set we collected, the goal of this research study is to construct a machine learning model that can determine whether a comment is offensive or not. Both good and negative messages are appropriately detected by the model. The data set is split into two parts: training and testing. The system will determine whether the user has entered a good or bad comment or offensive words if the user has entered a bad comment or offensive terms. This form's output can be used to make a news classification. The goal of this research is to develop a machine learning model that can reliably assess whether a comment or message is good or harmful based on the data available. If the system analyses the user entering any offensive terms and subsequently converts the offending words into a controlled format, the model must be able to discriminate between offensive and non-offensive data.



- Data Collection

As a dataset, we use our own csv and text files for this work. So we can use new words or sentences.

- Data preprocessing

Using the nltk library and profanity, the following preprocessing actions were implemented:

1. Word markup: A token is one thing that serves as the basis for a sentence or paragraph. Word encoding divides text into individual words in a list.

2.The stop words are filtered with python libraries ,nltk corpus stopwords .words('english'),which retrieves the list of stop words from the English dictionary and then removes them. Stop words are unimportant terms such as "the", "a", "an" and "in" that have no bearing on the interpretation of the data to be evaluated.

3) To get rid of punctuation, we only save characters that are not punctuation marks, as specified by the punctuation string.

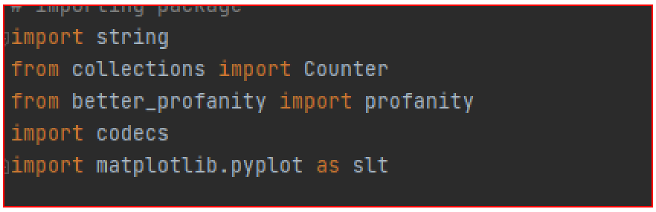
- Feature extraction

We then used the Python libraries and sentiments classification Transform function to extract the features so that they could be used with machine learning techniques.it determines if the words extracted as offensive or non offensive and then the offensive words are censored

IV. BUILD MODEL

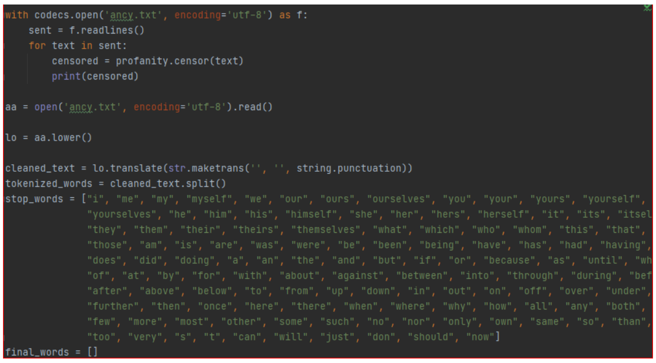
Model building is the main step involved in identifying cyberbullying. Here we use some commands for the purpose of building a model. The important steps are :

1.Importing Package



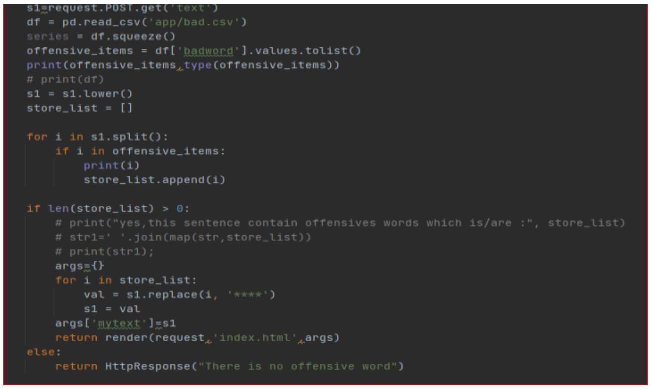
```
# importing package
import string
from collections import Counter
from better_profanity import profanity
import codecs
import matplotlib.pyplot as slt
```

2.Add csv file or text file of dataset and remove unwanted data,stopwords(eg,review data ,reviewer name,\n).



```
with codecs.open('ancy.txt', encoding='utf-8') as f:
    sent = f.readlines()
    for text in sent:
        censored = profanity.censor(text)
        print(censored)

aa = open('ancy.txt', encoding='utf-8').read()

lo = aa.lower()

cleaned_text = lo.translate(str.maketrans('', '', string.punctuation))
tokenized_words = cleaned_text.split()
stop_words = ["i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself",
              "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itsel
              "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that",
              "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having",
              "does", "did", "doing", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "wh
              "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "bef
              "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under",
              "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both",
              "few", "more", "most", "other", "some", "such", "no", "nor", "only", "own", "same", "so", "than",
              "too", "very", "s", "t", "can", "will", "just", "don", "should", "now"]
final_words = []
```

3. After preprocessing and feature exatraction of data, if we found any offensive words then that word should be censord.



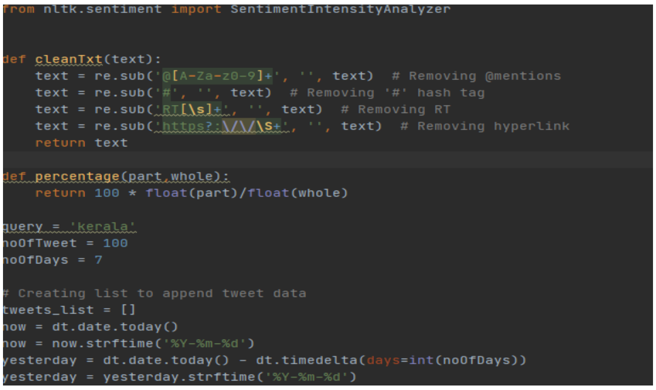
```
s1=request.POST.get('text')
df = pd.read_csv('app/bad.csv')
series = df.squeeze()
offensive_items = df['badword'].values.tolist()
print(offensive_items,type(offensive_items))
# print(df)
s1 = s1.lower()
store_list = []

for i in s1.split():
    if i in offensive_items:
        print(i)
        store_list.append(i)

if len(store_list) > 0:
    # print("yes,this sentence contain offensives words which is/are :", store_list)
    # str1=' '.join(map(str,store_list))
    # print(str1);
    args={}
    for i in store_list:
        val = s1.replace(i, '****')
        s1 = val
    args['mytext']=s1
    return render(request,'index.html',args)
else:
    return HttpResponse("There is no offensive word")
```

V. RESULTS

After multiple testing on network architecture and data optimization, we came up with some fascinating and inconsistent results. Our first network, the Sentiment Analysis Network, did an excellent job at assessing whether a message is good or terrible, but there is still space for improvement. Based on the supplied data set, the result reveals that offensive words can be classed as good or bad data. There are several areas of terrible and good in the dataset. This is classified, and the model successfully detects negative remarks, after which it is converted to a censored format.

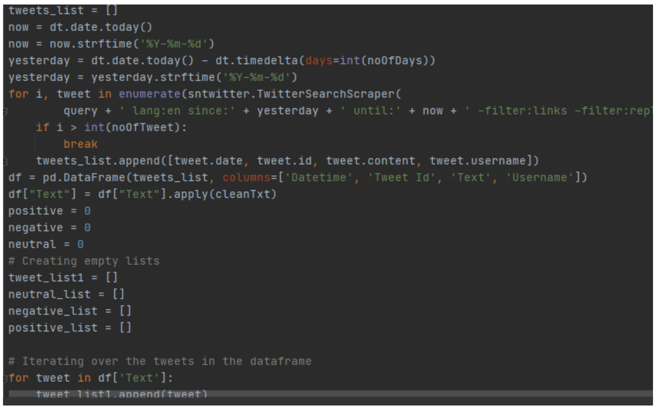
Cyberbullying detection in twitter

We use a dataset from Twitter and anticipate the comment or message we've provided to be used to pre-process the data, extract the features using emotional analysis, and then classify the given word as positive, negative, or neutral.



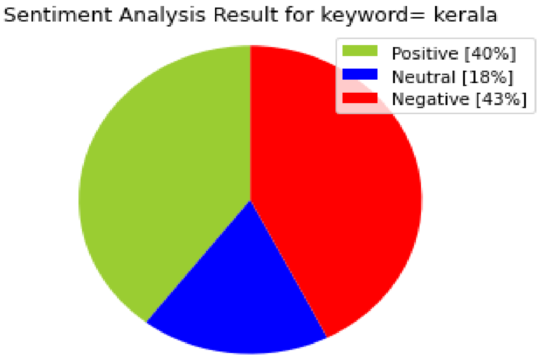
```
from nltk.sentiment import SentimentIntensityAnalyzer


def cleanTxt(text):
    text = re.sub('@[A-Za-z0-9]+', '', text)  # Removing @mentions
    text = re.sub('#', '', text)  # Removing '#' hash tag
    text = re.sub('RT[\s]+', '', text)  # Removing RT
    text = re.sub('https?:\/\/\S+', '', text)  # Removing hyperlink
    return text

def percentage(part,whole):
    return 100 * float(part)/float(whole)

query = 'kerala'
noOfTweet = 100
noOfDays = 7

# Creating list to append tweet data
tweets_list = []
now = dt.date.today()
now = now.strftime('%Y-%m-%d')
yesterday = dt.date.today() - dt.timedelta(days=int(noOfDays))
yesterday = yesterday.strftime('%Y-%m-%d')
```

- Using sentimental analyser we split ,train and test the dataset of twitter



```
tweets_list = []
now = dt.date.today()
now = now.strftime('%Y-%m-%d')
yesterday = dt.date.today() - dt.timedelta(days=int(noOfDays))
yesterday = yesterday.strftime('%Y-%m-%d')
for i, tweet in enumerate(sntwitter.TwitterSearchScraper(
        query + ' lang:en since:' + yesterday + ' until:' + now + ' -filter:links -filter:rep
    if i > int(noOfTweet):
        break
    tweets_list.append([tweet.date, tweet.id, tweet.content, tweet.username])
df = pd.DataFrame(tweets_list, columns=['Datetime', 'Tweet Id', 'Text', 'Username'])
df["Text"] = df["Text"].apply(cleanTxt)
positive = 0
negative = 0
neutral = 0
# Creating empty lists
tweet_list1 = []
neutral_list = []
negative_list = []
positive_list = []

# Iterating over the tweets in the dataframe
for tweet in df['Text']:
    tweet_list1.append(tweet)
```

- The result will plot in a graph



VI. CONCLUSION

As a result of the tremendous expansion in user-generated digital material, particularly on social networks, the volume of inflammatory words is continuously expanding. In recent years, research into detecting cyberbullying has grown in response to the spread of cyberbullying throughout social media and its detrimental influence on the younger generation. There is a growing amount of research on how to detect cyberbullying using automated methods. These methods compare textual data with chosen attributes and apply machine learning and natural language processing techniques to identify cyberbullying exchange features and automatically detect cyberbullying. We use numerous unpleasant phrases, sentences, and comments in this day and age, and sometimes it affects the other person or the victim, thus seeing these words directly hurts their feelings. To get around this, we determine that the word is offensive and replace it with another; we do this with nltk and emotional analysis. It's a simple code for identifying objectionable terms, so it'll come in handy in the future.

REFERENCE

[1] J. Yadav, D. Kumar and D. Chauhan, Cyberbullying Detection using Pre-Trained BERT Model, ICESC, pp.1096-1100, doi: 10.1109/ ICESC48915.2020.9155700. (2020)

[2] Yao, Mengfan, Charalambos Chelms and Daphne? Stavroula Zois. Cyberbullying ends here: Towards a robust detection of cyberbullying in social media. World Wide Web Conference. 2019.

[3] Y.N Silva, D.L Hall, C. Rich, BullyBlocker: Towards an interdisciplinary approach to identifying cyberbullying. Social network analysis and mining. 8 (2018), doi: 10.1007/s13278-018-0496-z.

[4] G.A León-Paredes et al. , Hypothetical Detection of Twitter Cyberbullying by Natural Language Processing and Machine Learning in Spanish, CHILECON pp.1-7, doi:10.1109/CHILECON47746.2019.8987684. (2019)

[5] R. R. Dalvi, S. Baliram Chavan and A. Halbe, Detecting A Twitter Cyberbullying Using Machine Learning, ICICCS, pp. 297-301, doi: 10.1109/ICICCS48265.2020.9120893.