International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 05 | May - 2023 SJIF 2023: 8.176 ISSN: 2582-3930

FIGHT BACK CYBERBULYYING WITH NLP ON SNS USING CODE TEXT APPROACH

Kunal Gupta[1], Kanchan Ghule[2], Ankita Dhondge[3]*, Shruti Dive[4], Aniruddha S. Rumale[5]

[1,2,3,4,] *Student,* [5]*Professor, Information Technology Department, Sandip Institute Of Technology And Research Centre, Nashik, India*

ABSTRACT: This research paper is about the important problem of rising hate and offensive Contents made against people or communities on social media is addressed by this proposed system. With the increasing use of social media, cyberbullying behaviour has received more and more attention. Cyberbullying may cause many serious and negative impacts on a person's life and even lead to teen suicide. To reduce and stop cyberbullying, one effective solution is to automatically detect bullying content based on appropriate machine learning and natural language processing techniques. The impacts of cyber bullying on social media are horrifying, sometimes leading to the death of some unfortunate victims. The behaviour of the victims also changes due to this, which affects their Emotions, self-confidence and a sense of fear is also seen in such people. Thus, a complete solution is required for this problem. Cyber bullying needs to stop. The problem can be tackled by detecting and preventing it by using a machine learning approach, this needs to be done using a different perspective.

Keywords: Hate Speech; Social Networking Site; Natural Language Processing; Text Classification; Machine Learning

I. INTRODUCTION

With the increasing use of social media, cyberbullying behaviour has received more and more attention. Cyberbullying may cause many serious and negative impacts on person's life and even lead to teen suicide. To reduce and stop cyberbullying, one effective solution is to automatically detect bullying content based on appropriate machine learning and natural language processing techniques. However, many existing approaches in the literature are just normal text classification models without considering bullying characteristics.

II. RELATED WORK

A social network perspective to the issue of cyber aggression or cyberbully, on the social media platform Twitter. Cyber aggression is particularly problematic because of its potential for anonymity, and the ease with which so many others can join the harassment of victims. Utilizing a comparative case study methodology, the authors examined thousands of Tweets to explore the use of denigrating slurs and insults contained in public tweets that target an individual's gender, race, or sexual orientation. Findings indicate cyber aggression on Twitter to be extensive and often extremely offensive, with the potential for serious, deleterious consequences for its victims [1].

The tweets are annotated with the language at word level and the class they belong to (Hate Speech or Normal Speech); a supervised classification system for detecting hate speech in the text using various character levels, word levels, and lexicon-based features. With the recent surge in the amount of user-generated social media data, there has been a tremendous scope in automated text analysis in the domain of computational linguistics. The popularity of opinion-rich online resources like review forums and microblogging sites has encouraged users to express and convey their thoughts all across the world in real time [2].

Code-mixed NLP has been extensively studied. As pre-trained transformer based architectures are gaining popularity, they observe that real code-mixing data are scarce to pre-train large language models. They present L3Cube-HingCorpus, the first large-scale real Hindi-English code mixed data in a Roman script. It consists of 52.93M sentences and 1.04B SITRC, Department of Information Technology Engineering 2022-23 Page 12 tokens, scraped from Twitter. They further present HingBERT, HingMBERT, HingRoBERTa, and HingGPT. The BERT models have been pre-trained on code mixed HingCorpus using masked language modelling objectives. They show the effectiveness of these BERT models on the subsequent downstream tasks like code-mixed sentiment analysis, POS tagging, NER, and LID from the GLUECoS benchmark. The HingGPT is a GPT2 based generative transformer model capable of generating full tweets [3].

Code-mixing is a linguistic phenomenon where multiple languages are used in the same occurrence that is increasingly common in multilingual societies. Code mixed content on social media is also on the rise, prompting the need for tools to automatically understand such content.

Automatic Parts-of-Speech (POS) tagging is an essential step in any Natural Language Processing (NLP) pipeline, but there is a lack of annotated data to train such models. In this work, they present a unique language tagged and POS-tagged dataset of code-mixed English Hindi tweets related to five incidents in India that led to a lot of Twitter activity [4].

Cyberbullying on social networking sites is an emerging societal issue that has drawn significant scholarly attention. The purpose of this study is to consolidate the existing knowledge through a literature review and analysts. They first discuss the nature, research patterns, and theoretical foundations. They then develop an Integrative framework based on social cognitive theory to synthesize what known and identify what remains to be learned, with a focus on the triadic reciprocal relationships between perpetration, victims, and bystander. They discuss the key findings and highlight opportunities for future research they conclude this paper by noting [5].

III. PROBLEM STATEMENT

Cyberbullying as the name implies is the use of cyberspace as a mechanism to bully others known or unknown to the bully. Cyberbullying has caused significant issues for those involved ranging from extreme displays of anger to suicide attempts. Hate speech detection in social media texts is an important Natural language processing task, which has several crucial applications like sentiment analysis, investigating cyberbullying, and examining socio-political controversies. While relevant research has been done independently on code-mixed social media texts and hates speech detection, the work on detecting hate speech in Hindi English code-mixed social media text is not feasible for the current time & age.

IV. METHODOLOGY

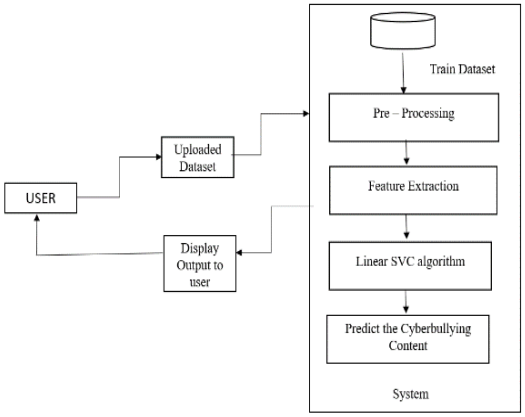

Fig 1. Proposed System

A. Data Collection

We Collected dataset from different social media platforms involves accessing their APIs (Application Programming Interfaces) or utilizing web scraping techniques, where applicable. Here's an overview of the process for collecting data from some popular social media platforms:

Twitter:

Use the Twitter API to access tweets, user profiles, and other relevant data. The API provides functionalities for searching tweets, filtering by specific criteria, and accessing user information.

Facebook:

Accessing data from Facebook requires complying with their platform policies and using their Graph API.

Instagram:

Instagram's API has limited access to public data and is primarily intended for business accounts and developers with specific use cases.

You can use the API to retrieve information like user profiles, posts, comments, and interactions.

		Label	Tweet
2	0	0	haa jaise tum bhi abhi
3	1	1	banti hai empowered woman feminism pe gyan pelti hai aur din bhar roti rehti hai pahle rona band kar madarchod
4	2	1	ab usko chhod mjse bat kr tera baap aa gya hai ab to ldki ko beech me la ra hai madarchod
5	3	1	punjab in madarchodon ko khila raha hai nokrian day raha hai aur yeh imran ma
6	4	1	agar koi bole ki ja ke chill maar to madarchod ki gand maar lene ka
7	5	1	main jutt punjabi hoon aur paka league madarchod imran ki punjab say nafrat clear hai
8	6	1	to bhosdike tere baap ka kya ja raha hai tu apna ghar dekh na madarchod
9	7	1	sunny leone bana ke chodege teri maa ko samjh me aaya madarchod neem ka patta kadva hai
10	8	1	pata nahi aise sanghi kutte ko bha
11	9	1	screw the law of the land if find this chutiya madarchod mulla will lynch him murder him cut into millions of pieces and ha
12	10	0	boy donated blood to his gf ntwo months later they broke up nboy mujhe mera khoon lauta de ngirl throw
13	11	1	bc madarchod south africa england main man of the match award le chuke
14	12	1	madarchod congresi or bjp dono ki sarkar bani baad me wade bhul gye rajasthan me hamri city tak singal
15	13	1	madarchod socialify wale log
16	14	1	madarchod kya hai yeh
17	15	1	kis madarchod ki he giri hui harkt
18	16	1	teri maa behan ko har rooj chood rahe hai tum khayal rakho wamna aise hi chudawati rahati hai madarchod
19	17	1	deewar superhit nai aaj ke time ke hisaab se atbb hai chutye dont compa

Fig 2. Snapshot of Dataset

B. Data Pre-processing

Data pre-processing plays a crucial role in fighting back against cyberbullying. By cleaning and preparing the data before analysis, you can improve the accuracy and effectiveness of models or algorithms used to detect and address instances of cyberbullying.

We have employed and replaced for the current work using pre-processing techniques such tokenization, stemming, and deleting stop words.

Each word in a sentence is divided into "tokens" using tokenization. This gives us specific words or terms.

Stemming locates the sentence's root word and replaces it with the original word. For instance, stemming will change the word "Leaves" to "Leaf."

Stop words are terms in natural language processing that don't help the machine learning algorithm learn new things. For instance, stop words like "is," "the," "an," and "in" need

to be eliminated from the dataset. So, we deleted all the stop words by running through our dataset.
The whole stop words in the manuscript, such as -is, am, and are, are eliminated during the stop wards removal process.

C. Feature Extraction

The majority of machine learning algorithms use mathematical concepts from areas like statistics, algebra, calculus, and others.
They assume the data to be numerical, such as a two-dimensional array with rows acting as instances and columns acting as features. The issue with natural language is that the data is in the form of raw text, so the text needs to be converted into a vector.
Text vectorization is the term used to describe the process of turning text into a vector. It's a key step in natural language processing because no machine learning method, not even computers, is capable of understanding a text.
Text can be converted into vectors with the help of the text vectorization technique TF-IDF vectorizer, which is a well-liked method for standard machine learning method.

• TF- Term Frequency
Term frequency simply means how many times or how frequently the term is occurring in document.

$$TF = \frac{NO.of\ times\ term\ appears\ in\ document}{Total\ terms\ in\ the\ document}$$

• IDF- Inverse document frequency.
The weight of uncommon words is IDF. High IDF scores are assigned to words that are infrequently used in the corpus.

$$IDF = \frac{\text{Log (Total no.of document)}}{\text{(no.of documents with term in it).}}$$

After that calculate TFIDF
TFIDF = TF*IDF
When the value of TFIDF is higher then the word is most significant.

D. Classification

An important issue for Supervised Machine is text classification. A supervised machine learning model is the support vector classifier (SVC). According to our analysis, SVC performs the best in classifying text. Linear SVC (Support Vector Classifier) is a common approach for classification tasks, including fighting back against cyberbullying. Linear SVC is a linear classification algorithm that works well with high-dimensional data and can effectively separate different classes. Text is transformed into an appropriate representation and supplied into SVC. When the training data and feature vector have very high dimensionalities, it performs well. TFIDF creates algorithms that can be used to mathematically model complicated patterns and prediction issues by using the way the brain processes information.

	Algorithm	F1 Score: Test	F1 Score: Train	Accuracy: Test	Precision: Test	Recall: Test	Prediction Time	Accuracy: Train	Precision: Train	Recall: Train	Training Time
0	LinearSVC	0.890110	0.998216	0.855942	0.893382	0.886861	0.001002	0.997635	0.998216	0.998216	0.006001
1	BaggingClassifier	0.886171	0.991007	0.854742	0.914563	0.859489	0.043875	0.988173	0.999093	0.983051	5.412974
2	AdaBoostClassifier	0.882969	0.916978	0.852341	0.922465	0.846715	0.094437	0.894737	0.960899	0.876896	2.438512
3	SGDClassifier	0.881132	0.998214	0.848739	0.912109	0.852190	0.002040	0.997635	0.999106	0.997324	0.007533
4	DecisionTreeClassifier	0.860294	0.999554	0.817527	0.866667	0.854015	0.001961	0.999409	1.000000	0.999108	0.324970
5	LogisticRegression	0.841584	0.924560	0.769508	0.768072	0.930657	0.001002	0.893554	0.871937	0.983943	0.047990
6	MultinomialNB	0.836570	0.903854	0.757503	0.751453	0.943431	0.001002	0.859846	0.828869	0.993756	0.001957

Fig 3. Accuracy Table

V. RESULT

To determine whether comments are considered bullying or not, the "Fight Back Cyberbullying" project would employ various techniques and algorithms.
Sentiment analysis involves determining the emotional tone of a comment. The project could employ natural language processing techniques to analyze the sentiment expressed in the comments, identifying those with negative or aggressive tones that may indicate bullying. Sentiment analysis involves determining the emotional tone of a comment. The project could employ natural language processing techniques to analyze the sentiment expressed in the comments, identifying those with negative or aggressive tones that may indicate bullying.
By training machine learning models on large datasets of labeled bullying comments, the project can develop algorithms that can automatically classify new comments as bullying or non-bullying based on learned patterns and features.

Fig 4. Snapshot of Result

VI. CONCLUSION

Cyberbullying is a serious issue, and likely any form of bullying it can have long term effects on its victims. Our project will grow and help individuals to be aware of Cyber bullies. Parents, teachers and children must work together to prevent Cyberbullying and to make Internet a safe place for all. To have social harmony and reduced depression and other mental illness caused by cyberbullying. Thus Creating counter measures needs an active funnel to take proper actions against the cyber bullies and to create an supervised classification system for detecting hate speech in the text

using various character levels, word level attributes as well as emoji in Hindi-English code text language to create an open source action and plugin system using cloud technology to take suitable actions on the results.

VII. FUTURE SCOPE

The fight against cyberbullying is an ongoing battle, and there are several potential future scopes for the "Fight Back Cyberbullying" project. Here are a few possibilities:

Collaboration with Social Media Platforms: Collaborating with social media platforms and online communities can be beneficial for implementing proactive measures to prevent cyberbullying.

Multilingual Support: Cyberbullying is a widespread issue worldwide, and adapting the project to different languages would increase its impact and effectiveness.

Natural Language Processing (NLP) Advancements: NLP techniques can play a crucial role in analyzing and understanding the context of cyberbullying incidents. Further advancements in NLP algorithms and models can enable more accurate sentiment analysis, and identification of implicit bullying in online conversations.

VIII. REFERENCES

[1]Glenn Sterner, Diane Felmlee. “The Social Networks of Cyberbullying on Twitter.”

[2]Aditya Bohra, Deepanshu Vijay, Vinay Singh. “A Dataset of Hindi -English Code-Mixed Social

Media Text for Hate Speech Detection.” [3]Ravindra Nayak and Raviraj Joshi. “L3Cube-HingCorpus and HingBERT: A Code Mixed Hindi-English Dataset and BERT Language Models.” [4]Kushagra Singh, Indira Sen, Ponnurangam Kumaraguru. “A Twitter Corpus for Hindi-English Code Mixed POS Tagging.”

[5] Tommy K.H. Chan, Christy M.K. Cheung, Zach W.Y. Lee. “Cyberbullying on social networking sites: A literature review and future.”

[6]Dr.A.K.Jaithunbi, Gollapudi Lavanya, Dondapati Vindhya Smitha , Bandi Yoshna. “Detecting Twitter Cyberbullying Using Machine Learning.”

[7]Rui Zhao, Anna Zhou, Kezhi Mao.“Automatic Detection of Cyberbullying on Social Networks based on Bullying Features.”

[8]Shivang Chopra, Ramit Sawhney. “Hindi-English Hate Speech Detection: Author Profiling, Debiasing, and Practical Perspectives.

[9]Pavitar Parkash Singh, Vijay Kumar, Majid Sadeeq. “Cyber Bullying as an Outcome of Social Media Usage: A Literature Review”

[10]Renee Garett, Lynwood R. Lord, and Sean D. Young. “Associations between social media and cyberbullying: a review of the literature.”