International Journal of Interdisciplinary Innovative Research &Development (IJIIRD)
ISSN: 2456-236X
Vol. 07 Issue 02 | 2023

Cyberbullying Detection Using Machine Learning and Deep Learning

Dr. Shailesh D. Nandgaonkar[1], Kusum Chauhan[2], Dnyaneshwari Kumthekar[3], Pravin Gaikwad[4], Vishal Dohale[5]

[1]*Associate Professor, Information Technology, Alamuri Ratnamala Institute of Engineering & Technology, Maharashtra, India*
[2,3,4,5]*UG scholar, Computer Engineering, Alamuri Ratnamala Institute of Engineering & Technology, Maharashtra, India*

ABSTRACT

Cyberbullying is a traumatic online misbehavior with disquieting consequences. It appears in different forms, and in utmost of the social networks, it's in textual format. further than 1. ninety-six billion are certain to have an ineluctable social life. still, the developing decade acts extreme demanding situations and the online geste of guests had been deposited to question. adding cases of importunity and bullying together with cases of casualty were a severe issue. Automatic discovery of similar incidents requires intelligent systems. utmost of the present exploration has approached this trouble with traditional contrivance studying fashions and the maturity of the developed models in these studies are adaptable to a single social network at a time. In recent studies, deep literacy grounded models have set up their way in the discovery of cyber-bullying Incidents, claiming that they're suitable to triumph over the constraints of the traditional models, and enhance the discovery performance. Though, numerous old- council fashions are to be had to govern the mishap, the want to rightly classify the bullying continues to be delicate. To efficiently screen the bullying withinside the digital area and to help the murderous fate with perpetration of Machine Learning and Language processing. In this paper, we suggest a system to offer a double class of cyberbullying. Our fashion makes use of a progressive idea of CNN for textual content evaluation but the present strategies use a naive system to offer the answer with much lower delicacy. An being dataset is used for trial and our frame is vindicated with other being procedures and is set up to give better delicacy and bracket.

Keyword: - Machine Learning, Deep Learning

1. INTRODUCTION

Cyberbullying has significantly endangered youths' psychosocial wellbeing over the past ten years. The intentional use of digital equipment to commit harmful acts against a target victim is termed as cyber bullying. In other words, it is the implementation of technology to annoy, threaten, embarrass, or make offensive comments about someone. Apart from children and teenagers, adults have also been witnessed to be engaged in such activities, who commit such crimes and then later face harsh legal sanctions, like prison terms. On the other hand, Cyberbullying does not necessarily involve using physical force or direct face-to-face communication, in contrast to more traditional "bullying" behaviors. Cyberbullying can be done by anyone using any device with an Internet connection. In such a scenario, bullies can come from close friends of children and young people as well as from an anonymous source. Following are the places where cyber bullying occurs the most:

- Social networking sites such as Twitter, Instagram
- Direct messages on a mobile device
- Emails

According to research scholars, the prevalence of cyber bullying encompasses the online behavior of attackers which describes their reason for impersonation and stalking. Therefore, they have created two major categories of cyber bullying; namely as:

- Written cyber bullying: this process involves bullying and offense being done verbally
- Visual cyber bullying: this process involves offense being done in the form of sharing videos and pictures0SS

Additionally, a number of empirical studies have considered the connection between the various online behaviors of teenagers and cyberbullying. Use of social networking sites and online gaming are two primary components in particular that are linked to an increased risk of cyberbullying. According to research done by the Cyberbullying Research Center (J. W. Patchin, 2014), one in six teens engaged in cyberbullying in 2018 and one in five teens were the target of it. Also, their research demonstrates that online bullying is on an increase across all studies. Inferring from the studies, they calculated those 2.85 million teenagers, across the country experienced cyberbullying in 2019. Further, this number is expected to exponentially increase in the coming years. In addition to this, the struggle for parents to recognize and identify the bullying is one of the main issues that contribute to cyberbullying (T. P. Pope). Only one in ten teenagers will ever report it to an adult, according to a research. Due to the absence of reporting and the lack of any outward signs of cyberbullying, it is very

likely that it will go unobserved if a child's online interactions are not physically monitored. This is one of the major reasons that the youth tend to commit suicide.

2. MACHINE LEARNING BASED CYBER BULLYING DETECTION

The system model was developed using various machine learning algorithms like Logistic Regression, random forest, and SVM's. To begin with, the required dataset was collected from the relevant data repository. The data was then subjected to pre-processing and feature extraction techniques. To verify the sentiments of the comments made, semantic analysis was performed, and a score was calculated, based on which the comment was classified as either bullied or non-bullied. Tokenization and N-Gram analysis were carried out during the feature extraction process. The model achieved an impressive accuracy of 93 percent.

In a study conducted by Patchin et al. (2006), the cyber bullying process was discussed in detail, along with its negative impacts. The literature survey presented in the research paper briefly touched upon the workings of this topic in various other domains. The primary focus of the proposed model was to define the different types of cyber bullying that occur on the internet platform and shed light on its adverse effects on the youth. The bullying took place on a public platform, and the harsh comments passed in the form of electronic texts had a direct negative impact on the mental state of the individual reading them. The author also delved into the psychological impacts of cyber bullying on individuals interacting on such platforms. Therefore, to prevent such attacks from occurring, the author proposed a machine learning-based method for detecting them on the internet. The fundamentals of machine learning were used to detect such comments and prevent their occurrence in the future. Feature extraction techniques were employed to extract and eliminate specific words that could be trained on the machine. The model was executed, and the data was split into a train and test set in an 80:20 ratio. Machine learning algorithms such as SVM, KNN, decision trees, and logistic regression were used to detect the cyber bullying. The results of the study showed that decision trees generated the best results, with the highest optimized accuracy.

In a similar vein, cyber bullying was observed to occur on YouTube, prompting K. Dinakar (2011) to conduct an experimental study to detect such cases. The authors noticed a growing trend of lewd commenting on certain sites that targeted individuals of a particular age group, causing significant mental distress and even suicidal tendencies. Therefore, the issue needed to be highlighted and resolved using computer-aided technology. Machine learning algorithms were used to collect data from the internet and depict the impact of bullying on youth. The data was classified as bullying or non-bullying, and it was observed that bullying also took place in the form of race, community, and religion-based comments directed towards college students. The authors proposed a model to detect and classify such comments as bullying or non-bullying, which could prevent cybercrime on the internet. The algorithms used in the study included logistic regression, decision trees, and ensemble learning. The authors conducted a comparative analysis and found that logistic regression produced the highest accuracy.

3. PROPOSED WORK

Decentralized Twitter Dapp is a software in which the customers can create bills on twitter and able to upload the tweets, delete the tweets and capable of ship the messages from one account to any other. As Block chain is decentralized generation and arise all the transactions with the assist of ethers for going on every and every transaction. In this paper, we used tools like Ethereum IDE, remix.ethereum.org and Ganache and MetaMask for deploying the contracts.

The smart contract will be created by the first user, who will then deploy it. After each new block is added to the chain, the Ethereum Virtual Machine, also known as the EVM, computes the state of the Ethereum network and executes smart contracts. The hardware layer and node network layer of Ethereum are on top of the EVM. The user will be able to log into Twitter using Metamask after the contract has been successfully deployed (the user must have Metamask with test ethereum on their wallet; they will log into Twitter after a transaction occurs; and the transaction will be stored on two platforms, one on Ether scan and the other on Sanity Database). After logging in, the user will be able to access the frontend, where they can easily add, delete, and tweet (although they still need ethereum for that). Additionally, we now have the mint option, which allows users to change their profiles to become NFTs. Overall, the user of our project needs Ethereum in order to complete all of their requests.

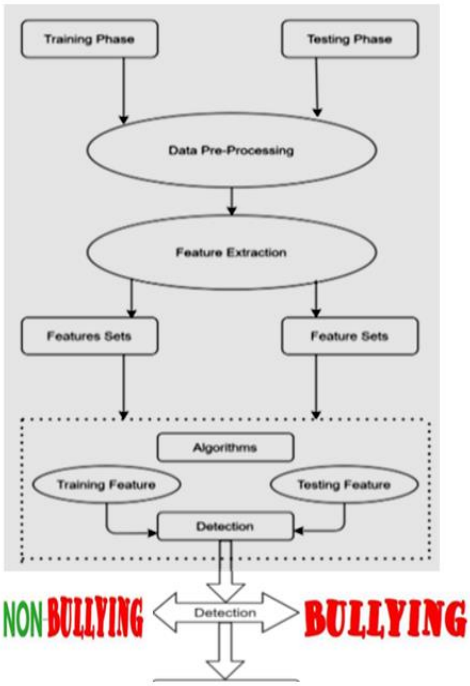

Fig -2: System Architecture of Cyberbullying Detection

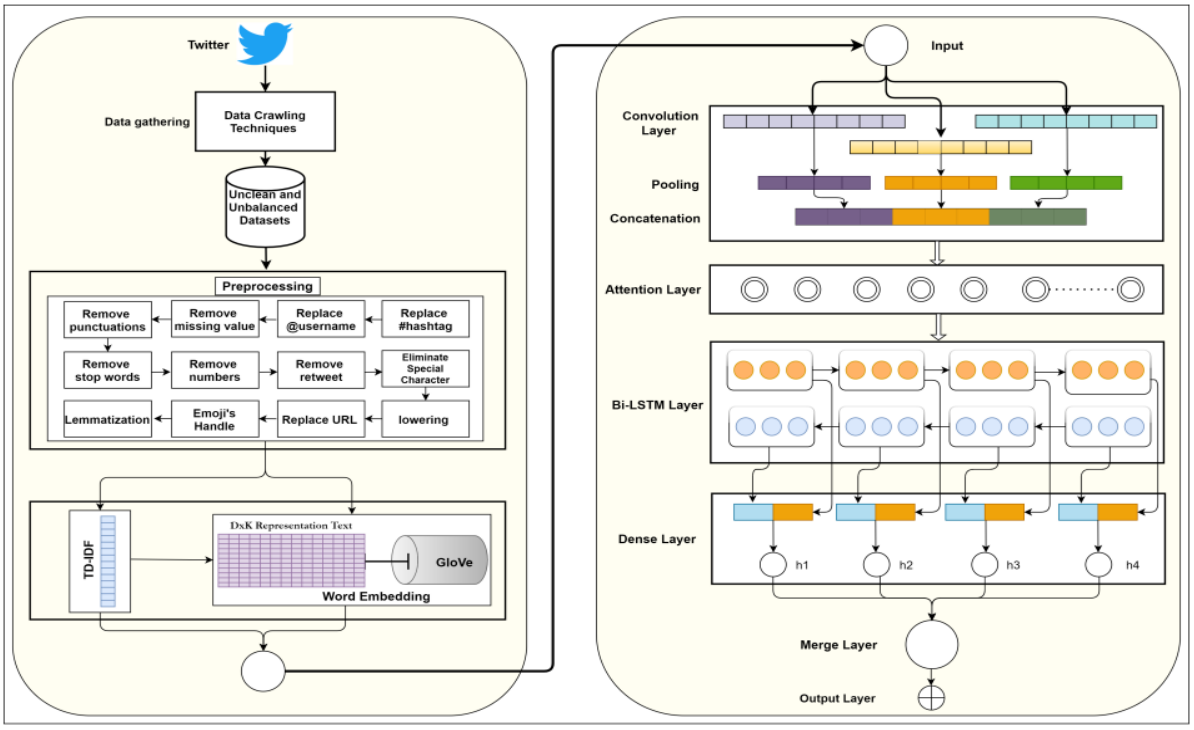

Fig -3: Working of Cyberbullying Detection

4. METHODOLOGY

The primary aim of the thesis is to detect the occurrence of online bullying that takes place on social media platforms such as Twitter. For this purpose, the authors have used the concepts of machine learning and deep learning algorithms to detect the same and further classify as bullying or non-bullying. Since the detection of bullying comments on Twitter depends on what sentiment a comment would depict, it is first necessary to comprehend different groups of terms that can be labelled as social features.

Following are the five main categories that are primarily used to classify an opinion based either on the sentiment targeted or the emotion of the comment so involved:

- Sentiment Features
- Sarcastic Features
- Syntactic Features
- Semantic Features
- Social Features

All the above mentioned features have been specifically categorized based on the existing systems and the literature survey so conducted. Apart from this it has also been recognized that the resulting features describes the comment to be either informative or descriptive. However, it is also important to note here that understanding what comment falls under what category is a major part as the final classification and pattern recognition depends on the features so mentioned.

The process of sentiment analysis includes the evaluation of sentiments such as anger, happiness, sadness, anxiety etc. evaluation of such sentiments is necessary to understand the meaning of a comment and further classify it as bullying or non-bullying. Once the detection of sentiments is done, the process is further followed by classification. This classification includes classifying the obtained sentiment as positive or negative. However, this classification of sentiment is done by calculating a sentiment score that could either be 0 or 1. If the sentiment score sums up to 1, the sentiment is calculated to be positive and is the score is evaluated as 0; the sentiment is classified as negative. On the other hand, context congruity is calculated using sarcastic features. The sarcastic features are responsible to portray a nonverbal behavior of an individual that is conveyed through a textual format. In such a scenario, a sentence may contain congruent as well as incongruent texts that would be embedded in the words and further needs to be detected for bullying cases. Sarcastic features are most likely to be present in a hidden format under a sentence and needs to be clearly detected. In some cases sarcastic features can also be represented through emoticons.

The detection of syntactic features majorly includes the identification of bad and insulting words that might occur in a sentence. Hence, the monitoring of such words is important and needs to be listed differently. Another important characteristic of syntactic features is the range of density in which the words might occur in a sentence. This enables an emphasis to be made on upper case and lower case letters as well. Since the usage of upper case letters indicates making a hate statement rudely, a comment written in lower case generally indicates an informative conversation being made. Similarly, the use of special characters plays a same role in determining syntactic features in a comment.

Semantic features on the other hand are majorly used to determine and establish a lexical relationship between two words in a sentence. Such features also includes mapping of different pronouns that can be used to explicitly refer to the comments being made by other individuals.

A social feature on the other hand, refers to the social behavior of the bully and includes analyzing the posts and comments made by him. However, analyzing only the posts would not suffice to detect the working nature and behavior of the bully, therefore analyzing his working and writing patterns becomes a must. In such a scenario, the social features of the bully come into picture and help to gain information of whom the bully would target next. In addition to this, identifying the victims based on the targets the bully makes also results into detecting and classifying hate speech.

This section of the thesis highlights the methodologies adopted to develop a detection framework. However, a general implementation of the framework includes three major chunks of implementation. The first chunk is labelled as NLP (Natural Language Processing), the second includes implementation through machine learning and the third involves the framework of deep learning. The first phase of implementation includes the process of data collection from certain repositories that offer Twitter dataset and contains those texts that involve bullying comments. In the next stage the comments and posts are detected for lewd comments and further labelled so that they can be fed to ML and DL based algorithms. This entire process is carried out through NLP and the data is further fed to the machine to detect the comments so collected.

Natural Language Processing (NLP)

In the physical world, comments and text often contain external characters and message. For instance, punctuation or numbers have no impact on the detection of bullying. Hence the data needs to be cleaned and the extracted comments needs to be prepared so that it can be further fed to the machine and deep learning algorithms. However, this phase includes cleaning the data in a variety of ways such as performing the process of tokenization, stemming and the elimination of unnecessary words such as stop words and punctuations. Following are the derived pre-processing techniques that need to be performed on the texts:

Bag of Words: Since, the raw texts cannot be used by machine learning algorithms directly. Therefore, they need to be converted to appropriate vectors before they are fed to the algorithms. Thus, for the subsequent stage, the processed data is transformed into a Bag of Words (BoW). It is a textual illustration that shows where specific words appear throughout a document. We use the Boolean values 0 for absence and 1 for presence when retrieving document vectors, and it involves a lexicon of known words and a measurement of the presence of known words. Each tweet will further be converted into a list of word counts. With CountVectorizer, this can be completed quickly and easily ()

Lexicon Approach: The lexicon-based approach, as its name suggests, makes use of definitions or vocabulary. In this stage, the semantic orientation or polarity of terms or phrases is used to determine the alignment for a document. The lexicon-based method does not necessarily call for storing a huge amount of data, unlike a machine learning approach. Using a vocabulary or dictionaries, it assesses the orientation of a piece of writing. The Semantic Orientation (SO) of a text serves as a gauge of individuality and viewpoint and quantifies the polarity and impact of individual sentences and phrases. These paragraphs collectively control the document's entire sentimental orientation. Opinion lexicon can be generated either manually or automatically. However, I t may take some time to develop the opinion lexicon manually. These details must be merged with other automated methods and fall primarily into one of two types of manual lexicons: the prevalent lexicon or the classification lexicon. The selection of sentiment lexicon and their conductivity can be quickly searched. Split terms, contraction words, and blind boolean words are included in the standard lexicon along with standard opinion words with the same sentiment value. For quick, precise sentiment analysis on large scales, a complete, high-quality lexicon is frequently required

Natural Language Toolkit (NLTK): The NLTK Python package is a complimentary, open-source collection of tools for software and data classification. The use of NLTK will be advantageous to linguists, engineers, instructors, teachers, experts, and designers who deal with textual data in naturally occurring language processing and text analytics. Accessing the interfaces of more than 50 textual and lexical resources is made simple by NLTK

TF-IDF: Term Frequency is referred to as an IDF. records with Inverse Document Frequency. It can be summed up as determining how pertinent a word is to a corpus or series of words in a text. The frequency of a word in the corpus offsets the meaning increase that occurs when a word appears more frequently in the text (data-set). The abbreviation of the word TF-IDF stands for Term Frequency Inverse Document Frequency which is responsible to maintain a set of records. This is a different feature that can be take into account for the implementation of our model. A statistical measure called TF-IDF (Term Frequency-Inverse Document Frequency) can assess how pertinent a word is to a file within a collection of documents. Every word in a bag of words is given equal weight, but in a TF-IDF, the words that appear more frequently should be weighed more heavily because they are better suited for classification.

5. CONCLUSIONS

The advancement of technology has undoubtedly improved the quality of life for many people. However, it has also created opportunities for predators to carry out harmful crimes. Internet crimes, in particular, have become a significant issue, with victims being targeted relentlessly and having no means of escape. Cyberbullying is one of the most severe internet crimes, and research has shown its detrimental consequences on victims, ranging from suicide to lowered self-esteem. As a result, cyberbullying prevention and control have become the focus of many psychological and technical studies.

This paper proposes a novel idea for identifying cyberbullying remarks in tweets using a precise method of Convolutional Neural Network (CNN) implementation with the Keras framework, achieving accurate results. The proposed system can be used by government bodies, organizations, parents, guardians, institutions, policy makers, and law enforcement agencies to help prevent individuals from becoming victims of cyberbullying.

Since the domain of virtual bullying is constantly evolving, the methodology requires constant updating and upgrading to match the current scenario. The proposed methodology can handle crisis situations and can be enhanced to provide full-time support, preventing potential crises. Some potential enhancements that can be incorporated in the future include:

Integration with natural language processing (NLP) techniques to improve accuracy and handle more complex language structures. Incorporating real-time monitoring of social media platforms to detect cyberbullying in its early stages. Integrating with existing tools that can provide counseling and support to victims of cyberbullying. Overall, the proposed methodology has the potential to significantly reduce the harmful effects of cyberbullying and provide much-needed support to victims.

6. REFERENCES

[1] “Text classification using convolution neural networks. (2017).
[2] 2.Keras academic deep-studying in python.
[3] B. Sri Nandhinia, J. (2015).“Online social community bullying detection the use of intelligence techniques”.
[4] K. Dinakar, R. R. and Lieberman, H. (2011). “Modelling the detection of textual cyberbullying”.
[5] K. Reynolds, A. K. and Edwards, L. (2011).“Using device studying to discover cyberbullying”.
[6] Mohammed Ali Al-garadi*, Kasturi Dewi Varathan, S. D. R. (2016).
[7] “Automatic detection of cyberbullying on social networks primarily based totally on bullying features”.
[8] V. Nahar, X. L. and Pang, C. (2013).“An powerful method for cyberbullying detection”.
[9] Whittaker, E.,. K. R. M. (2015).“Cyberbullying thru social media”.

[10] Archer (2018) B.Sri Nandhinia (2015) Mohammed Ali Al garadi* (2016)
[11] Rui Zhao (Rui Zhao) K.Dinakar and Lieberman (2011) K.Reynolds and Edwards (2011) Whittaker (2015) V.Nahar and Pang (2013) lin (2017) "Tweet class of soft evaluation the use of keras in python"
[12] Deep Learning for detecting cyberbullying throughout social media systems with the aid of using S Agarwal A Awekar.
[13] "Detecting kingdom of aggression in sentence" By R potapova
[14] Hate speech detection on Facebook (Blog)
[15] Analytics Vidya (Website for python and CNN)
[16] S. Salawu, Y. He, and J.Lumsden, "Approaches to Automated Detection of Cyberbullying : A Survey," vol. 3045, no. c, pp. 1–20, 2017.
[17] T. Wu, S. Liu, J. Zhang, and Y.Xiang, "Twitter unsolicited mail detection primarily based totally on deep learning," Proc. Australas. Comput. Sci.Week Multiconference - ACSW "17, pp. 1–8, 2017.
[18] MIT Technology Review, "These are the 10 breakthrough technologies you need to know about right now," 2018. [Online]. Available: https://www.technologyreview.com/lists/technologies/2017/. [Accessed: 02-Mar-2018].