# Cyberbullying Detection in Instagram Comments

[1].C. Valarmathi, Assistant Professor, Department of Computer Science and Engineering, Sri Sairam College of Engineering, Anekal, Bangalore,vinmathi20@gmail.com.

[2].S. Sharanya, Student, Department of Computer Science and Engineering, Sri Sairam College of Engineering, Anekal, Bangalore, sridharan784@gmail.com.

## Abstract:

Bullying or harassment that takes place online is known as cyberbullying or cyber harassment. Online bullying also refers to cyberbullying and cyber harassment. Repeated behaviour and an intention to cause harm are indicators of bullying or harassment. Cyberbullying victims may have reduced self- esteem, more suicide thoughts, and other unfavourable emotions including anxiety, frustration, anger, or depression. Our project is about detecting these hateful comments using a machine learning algorithm and classifying them into 2 levels based on their severity like “low” and “High” levels. Social media has become a very important part of our lives, especially after the pandemic. It has been used by people in all age groups as a way of communication. Even though it is making our life so much easier, it also has its downsides. In this project, we will address one of such issues, it is evident that social media is to share opinions but sometimes it gets out of hand. Hateful comments can affect people’s physical and mental health. This is something we have come across in our day-to-day life, and the solution we have come up with is to detect such comments and take action against them. We are using a Machine Learning algorithm to detect hateful comments and their severity to take action depending on it.

**Keywords: cyber bullying, Machine Learning, Random Forest algorithm**.

## I.INTRODUCTION

Bullying of children and teenagers is primarily conducted on social media. In their daily lives, people pay close attention to everything. Social networking is being used by many people to boost their careers. Prefer putting their skills to use and sharing those things on various social media networks. Such as social networking sites like Instagram. Simply leaving abusive remarks on someone else's posts qualifies as cyberbullying. their mental health is so disturbed. Bullying is affecting a lot of young people. Cyberbullying is on the rise as social networking usage grows. Our objective here is to identify and suppress such hostile remarks. In order to identify such remarks based on severity, we employ certain machine learning models here. The severity is therefore divided into three categories: High, Low, and Medium. Those remarks ought to be classified as hateful or not depending on their seriousness. Numerous media bullying strategies have been used, although many of them were text-based. The purpose of this paper is to demonstrate the software solution that will be used to identify hostile remarks made by bullies. Creating a false identity and sharing an embarrassing image or video are just two examples of cyberbullying; it also involves spreading unfavorable rumors and making threats. The victim's conduct is altered by the poster of such cruel comments. This has an impact on their emotions, as well as their self- confidence and sensation of terror. A comprehensive solution is therefore needed for this situation. Cyberbullying must be stopped. Using machine learning models, the issue can be resolved by identifying and preventing it.

## II.BACKGROUND OF THE WORK

[1] Varun Jain and et.al, "Cyberbullying Detection on Social Networks Using Machine Learning Approaches", 2021.Social media use has grown rapidly along with the Internet's expansion, becoming the primary networking tool of the twenty-first century. Increased social networking, however, frequently has negative repercussions on society that encourage a range of unwanted phenomena such online abuse, harassment, cyberbullying, criminality, and trolling. Cyberbullying frequently results in serious emotional suffering and bodily harm, especially for women and children. On occasion, it even drives victims to attempt suicide. [2] Rahul Ramesh Dalvi and et al "Detecting A Twitter Cyberbullying Using Machine Learning", 2020.Bullying affects a lot of young people on social media. Cyberbullying is as prevalent as social networking sites, and it is getting worse daily. The abusive tweets' phrase patterns can be discovered using machine learning, which can also be utilized to create a model that can automatically identify abusive behavior on social media. There have been numerous techniques for detecting social media bullying, but many of them mainly rely on text. This article's main goal is to show how the software may be used to find offensive tweets, posts, etc. It is suggested that bullying on Twitter be identified and stopped using machine learning. The two classifiers used to train and test the social media bullying content are SVM and Naive Bayes. SVM and Naive Bayes were both.[3] Saloni Mahesh Kargutkar, Prof. Vidya Chitre, "A Study of Cyberbullying Detection Using Machine Learning Techniques,2020.These days, people make extensive use of social media platforms to use their expertise to further their careers. It will be taking place in a daily manner for folks. Many people use social media for career purposes, but some of them demotivate them by posting harsh comments. People's mental health, among other things, are affected by those cruel remarks. It was referred to as online harassment. Bullying on social media platforms was suggested as a solution here. Therefore, they employed certain strategies in this instance to lessen harassment through online networking. Perceptron's are utilized with CNNs (convolutional neural networks).

## III. PROPOSED SYSTEM

Each system component is described in the system architecture. The following list of modules makes up this project: 1.Dataset: The dataset includes over 40 thousand comments that have been categorised as hateful. 2.Pre-processing: Before the input is fed to the algorithm, several modifications are made to it. Data rescaling, binary, and standardisation are all included in this pre-processing. 3.Feature extraction: We are using TF-IDF(Term Frequency-Inverse Document Frequency) algorithm for feature extraction from natural language comments. TF- IDF algorithm is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus. This converts each word in the comments into a vector of integers which is forwarded into the next module. 4.Classification: RFC [Random Forest Classifier] The classification algorithm Random Forest Classifier is employed. A random forest is a machine learning method for tackling classification and regression issues. It makes use of ensemble learning, a method for solving complicated issues by combining a number of classifiers. Based on the predictions of the decision trees, the (random forest classifier) algorithm determines the result. It makes predictions by averaging or averaging out the results from different trees. The accuracy of the result grows as the number of trees increases.

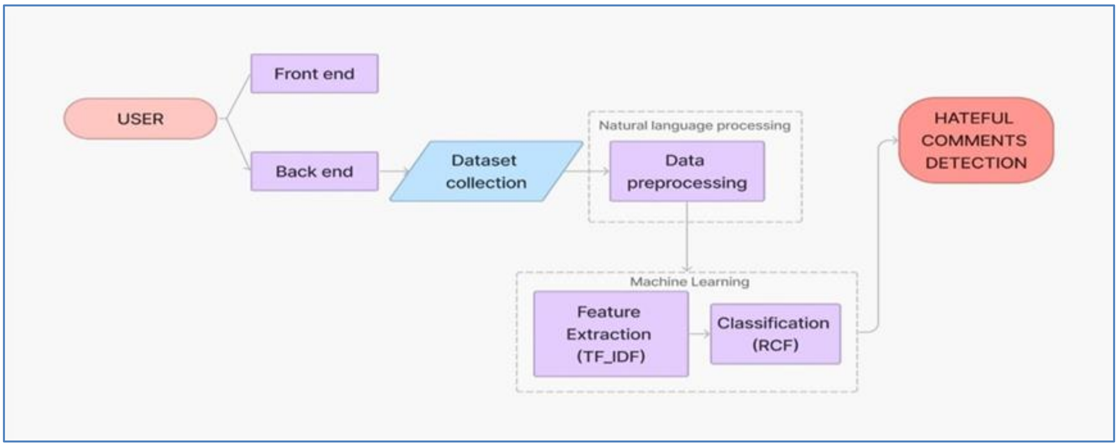


Fig.3.1. System architecture of Cyber bullying proposed model

The DFD also provides information about the inputs and outputs of each entity as well as the process itself. Loops, decision-making processes, and control flows are absent from a data-flow diagram. Certain operations based on the data can be represented using a flowchart. System architecture, and many designs that are involved in the suggested system.

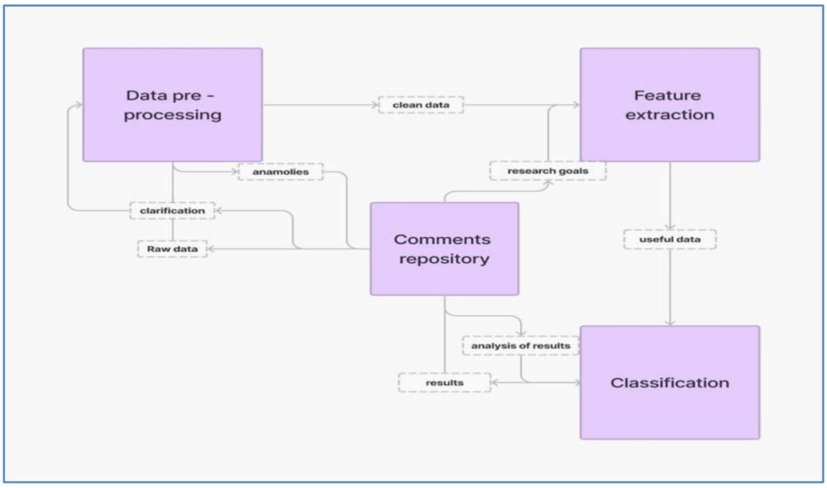


Fig.3.2. Data Flow Diagram

3.1. Feature Extraction

In the TF-IDF algorithm, the Term Frequency (TF) registers the frequency of a specific word relative to the comments. Inverse Document Frequency (IDF) looks at how common a word is amongst the corpus. For example a post Instagram like “i hate you” is processed as follows,

Let *w* is a term or word in the cyberbullying instance and *j* is a specific cyberbullying instance (e.g., a comment, message, or post)

$TF\_IDF(w,j) \ = count(w,j) * \ log \left(\frac{N}{df(w)}\right)$ ------------------------------ (Eqn.1)

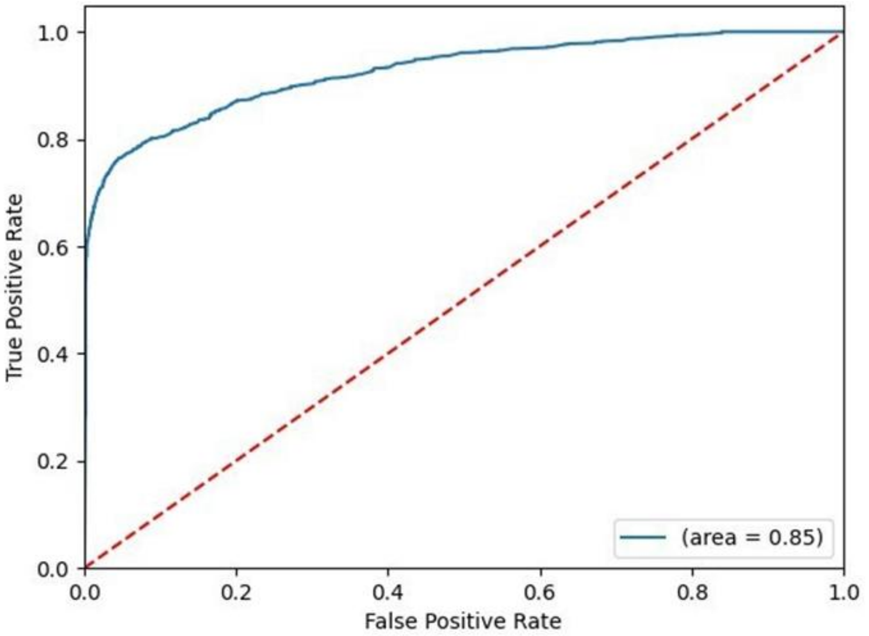
$N$ => the total number of cyberbullying instances in the dataset.

3.2. Random Forest Classifier

The classification algorithm Random Forest is employed. The strategy includes employing a classification procedure for cyberbullying recognizable proof on social media, applying it to 50 bunches of Instagram posts comments, utilizing directed learning with choice trees, building choice trees based on the Gini measure, and assessing expectation comes about against past information labelling. The accuracy and number of trees for classification are straightforwardly proportional to each other. Consequently, the accuracy of the result may be improved when the number of trees increases.

## IV. RESULT ANALYSIS

The performance of a well-liked machine learning method called Random Forest Classifier is evaluated with the input file. It is an approach for ensemble learning that creates numerous decision trees and combines them to increase accuracy and decrease over fitting. The below indicators are used to assess the performance of a Random Forest Classifier model. Here are a few common measurements to analyse the execution with confusion matrix along with the output values. It is often necessary to maintain balanced performance metric to achieve desired page in cyberbullying detection task. The values which are achieved in random forest classifier used in this model are calculated as per the formula.



These metrics are used to assess a Random Forest Classifier model's performance and comparing it against other models or iterations of the same model. It's critical to select the statistic that best fits the challenge at hand and the project's particular needs. The overall distribution of hateful and non- hateful comments is shown in Fig4.1.

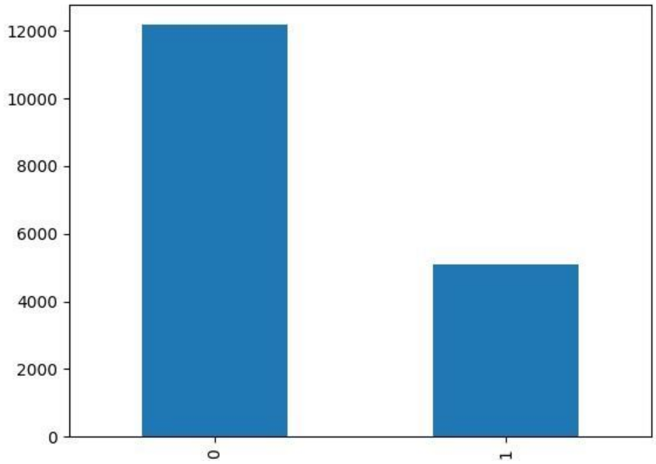


Fig 4.1 Distribution of hateful comments in the dataset

## V.CONCLUSION:

It’s not that all the users wish to see only the positive side of the comments on the posts. Also the entire comments needs to be analyzed to determine if the comment is a positive or a negative one. The analysis of the comments will be done using Machine Learning Algorithm(tf-idf) and provides a pop up if the comments section contains any hateful comments so the user need not to analyze the comments. In few cases the person who adds a harmful/hateful comment may not be aware that his comment would affect the other person (i.e, Few comments may also be unintentional) so adding a tag next to an hateful comment will be useful for the users to know that their comment is harmful and also reduces the amount of cyber bullying which will be the main scope of this project.

## REFERNCES:

[1] Varun Jain, Vishant Kumar, Vivek Pal, “Cyberbullying Detection on Social Networks Using Machine Learning Approaches”, 2021.

[2] Rahul Ramesh Dalvi, Sudhanshu Baliram Chavan, Aparna Halbe, “Detecting A Twitter Cyberbullying Using Machine Learning”, 2020

[3] Saloni Mahesh Kargutkar, Prof. Vidya Chitre, “A Study of Cyberbullying Detection Using Machine Learning Techniques,2020

[4]. Dalvi, R. R., Baliram Chavan, S., & Halbe, A. (2020, May). Detecting A twitter cyberbullying using machine learning. 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS). Presented at the 2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India. doi:10.1109/iciccs48265.2020.9120893

[5]. Karthikeyan. (2022). The escalating cyberbullying menace through social media platforms in developing countries including India. In Handbook of Research on Digital Violence and Discrimination Studies (pp. 526–546). doi:10.4018/978-1-7998-9187-1.ch023

[6]. Saloni Mahesh Kargutkar, Prof. Vidya Chitre, “A Study of Cyberbullying Detection Using Machine Learning Techniques,2020.

[7]. Kazi Saeed Ala, Shovan Bhowmik, Priyo Ranjan Kundu Prosun,“Cyberbullying Detection: An Ensemble Based Machine Learning Approach ”, 2021.

[8] Novalita, N., Herdiani, A., Lukmana, I., & Puspandari, D. (2019). Cyberbullying identification on twitter using random forest classifier. Journal of Physics. Conference Series, 1192, 012029. doi:10.1088/1742-6596/1192/1/012029.

[9] Altay, E. V., & Alatas, B. (2018, December). Detection of cyberbullying in social networks using machine learning methods. 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT). Presented at the 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT), Ankara, Turkey. doi:10.1109/ibigdelft.2018.8625321.

[10] Thangarasu, G., & Alla, K. R. (2023, May 20). Detection of cyberbullying tweets in twitter media using random forest classification. 2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE). Presented at the 2023 IEEE 13th Symposium on Computer Applications & Industrial Electronics (ISCAIE), Penang, Malaysia. doi:10.1109/iscaie57739.2023.10165118.

[11] Valarmathi. C and S. Sharanya, “Scam Call Detection Using NLP and Naïve Bayes Classifier,” INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT, vol. 08, no. 07. Indospace Publications, pp. 1–6, Jul. 22, 2024. doi: 10.55041/ijsrem36688.