

# **AUTOMATED CYBERBULLYING DETECTION SYSTEM**

## **Capstone Project Proposal**

**Submitted by:**

**(101903365)      Bhavya Kakwani**  
**(102083036)      Ananya Agarwal**  
**(102083061)      Moti Rattan Gupta**  
**(102083037)      Prabhnoor Singh**

**BE Third Year- COE**  
**CPG No. 179**

**Under the Mentorship of**  
**Dr. Jayendra Barua**  
**Assistant Professor, CSED**




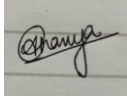

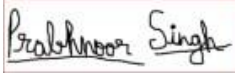
**Computer Science and Engineering Department**  
**Thapar Institute of Engineering and Technology, Patiala**  
**March 2022**

## TABLE OF CONTENTS

S. No.	Topic	Page No.
1	Mentor Consent Form	1
2	Project Overview	2
3	Problem Statement	3
4	Need Analysis	4
5	Literature Survey	5
6	Objectives	7
7	Methodology	8
8	Work Plan	9
9	Project Outcomes & Individual Roles	10
10	Course Subjects	11
11	References	12

## 1. Mentor Consent Form

I hereby agree to be the mentor of the following Capstone Project Team:

Project Title: AUTOMATED CYBERBULLYING DETECTION SYSTEM		
Roll No	Name	Signatures
101903365	Bhavya Kakwani	
102083036	Ananya Agarwal	
102083061	Moti Rattan Gupta	
102083037	Prabhnoor Singh	

NAME of Mentor: **Dr. Jayendra Barua**

SIGNATURE of Mentor:

## **2. Project Overview**

Cyberbullying or Cyber-harassment is a form of bullying or harassment done by using electronic means. This is also known as online bullying. It has become increasingly common among teenagers, because the digital sphere has expanded and technology has advanced. It is when someone bullies or harasses others on the internet, particularly on social media. Harmful bullying behavior can include posting rumors, threats, sexual remarks, personal information of victim, or hate speeches. Bullying or harassment can be identified by repeated behavior and an intent to harm someone. Victims of cyberbullying may experience lower self-esteem, increased suicidal ideation, and various negative emotional responses, including angry, or depressed.

According to a study[10], children and young people under 25 who are victims of cyberbullying tend to cause self-harm and enact suicidal behavior twice more likely. Also, in another study[11] over 14% admitted to cyberbullying another person, with spreading rumors online, where it was found that text, or email are the most common form of bullying.

We are going to develop a machine learning model, which will classify any text into 6 categories which are as follows: age-based cyberbullying, ethnicity-based cyberbullying, gender-based cyberbullying, religion-based cyberbullying, any other form of cyberbullying, and not cyberbullying. Further, we are going to develop chatbots for various social media platforms like discord to try to detect cyberbullying using the above machine learning model, and take appropriate measures.

### 3. Problem Statement:

To develop an automated Cyberbullying detection system which will reliably detect any cyberbullying activity happening on the social media platforms over textual conversations and will take appropriate actions when it detects such activity.

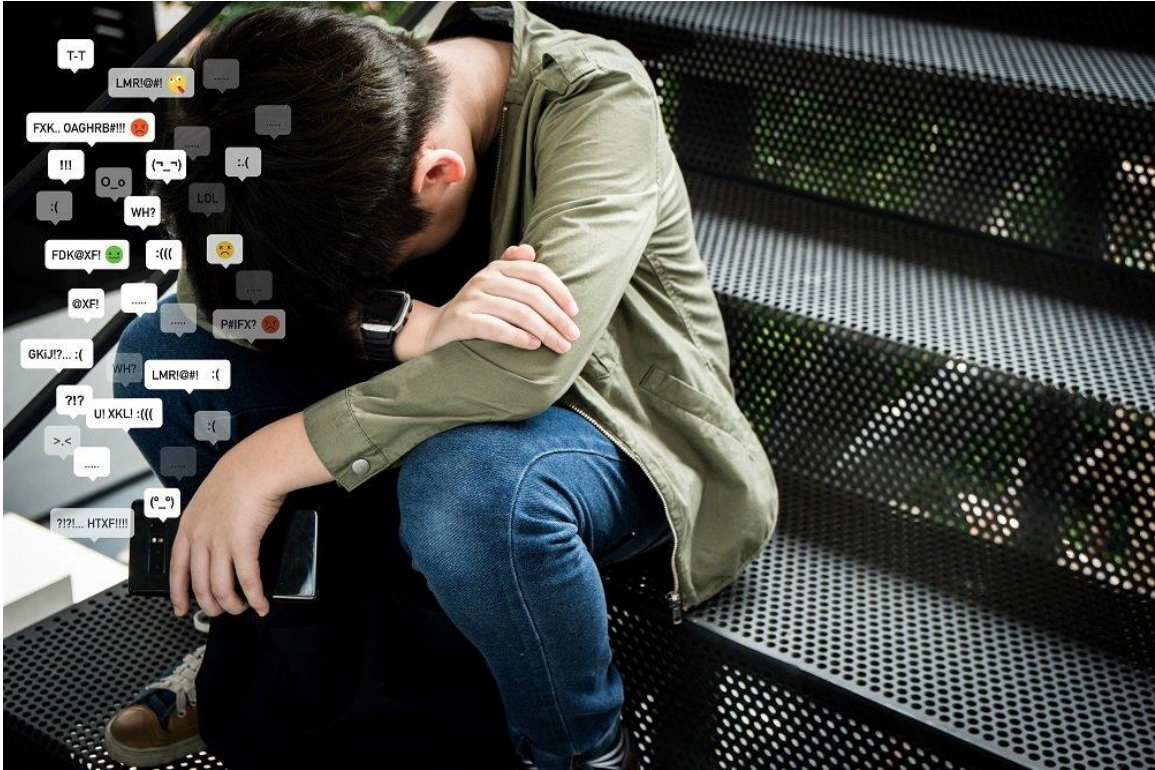


Figure 1: Cyberbullying impact on a kid[12]

## 4. Need Analysis

Recent research[11] has shown that large numbers of people are victims of cyberbullying, leading to a growing perception that cyberbullying is becoming a major problem. Some of the statistics proving the same are as follows:

- A study of more than 6,000 children aged 10-18 from June to August last year found that about 50% of children had experienced at least one form of cyberbullying in their lifetime, according to a report published in February by the Joint of the European Commission. Research Centre (JRC).
- Of the 11 European countries included in the report, 44% of children who had been abused online before the Lockdown closure said it happened the most during the closure (February 2021).
- Children and young people under the age of 25 who are victims of cyberbullying are twice as likely to injure themselves and commit suicide, according to the study. The perpetrators also have a high risk of suicidal thoughts and behavior.
- A 2016 report from the Cyberbullying Research Center indicates that 33.8% of students of age in-between 12 and 17 were victims of cyberbullying in their lifetime. Also, 11.5% of students of age in-between 12 and 17 indicated that they had engaged in cyberbullying in their lifetime.

Therefore, there is a dire need for a solution to combat this evil.

## 5. Literature Survey

The term cyberbullying[1] is widely attributed to Bill Belsey, although its earliest usage dates back to a New York Times article in 1995[1]. In the research done by Cynthia Van Hee et al.[2], They focused on cyberbullying as a particular form of cybervictimization and explore its automatic detection and fine-grained classification. Data containing cyberbullying was collected from the social networking site Ask.fm. They developed and applied a new scheme for cyberbullying annotation, which describes the presence and severity of cyberbullying, a post author's role (harasser, victim or bystander) and a number of fine-grained categories related to cyberbullying, such as insults and threats. They presented experimental results on the automatic detection of cyberbullying and explore the feasibility of detecting the more fine-grained cyberbullying categories in online posts. For the first task, an F-score of 55.39% is obtained. They observed that the detection of the fine-grained categories (e.g., threats) is more challenging, presumably due to data sparsity, and because they are often expressed in a subtle and implicit way.

In another research Vivek K. Singh et al.[3], the authors proposed a novel probabilistic information fusion framework that utilizes confidence score and interdependencies associated with different social and textual features and uses those to build better predictors for cyberbullying. The performance of the proposed approach was compared to a recent approach in literature which used a similar dataset and features and the proposed approach resulted in significant improvements in terms of cyberbullying detection. There exist some limitations like does not employ a representative user sample, the balanced dataset used in this work may not mimic the real-world.

In yet another research by Walisa Romsaiyud et al.[4], the Naïve Bayes classifier is used for learning word extraction and clustering of loaded patterns. The algorithm includes two main methods: (1) using k mean clustering to iteratively move from the full data set to clusters to create partitions as a polynomial feature vector of the model (2) to predict 8 classes by capturing a random of a particular section as frequency of words and plotting the probability of words in the document. The proposed method (3) improved the accuracy and reliability of the experiment.

Another research by Pei-Ju Lee et al.[5], they used three kinds of classifier which are k-Nearest Neighbors (KNN), Support Vector Machine (SVM), and C4.5 Decision Trees (DT) and uses the automatically adjusted default values for analyzation. The data mining software Weka 3.8 is used to construct the models of KNN, SVM, and DT of cyberbullying prediction. The information gain and chi-square test are used for feature selection from textual features and user features and set up a new feature set for the models. The 10-fold cross validation is used for evaluations of the three models.

Some other research by Semiu Salawu et al.[6], They identified that the classification of messages as bullying or bully-victim identification are the most common tasks performed by researchers today. This task can be considered well-researched and researchers now need to research on more advanced tasks, such as detecting cyberbullying via social

exclusion, proving power differential and repeatability criteria, during a cyberbullying episode. They also discovered that swear words and abusive text is often equated to bullying. This is not always the case; hence more effort should be directed at detecting cyberbullying in text devoid of profanity and insults.

In another research by Jason Wang et al.[7], they have a framework for the automatic generation of balanced data by using a semi-supervised online Dynamic Query Expansion (DQE) process to extract more natural data points of a specific class from Twitter. We also propose a Graph Convolutional Network (GCN) classifier, using a graph constructed from the threshold cosine similarities between tweet embeddings.

In yet another research by Nabi Rezvani et al.[8], they presented an intelligent Cyberbullying detection pipeline to: (i) extract features (from an image, image meta-data and textual content generate); (ii) contextualize the extracted features by developing a crowdsourced feedback loop and drawing on human knowledge; and (iii) combining features using a Neural Network to identify and build potentially useful features. They adopted a typical scenario for analyzing social media content generated on Instagram and Twitter to identify Cyberbullying activities hence significantly improving the quality of extracted knowledge.

Another research by Seunghyun Kim et al.[9], presents finding of the corpus of papers and helped establishing the basic terms of cyberbullying widely used. Organized with Baumer's three-dimensional human-centered algorithm design framework, each of the following subsections interprets the cyberbullying detection algorithms in the papers using the coding rubric. They have analyzed the human involvement in the development of these models using an established human-centered algorithmic design framework and revealed that despite extensive research on developing cyberbullying detection models that optimize for statistical performance and methodological innovation, there were clear gaps.



## 6. Objectives

- To study existing systems and techniques for cyberbullying detection and prevention.
- To make a ML model which will reliably detect any cyberbullying activity over textual conversations on social media platforms using NLP.
- To evaluate ML model on the various model evaluation parameters like accuracy, precision, recall, F1-score, etc.
- To develop a discord bot (<http://discord.com>) which uses our ML model to detect Cyberbullying and take appropriate actions when it detects such activity.

## 7. Methodology

As a common methodology for both the ML model creation and Discord bot, the following steps are to be executed:

- Prerequisite Learning:
  - Learning about existing systems and techniques for cyberbullying detection and prevention.
  - Learning Natural Language Processing (NLP).
  - Learning about Cyberbullying.
- Creation of Machine Learning model:
  - Gathering various datasets from the internet.
  - Applying data preprocessing like stop-word removal, Lemmatization, Tokenization, usernames removal, Punctuation removal including emojis on the chosen dataset.
  - Trying various models used to attain the best accuracy including SVM (Linear-Kernel), Naive Bayes, Logistic Regression, Random Forest, xgboost, and LSTM.
- Calculating various model evaluation parameters like accuracy, precision, recall, F1-score, etc. for our ML model.
- Trying to find the:
  - Best ML model on the selected dataset.
  - Best social media platform with respect to our ML model.
- Creation of the discord bot:
  - Using the discord.py API for creating our discord bot.
  - Using our ML model in the bot for detecting cyberbullying on the discord platform and taking appropriate actions when it detects such activity.

## 8. Work Plan

S.No.	Activity	Month	February				March				April				May				August				September				October				November				December			
		Week No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36
1	Identification, Formulation and Planning of Project	Plan																																				
		Actual																																				
2	Studying NLP, ML, & Cyberbullying	Plan																																				
		Actual																																				
3	Data Collection & Preprocessing	Plan																																				
		Actual																																				
4	Creating ML models	Plan																																				
		Actual																																				
5	Evaluation of ML models on conversations from various social media platforms	Plan																																				
		Actual																																				
6	Developing Discord Bot	Plan																																				
		Actual																																				
7	Testing	Plan																																				
		Actual																																				
8	Final Report	Plan																																				
		Actual																																				

Figure 2: Project Gantt Chart

## 9. Project Outcomes & Individual Roles

The final execution of the project will have the following outcomes:

- A ML model which will reliably detect any cyberbullying activity over textual conversations on social media platforms using NLP.
- A fully functional discord bot having administrative capabilities which uses the above ML model to detect Cyberbullying and take appropriate actions when it detects such activity.

The individual roles for our project are as follows:

S. No.	Member Name	Role
1	Bhavya Kakwani	Programming, Data Collection & Studying
2	Ananya Agarwal	Programming, Documentation
3	Moti Rattan Gupta	Programming, Data Collection & Studying
4	Prabhnoor Singh	Programming, Documentation

## **10. Course Subjects**

Some of the course subjects used in our project include:

- Artificial Intelligence (AI)
- Natural Language Processing (NLP)
- Machine Learning (ML)
- Software Engineering (SE)
- Database Management System (DBMS)
- Object Oriented Programming (OOP)
- Data Structures (DS)

## 11. References

- [1] Cyberbullying - an overview | ScienceDirect Topics, accessed 16 March 2022, <https://www.sciencedirect.com/topics/social-sciences/cyberbullying>
- [2] Van Hee, Cynthia, et al. "Detection and fine-grained classification of cyberbullying events." *Proceedings of the international conference recent advances in natural language processing*. 2015.
- [3] Singh, Vivek K., Qianjia Huang, and Pradeep K. Atrey. "Cyberbullying detection using probabilistic socio-textual information fusion." *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2016.
- [4] Romsaiyud, Walisa, et al. "Automated cyberbullying detection using clustering appearance patterns." *2017 9th International Conference on Knowledge and smart Technology (KST)*. IEEE, 2017.
- [5] Lee, Pei-Ju, et al. "Cyberbullying detection on social network services." (2018).
- [6] Salawu, Semi, Yulan He, and Joanna Lumsden. "Approaches to automated detection of cyberbullying: A survey." *IEEE Transactions on Affective Computing* 11.1 (2017): 3-24.
- [7] Wang, Jason, Kaiqun Fu, and Chang-Tien Lu. "Sosnet: A graph convolutional network approach to fine-grained cyberbullying detection." *2020 IEEE International Conference on Big Data (Big Data)*. IEEE, 2020.
- [8] Rezvani, Nabi, Amin Beheshti, and Alireza Tabebordbar. "Linking textual and contextual features for intelligent cyberbullying detection in social media." *Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia*. 2020.
- [9] Kim, Seunghyun, et al. "A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms." *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2 (2021): 1-34.
- [10] Young victims of cyberbullying twice as likely to attempt suicide and self-harm, study finds, accessed 16 March 2022, <https://www.sciencedaily.com/releases/2018/04/180419130923.htm>
- [11] Enough is Enough: Cyberbullying, accessed 16 March 2022, [https://enough.org/stats\\_cyberbullying](https://enough.org/stats_cyberbullying)
- [12] Cyberbullying.jpg, accessed on 16 March 2022, <https://www.allbusiness.com/asset/2021/09/Cyberbullying.jpg>