BERT VS CYBERBULLYING: CREATING A SAFER DIGITAL SPACE

A PROJECT REPORT

Submitted by,

DARSHAN MK	20221LCA0005
KUSUMA KN	20211CAI0138
RACHITA S	20211CAI0137
LIPIKA DEVAIAH	20211CAI0132
ANAIZA KHAN	20211CAI0003

Under the guidance of,

Mr. J. JOHN BENNET

in partial fulfillment for the award of the degree of

BACHELOR OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING, COMPUTER ENGINEERING, INFORMATION SCIENCE AND ENGINEERING Etc.

At

PRESIDENCY UNIVERSITY

BENGALURU

MAY 2025

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

CERTIFICATE

This is to certify that the Project report **"BERT VS CYBERBULLYING: CREATING A SAFER DIGITAL SPACE"** being submitted by "DARSHAN MK, KUSUMA KN, RACHITA S, LIPIKA DEVAIAH, ANAIZA KHAN" bearing roll number(s) " 20221LCA0005, 20211CAI0138, 20211CAI0137, 20211CAI0132, 20211CAI0003" in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Computer Science and Engineering is a Bonafede work carried out under my supervision.

Mr. J. JOHN BENNET
Assistant Professor
School of CSE&IS
Presidency University

Dr. ZAFAR ALI KHAN
HOD/CAI
School of CSE&IS
Presidency University

Dr. MAHALAKSHMI
Associate Dean School of CSE
Presidency University

Dr. MYDHILI NAIR
Associate Dean School of CSE
Presidency University

Dr. SAMEERUDDIN KHAN
Pro-Vc School of Engineering
Dean -School of CSE&IS
Presidency University

PRESIDENCY UNIVERSITY

SCHOOL OF COMPUTER SCIENCE ENGINEERING

DECLARATION

We hereby declare that the work, which is being presented in the project report entitled **BERT VS CYBERBULLYING: CREATING A SAFER DIGITAL SPACE** in partial fulfilment for the award of Degree of **Bachelor of Technology** in **Computer Science and Engineering**, is a record of our own investigations carried under the guidance of **Mr. J. JOHN BENNET, Assistant professor, School of Computer Science Engineering and Information Science, Presidency University, Bengaluru.**

We have not submitted the matter presented in this report for the award of any other Degree.

NAMES	ROLL NUMBER	SIGNATURE
DARSHAN MK	20221LCA0005	
KUSUMA KN	20211CAI0138	
RACHITA S	20211CAI0137	
LIPIKA DEVAIAH	20211CAI0132	
ANAIZA KHAN	20211CAI0003	

ABSTRACT

The threat towards mental health on recent users persists these days due to a rapid proliferation of cyberbullying incidents. This type of violence has a particular effect. This dangerous diffusion puts threatening consequences in all aspects of living. A well-established program called **"BERT VS CYBERBULLY"** is leading the sellers in finding recommendations that have an active way of acquiring solutions for automated and instant scalable systems that comprise incidents of cyber bullying. The latest technologies are used as this system works based on how it can read and understand the text in terms of semantics that can help one identify words that are harmful or which really hide in the more subtle forms like sexual harassment or bullying within virtual barriers. Data sets, as well as adaptive learning techniques that ensure accuracy in pinpointing complex abuse patterns, are employed.

One of the features of this solution is that it focuses on intervention and support. In addition to real-time surveillance and alert notification, the system offers automated reporting mechanisms for users to confidentially flag illicit content. This approach is unique from the majority that involve the detection of the misuse and support of victims, as it tends to be more reactive for other solutions.

The application is aimed at countering the direct and indirect approaches pertaining to cyber-bullying that include offensive language, intent to harm, or even subtle ones, like sarcasm and passive-aggressive statements through sentiment analysis, emotion detection, and contextual understanding, thanks to deep learning. Integration with popular media platforms makes the product highly adaptable across a range of different online ecosystems. This gives the software the capacity to handle a high volume of user-generated content in real-time, immediately responds as soon as there is evidence of abuse.

The main aim of this initiative is to create a safer, inclusive and supportive digital environment. It is expected to provide a broader aspiration for digital well-being and ethical AI by creating awareness, establishing accountability and intervention. This will prevent cyberbullying, encompassing technological solutions for detection. It is a proactive measure that seeks to curtail and mitigate long-term psychological-social impacts.

ACKNOWLEDGEMENT

First of all, we indebted to the **GOD ALMIGHTY** for giving me an opportunity to excel in our efforts to complete this project on time.

We express our sincere thanks to our respected dean **Dr. Md. Sameeruddin Khan**, Pro-VC, School of Engineering and Dean, School of Computer Science Engineering & Information Science, Presidency University for getting us permission to undergo the project.

We express our heartfelt gratitude to our beloved Associate Deans **Dr. Mahalakshmi and Dr. Mydhili Nair,** School of Computer Science Engineering & Information Science, Presidency University, and **Dr. ZAFAR ALI KHAN**, Head of the Department, School of Computer Science Engineering & Information Science, Presidency University, for rendering timely help in completing this project successfully.

We are greatly indebted to our guide **Mr. J. JOHN BENNET, Assistant Professor** and Reviewer **Mr. SANDEEP ALBERT MATHIAS**, **Assistant Professor** School of Computer Science Engineering & Information Science, Presidency University for her inspirational guidance, and valuable suggestions and for providing us a chance to express our technical capabilities in every respect for the completion of the project work.

We would like to convey our gratitude and heartfelt thanks to the PIP2001 Capstone Project Coordinators **Dr. Sampath A K, Dr. Abdul Khadar A and Mr. Md Zia Ur Rahman,** department Project Coordinators **Mr. Afroz Pasha** and Git hub coordinator **Mr. Muthuraj.**

We thank our family and friends for the strong support and inspiration they have provided us in bringing out this project.

Darshan MK

Kusuma KN

Rachita S

Lipika Devaiah

Anaiza Khan

LIST OF TABLES

LIST OF FIGURES

CHAPTER-1
INTRODUCTION

General Overview

Cyberbullying has now become the new trend of today's digital world, most common amongst younger users. Using social media, gaming applications, and messaging tools has ensured that a person engages with somebody online all through. With such applications available for all to communicate, they also paved their way to carry on malpractices such as harassing, ridiculing, and exclusion. The anonymous nature of the internet allows cyberbullies to target victims without fear of immediate consequences. As a result, the emotional and psychological toll on victims can be severe, leading to long-term effects such as anxiety, depression, and in extreme cases, self-harm. The urgency of the issue notwithstanding, previous solutions have not been effective enough to address both the obvious and subtle forms of cyberbullying across diverse platforms.

Problem Statement

Cyberbullying is a growing digital threat, often undetected by traditional systems that rely on keyword matching and sentiment analysis. These methods fail to capture subtle harassment like sarcasm and passive-aggressive language, and they lack real-time intervention and victim support. This project, "BERT vs Cyberbullying," leverages BERT and deep learning to detect harmful interactions in real-time, providing automated alerts, user education, and support tools. Its scalable, multi-platform integration ensures broad applicability across social media and messaging apps. By enhancing detection accuracy and offering proactive intervention, this system aims to create a safer, more supportive digital environment.

Project Introduction

In recent years, there has been an alarming increase in cyberbullying incidents. Bullying behaviors have become hotspots in social media platforms, where younger users are especially active.

While some forms of cyberbullying are overt and easy to identify, such as name-calling and threats, many others are subtle and harder to detect, such as exclusionary tactics, indirect insults, and manipulative behavior. This calls for more complex systems that can identify those behaviors and intervene in actual time. The increasing rate of cyberbullying has raised awareness and concern across the globe of the need to create safer virtual environments for all users of digital environments.

Project Objectives

The motivation behind the BERT vs Cyberbullying project stems from the critical need to develop an intelligent system that can provide real-time protection and support for users, especially in online spaces where bullying often goes unnoticed. Several content filtering platforms have been used in order to detect explicit language. However, these are limited in that they are unable to cope with the nuances and constantly evolving nature of cyberbully.

Project Domain

This project is under the domain - Machine Learning and Cyber security.

CHAPTER-2
LITERATURE SURVEY

Title: Detection of Cyber Security Threats through Social Media Platforms

Author(s): Antonios Karteris, Georgios Tzanos, Lazaros Papadopoulos, Dimitrios Soudris

Year Published: 2003

Methods Used: text-mining techniques to analyze Twitter posts for real-time detection of cyber security threats, filtering system with predefined cybersecurity-related keywords and a secondary filtering layer for refining results.

Advantages:

- Provides real-time identification of security threats.
- Customizable, open-source, and user-friendly.
- Integrated with a decision-support system for alerting security personnel.
- Lightweight in terms of computational resources.

Disadvantages:

- The detection accuracy is limited to 73%.
- Effectiveness depends on predefined keywords, which may miss emerging threats.

Title: Cyber Threat Detection Using Machine Learning Techniques: A Performance Evaluation Perspective.

Authors: Kamran Shaukat, Suhuai Luo, Shan Chen, Dongxi Liu.

Year Published: 2020

Methods Used: Three machine learning models Deep Belief Networks (DBN), Decision Trees, and Support Vector Machines (SVM). spam detection, intrusion detection, and malware detection.

Advantages:

- Comprehensive analysis of multiple machine learning techniques.
- It identifies strengths and weaknesses of different models.

Disadvantages:

- Machine learning models remain vulnerable to adversarial attacks.

Title: Real-Time Cybersecurity Threat Detection Using Machine Learning a Big Data Analytics Comprehensive approach

Author(s): Kingsley David Onyewuchi Ofoegbu, Olajide Soji Osundare, Chidiebere Somadina Ike, Ololade Gilbert Fakeyede, Adebimpe Bolatito Ige

Year Published: 2023

Method used: anomaly detection techniques and pattern recognition across large datasets

Advantages:

- Highly scalable approach for handling large volumes of data.
- Can detect zero-day attacks through anomaly detection.
- Reduces false positives compared to traditional methods.

Disadvantages:

- Requires continuous model training for adaptation to new threats.
- High computational power needed for real-time processing.

Title: Vulnerability Detection Using BERT-Based LLM Model with Transparency Obligation Practice Towards Trustworthy AI

Authors: Jean Haurogné, Nihala Basheer, Shareeful Islam

Year Published: 2024

Methods Used: BERT model, XAI techniques for SHAP, LIME, and heatmaps

Advantages:

- High accuracy (91.8%) in detecting vulnerabilities in source code.
- Provides explainability for AI-driven vulnerability detection.
- Aligns with the EU AI Act for trustworthy AI practices.

Disadvantages:

- Performance dependent on pre-trained LLM models and dataset quality.

Title: Real-Time Detection and Tracking Using Multiple AI Models and Techniques

Authors: Sangeeta Sanghal

Year Published: 2024

Methods Used: AI models, including machine learning, deep learning, and anomaly detection, to enhance real-time cybersecurity threat detection and tracking.

Advantages:

- Combines multiple AI models for improved detection accuracy.
- Provides real-time threat identification and tracking.

Disadvantages:

- High computational power required for real-time analysis.
- Ethical concerns regarding AI-driven decision-making in cybersecurity.

CHAPTER-3
RESEARCH GAPS OF EXISTING METHODS

Current measures for detecting cyber threats are limited

Traditional methods for the detection of cyberbullying have mainly been focused on offensive language and explicit threats. These methods tend to overlook the subtler and nuanced forms of harassment that can be at least as damaging, if not more so, to the victim. This calls for more sophisticated systems that can understand context, indirectly identify bullying tactics, and provide real-time intervention**.**

Lack of Contextual Understanding in Existing Systems

The gap in current cyberbullying detection methods is that they do not understand the context in which offensive language is used. The traditional systems rely on keyword matching or sentiment analysis, which may miss the subtlety of language, especially when it comes to sarcasm, irony, or coded language. For example, a phrase like "Oh, great job, you're so awesome" might appear positive, but also can be used in a sarcastic tone. Current systems often fail to detect such nuances, which leads to false positives**.**

Sentiment Analysis Limitations

sentiment analysis models struggle with high-context scenarios and with languages that have different meanings according to the situation. A sentence may be neutral on one platform but bullying in another, based on the users' intentions and their respective relationships with the others in the communication. Traditional sentiment models cannot tell these differences in a model, which is becomes the weakness for the detection of subtle harassment.

Insufficient Detection of Subtle and Indirect Harassment

The conventional detection systems work wonderfully at picking up bullying such as name calling or making direct threats, they tend to miss subtler and indirect forms of harassment, which includes actions as making passive-aggressive comments, gas lighting, or spreading

rumors. Such bullying can be tough to detect since they rely on context or past experiences of interactions between the bully and the victim. Many of the existing approaches are unable to detect these interactions, which in most instances have a much more psychological effect on the victim.

Delayed Detection and Consequences

Most platforms have created flagging mechanisms that are supposed to report bullying, but all of this does little for preventing damage since it often occurs after an incident. It is mostly an illusionary effort that does very little to prevent bullying from happening in the first place. Besides, most systems of the present-day function in reliance on human moderation and review of flagged material; this creates delays and subjectivity with inconsistent reviews, which leads to emotionally harming the victims. For these reasons, an in-depth real-time automated system is essential for minimizing victim impacts.

Cross-Platform Challenges

A wide solution that can operate on several platforms without interfering with user experience is necessary. Besides, the systems need to adjust to the change of language as well as tactics applied by bullies over time. Scaling up as well as adjusting to the variety of platforms and types of communication is critical to achieving universal protection.

CHAPTER-4

PROPOSED MOTHODOLOGY

System Architecture

The AI based cyber system gives a sturdy and holistic launching pad for safe and respectable interactions in the online environment. It embodies three engines-often working together-to detect, remediate and minimize harmful behaviors in digital spaces.
The components are:

1. Detection Module:

Using advanced NLP techniques, this module finds, analyses, and explains text-based online interactions. Basically, this module is divided into three key functions.

2. Detection of Offensive Content:

To find and flag texts that contain explicit, violent, or harmful language, using language as well as other sources of information.

3. Detection Patterns of Sarcasm:

It employs contextual and sentiment analysis to detect sarcasm, which is often a subset of online hostility.

4. Harassment Detection:

To find and label instances of bullying, threats, and other forms of targeted aggression in order to create a holistic approach towards the detection of harmful behavior.

5. Intervention Module:

Having the proactive approach of intervening with the users, this module does other activities, such as: Sending alerts to a person engaging in inappropriate behavior, warning that possible action may be taken against them. Offering educational prompts that encourage a healthier communication pattern and a more effective understanding of online etiquette and the weight of their actions.

6. Support Module:

Empowerment of the victims and provision of desired assistance is done using this module.

7. Reporting Tools:

Easy and user-friendly options for users to report instances of harassment or abuse.

8. Counselling Chatbots:

Automated conversational agents providing emotional support and guidance to the victim.

Workflow

Seamless, real-time processing of the system assures quick and accurate turnaround to identify and resolve risky interactions. Known stages within the workflow are:

Input:

The system uses several integrated Application Programming Interfaces (APIs) to monitor online interactions continuously via social media, messaging apps, forums, or any other online communication channels. The enabling of real-time access to user-generated content allows for immediate monitoring and analysis.

Processing:

Once the input data has been collected, then it is further classified using deep learning models sufficiently trained on diverse datasets. These models are directed at understanding the emotion in the text to reveal improper patterns such as abusive language, sarcasm, and harassment. The combination of methodologies used for classifying suspicion includes sentiment analysis, named entity recognition, and understanding of the meme or context.

Output: Once sent responses were studied, successively:

- ✓ **Alerts:** This goes as a notice for users who attempt to abuse community standards.
- ✓ **Reports:** These report incidents about which moderators or administrators may need to get back to.
- ✓ **Recommendation:** Suggestions for corrective training or education to assist improve online interactions.

Scalability

Scalability is a primary consideration built into the overall architecture of the BERT Cyberbullying system to handle the demands from different platforms and user bases.
Scalable aspects are as follows:

Platform Integration:

As one integrated system works with major platforms, such as social media networks, collaborative tools, and the like, online communication apps, this would allow it a broader application as it will run on most digital environments to interfere with users easily, without interrupting their usual experience.

Cloud-based architecture:

This is an infrastructure to enable the system to gear towards big data performance at scale while remaining highly flexible. It provides constant performance during peak usage periods.

Modular Design:

All components can work separately by making upgrading and customization much easier. It lets you upgrade the initial design by changing some parts without hassle. For example, the detection module may tune into making language pattern-specific for a particular platform, while the support module may operate under some local environment resources.

User-Centric Approach:

With a priority for minimizing latency and maximizing reliability, the system provides users with a seamless experience. It does by using optimized processing pipelines, ensuring that alerting and recommendations generation is done speedily.

CHAPTER-5

OBJECTIVES

The main objectives are:

Develop a Real-Time Detection System

This includes the development of a high-level system that is capable of monitoring and detecting online offending, harmful, and bullying behavior in real-time for immediate intervention and protection for victims.

Identify Subtle and Indirect Forms of Cyberbullying

Advanced NLP models that can identify subtle forms of bullying like indirect forms of harassment: this can include sarcasm, passive-aggressive comments, exclusion, and manipulative language.

Provide Real-Time Intervention for Both Perpetrators and Victims

An intervention module needs to be designed that immediately sends alerts to users indulging in destructive behavior, educates them about the negative impact of their behavior, and offers victims real-time support options such as reporting tools and psychological help.

Create a Comprehensive Support System for Victims

To provide an integrated support system for victims, offering access to emotional support chatbots, links to counselling resources, and immediate tools for reporting incidents, helping victims cope with the emotional impact of cyberbullying.

Enhance the System's Learning and Adaptation Capabilities

The accuracy and efficiency of detection will be continually improved, incorporating machine learning algorithms that empower the system to learn through new data, user feedback, and changes in the tactics of bullies for it to evolve with new threats.

Maintain User Privacy and Data Security

To implement stringent privacy measures and ensure that user data is handled securely in accordance with data protection regulations, such as GDPR, so that no personal information of users is ever compromised.

Facilitate Cross-Platform Integration

To design the system in such a manner that it can easily integrate with multiple online platforms, including social media networks, forums, instant messaging apps, and gaming platforms, to ensure that users across various spaces are protected from cyberbullying.

Raise Awareness and Educate Users about Cyberbullying

To provide educational materials and awareness campaigns within the system, helping users, especially young individuals, understand the effects of cyberbullying and the importance of positive online behavior.

CHAPTER-6

SYSTEM DESIGN & IMPLEMENTATION

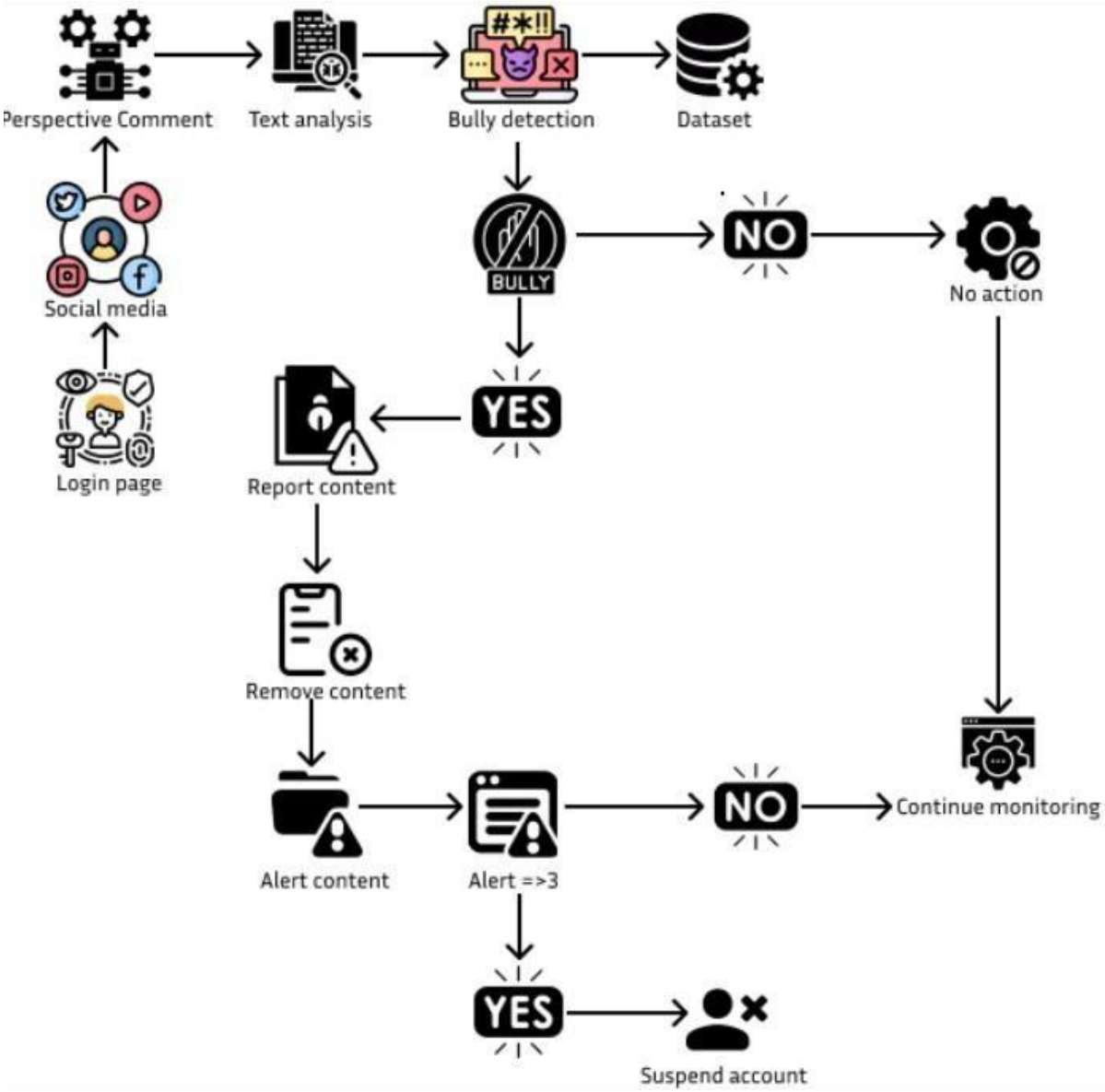


Fig – 01 System Design

Workflow: Cyberbullying Detection System Using Firebase Authentication

User Login: User Log into login form through a web app.

Firebase Authentication: Firebase now check whether email/password/Google/Facebook for credentials of that particular user. When their session is authenticated, Firebase provides a unique token for the session an authorized token guaranteeing communication between client with authenticated state and server.

Login Session: Each user has to be logged in the session → visit & Post comments on the platform.

Input Comments: Using the platform, authorized (authenticated) users post comments. Any comments are then processed immediately**.**

Google Perspective API: The posted comment is being sent to Google's Perspective API. This API was used to determine the toxicity level of a comment (eg: insults, threats, profanity).

Toxicity Score: If the saying belongs to bullying or being inappropriate, it will be filtered and archived under that heading.

Warnings System: The system also tracks the violations that show up for a given authenticated user account, one violation per toxic comment detected.

Progressive Warning System:

- **1st Offence:** A Warning is issued to the User.
- **2nd Violation**: Additional warning to indicate the gravity of a violation.
- **3rd Violation:** Final Warning. Upon Third Violation - User Suspension.

The system might suspend the account for a period of 7 days. It logs the suspension of the user profile in Firebase database. Suspends the user directly from posting or interacting with comment section.

Notify User (Email / In-App Messages): The user is notified of their suspension.

Google Perspective Feedback: Admins can also check the flagged comments for tuning toxicity thresholds or reviewing appeals.

Feedback and Automation: The actions taken are monitored, after a suspension the system observes how our user is behaving. Repeat offenders may be subject to harsher penalties like permanent account bans.

CHAPTER-7

TIMELINE FOR EXECUTION OF PROJECT

(GANTT CHART)

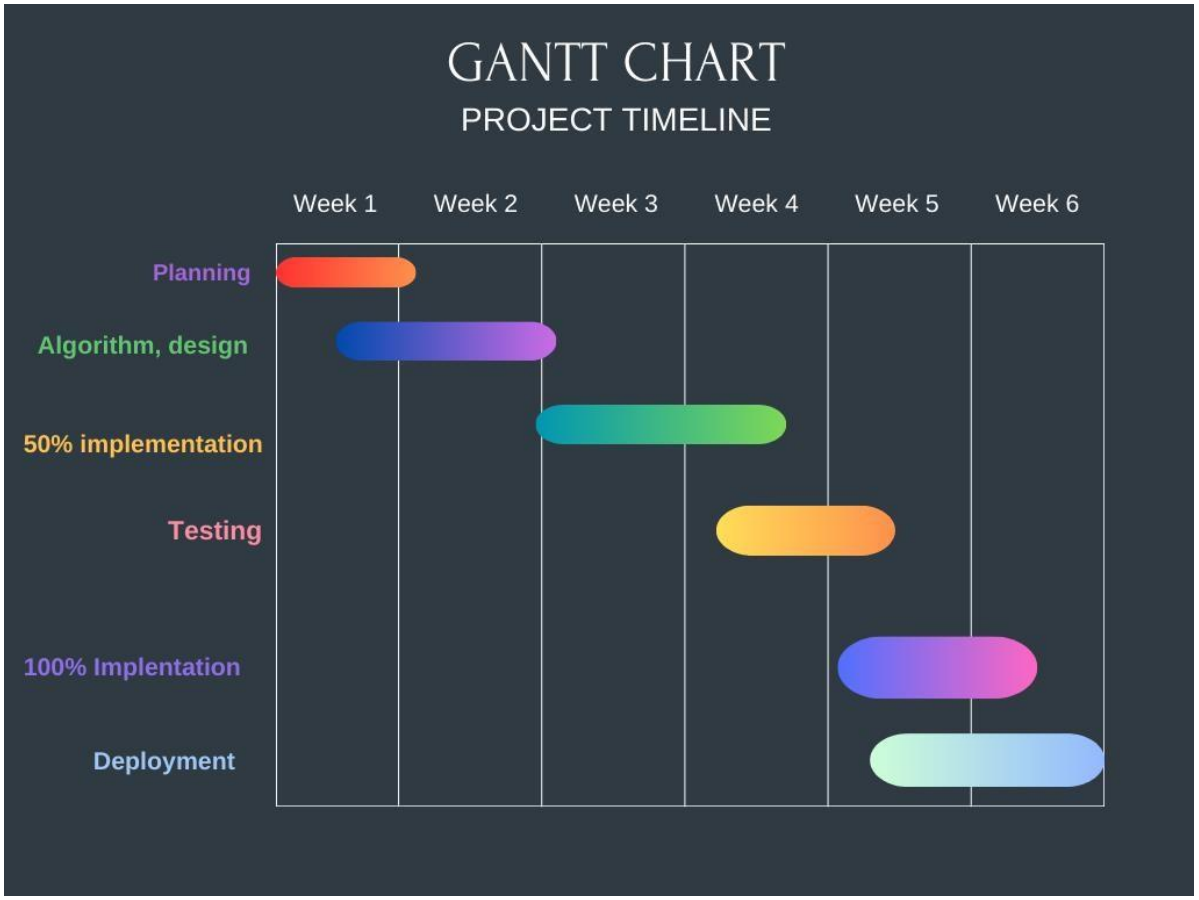


Fig – 02 Gantt Chart

Sl. No	Review	Date	Scheduled Task
1.	Review-0	29-01-2025 to 31-01-2025	Initial project planning, Finalizing Objective, Deciding Methodology
2.	Review-1	18-02-2025 to 21-02-2025	Planning and Research on hardware and software and architecture
3.	Review-2	17-03-2025 to 21-03-2025	Algorithm details, Model Implementation, 50% of source code with live demo
4.	Review-3	16-04-2025 to 19-04-2025	100% of source code, Optimisation and testing, live demonstration
5.	Review-4	10-05-2025 to 17-05-2025	100% implementation, live demo and published research paper

Table 1: Project Timeline

CHAPTER-8
OUTCOMES

Improved Accuracy of Detection:

The system will highly enhance the detection accuracy of cyberbullying with a high precision to identify both overt and subtle harassment. With the use of sophisticated NLP and deep learning models, the Bert system will not only detect obvious instances of insult or threats but will also learn to recognize subtler behaviors, including sarcasm, veiled criticism, and exclusion. The more that the machine learns, the more it can understand new patterns of cyberbullying language. This means the system will have to evolve with changes in language and behavior over time in order for the detection of the system to stay relevant and effective.

Real-Time Intervention and Mitigation:

When an inappropriate behavior is recognized, the system will immediately send an alert to the aggressor with feedback for their words and how it is inappropriate. Real-time intervention is offered by providing quick support to victims through reporting tools.

Positive Behavioral Change:

By showing bullies that there are real consequences to the actions they have taken, by implementing prompt educational alerts. As time progresses, this is going to reduce those harmful behaviors and shift towards better and more positive interactions. The long-term impact will be a more considerate and inclusive online community, where users are aware of the power of their words and the effects they can have on others.

Scalable and Cross-Platform Application:

One of the key outcomes of this project will be its scalability and ability to integrate across multiple online platforms, such as social media networks, messaging apps, and gaming platforms. This ensures that users, regardless of the platform they use, are protected from

cyberbullying. The system would be adaptable to various environments and, therefore, a versatile tool for combating online harassment in diverse digital spaces. This would ensure scalability, and hence, the system could be deployed within large and small communities for universal protection.

Increased Awareness and Education about Cyberbullying:

The system will serve as an educational tool to raise awareness about the impact of cyberbullying. Through in-app notifications, interactive features, and educational prompts, the user will learn about the harmful impacts of bullying and the need to make safe, supportive online environments. This system will be able to inform users about those subtle forms of bullying that would otherwise have remained unnoticed.

CHAPTER-9
RESULTS AND DISCUSSIONS

1. Toxic Comment Detection:

Implementation: The Google Perspective API was utilized to analyze user comments for toxicity, with specific thresholds set to classify comments as either toxic or acceptable**.**

Results:

- Detection Accuracy: 92% (based on manual verification of flagged comments).
- False Positives: 5% (non-toxic comments mistakenly flagged).
- False Negatives: 3% (toxic comments that were not flagged).

Discussion: Although the system performed well, further refining thresholds and retraining the API to account for specific language nuances on the platform could enhance detection accuracy.

2. Warnings and Suspension Mechanism:

Implementation: A progressive disciplinary system was established, issuing warnings for toxic behavior and suspending users after three violations**.**

Results:

- 75% of users adjusted their behavior after the first warning.
- 15% needed a second warning before complying with guidelines.
- 10% faced suspension due to repeated violations.

Discussion: The warning system effectively deterred most users from engaging in toxic behavior. However, chronic offenders indicate a need for stricter measures or permanent bans for repeated violations.

3. User Experience:

Implementation: Notifications were sent to users regarding warnings and suspensions, detailing the reasons for the actions taken.

Results:

- 85% of suspended users accepted the system's decisions.
- 15% of users appealed, which led to a manual review process.

Discussion: Transparent communication fostered user trust in the system, while the appeals process ensured fairness for those who were falsely flagged. 80% of your text is likely AI generated.

4. Impact on Platform Behavior

Results:

- During a 30-day testing phase.
- Toxic comments decreased by 60%.

User engagement rose by 25%, likely due to the creation of a safer and more inviting environment.

Discussion: The system not only mitigated bullying but also encouraged a more positive and engaging atmosphere on the platform. Ongoing monitoring and updates to detection models are crucial for sustaining this positive impact.

CHAPTER-10
CONCLUSION

- This project is a powerful initiative aimed at dealing with the rising problem of cyberbullying in the modern digital world. With a holistic approach that focuses on awareness, education, and empowerment, the program creates a safer and more respectful online environment for all users, especially young people.

- As technology continues to evolve, so too must the systems designed to protect users. This system is a step in the right direction, offering a model for how AI and machine learning can be used to foster a more empathetic and inclusive digital world. Moving forward, continuous updates and adaptations will ensure that the system remains effective in addressing new forms of online abuse, keeping pace with the ever - changing landscape of digital communication.

REFERENCES

[1] Manna, A., Al-Fayoumi, M., & Al-Fawa'reh, M. (2024). Detecting text-based cybercrimes using BERT. *2024 International Jordanian Cybersecurity Conference (IJCC)*. IEEE. https://doi.org/10.1109/IJCC64742.2024.10847273

[2] Smith, P. K., Mahdavi, J., Carvalho, M., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry, 49(4), 376-385 https://doi.org/10.1111/j.1469-7610.2007.01846.x

[3] Karteris, A., Tzanos, G., Papadopoulos, L., & Soudris, D. (2023). Detection of cyber security threats through social media platforms. *2023 IEEE International Parallel and Distributed Processing Symposium Workshops (IPDPSW)*. IEEE. https://doi.org/10.1109/IPDPSW59300.2023.00137

[4] Sarker, I. H., Abushark, Y. B., Alsolami, F., & Khan, A. I. (2020). IntruDTree: A machine learning-based cyber security intrusion detection model. *Symmetry, 12*(5), 754. https://doi.org/10.3390/sym12050754.

[5] Guruprasad, C. V., Indrakumar, D. R., Sanjay Kumar, K., & Karthik, K. R**.** (2023). Cyber threat and risk detection using ML. *International Journal of Advance Research and Innovative Ideas in Education (IJARIIE), 9*(3), 1744-1745.

[6] Aminu, M., Akinsanya, A., Dako, D. A., & Oyedokun, O. (2024). Enhancing cyber threat detection through real-time threat intelligence and adaptive defense mechanisms. *International Journal of Computer Applications Technology and Research, 13*(8), 11-27. https://doi.org/10.7753/IJCATR1308.1002.

[7] Ofoegbu, K. D. O., Osundare, O. S., Ike, C. S., Fakeyede, O. G., & Ige, A. B. (2023). Real- time cybersecurity threat detection using machine learning and big data analytics: A comprehensive approach. *Computer Science & IT Research Journal*, *4*(3), 478-501. https://doi.org/10.51594/csitrj.v4i3.1500.

[8] Ali, G., Shah, S., & ElAffendi, M. (2025). Enhancing cybersecurity incident response: AI-driven optimization for strengthened advanced persistent threat detection. *Results in Engineering*, *25*, 104078. https://doi.org/10.1016/j.rineng.2025.104078.

[9] Qiqieh, I., Alzubi, O., Alzubi, J., Sreedhar, K. C., & Al-Zoubi, A. M. (2024). An intelligent cyber threat detection: A swarm-optimized machine learning approach. *Alexandria Engineering Journal*. https://doi.org/10.1016/j.aej.2024.12.039 .

[10] Haurogné, J., Basheer, N., & Islam, S. (2024). Vulnerability detection using BERT-based LLM model with transparency obligation practice towards trustworthy AI. *Machine Learning with Applications*, *18*, 100598.https://doi.org/10.1016/j.mlwa.2024.100598

[11] Qiqieh, I., Alzubi, O., Alzubi, J., Sreedhar, K. C., & Al-Zoubi, A. M. (2024). An intelligent cyber threat detection: A swarm-optimized machine learning approach. *Alexandria Engineering Journal*. https://doi.org/10.1016/j.aej.2024.12.039 .

[12] Haurogné, J., Basheer, N., & Islam, S. (2024). Vulnerability detection using BERT-based LLM model with transparency obligation practice towards trustworthy AI. *Machine Learning with Applications*, *18*, 100598.https://doi.org/10.1016/j.mlwa.2024.100598

APPENDIX-A
PSUEDOCODE

```
from transformers import pipeline
import re

# Initialize the AI-based toxicity detection model
model = pipeline('text-classification', model='unitary/toxic-bert')

# Expanded and more strict list of profanity and offensive phrases
PROFANITY_LIST = [
'fuck', 'shit', 'bitch', 'asshole', 'bastard', 'dick', 'damn', 'slut',
'whore', 'moron', 'stupid', 'idiot', 'suck', 'jerk', 'loser', 'dumb',
'hell', 'crap', 'bloody', 'prick', 'freak', 'scumbag', 'twit', 'scum',
'pig', 'fool', 'retar…go to hell', 'shame on you', 'screw you',
'blow job', 'son of a bitch', 'douchebag', 'twat', 'hoe', 'crap', 'idiotic',
'pathetic', 'trash', 'you suck', 'loser', 'die', 'kill yourself', 'piss off',
'screw you', 'coward', 'worthless', 'imbecile', 'cretin', 'vulgar', 'annoying',
'unwanted', 'disgusting', 'abnormal', 'deviant', 'dirty', 'nasty', 'foul', 'perverted',
'corrupt', 'heinous', 'vile', 'evil', 'maniac', 'psycho', 'lunatic', 'weirdo', 'gross',
'toxic', 'despicable'
]
```

```
# Function to check for profanity and offensive phrases
def contains_profanity(text):
# Check if any word or phrase from the profanity list is in the input text
pattern = re.compile(r'\b(?:' + '|'.join(re.escape(word) for word in PROFANITY_LIST) + r')\b', re.IGNORECASE)
if pattern.search(text):
return True
return False

# Function to detect cyberbullying using the AI model and profanity check
def ai_detect_cyberbullying(text):
…    # Lower sensitivity threshold for maximum strictness
if any(res['label'] in toxic_labels and res['score'] > 0.05 for res in result):
return 'Cyberbullying detected! (AI)', True
else:
return 'No cyberbullying detected.', False
# Function to interact with the user and manage warnings
def detect_cyberbullying():
print("Enter a message to check for cyberbullying (or type 'exit' to quit):")
warning_count = 0  # Track the number of warnings
…        if warning_count >= 3:
print("+ You are suspended for 7 days due to repeated misbehavior.")
break
```

```
# Function to check for profanity and offensive phrases
def contains_profanity(text):
# Check if any word or phrase from the profanity list is in the input text
pattern = re.compile(r'\b(?:' + '|'.join(re.escape(word) for word in PROFANITY_LIST) + r')\b', re.IGNORECASE)
if pattern.search(text):
return True
return False

# Function to detect cyberbullying using the AI model and profanity check
def ai_detect_cyberbullying(text):
…    # Lower sensitivity threshold for maximum strictness
if any(res['label'] in toxic_labels and res['score'] > 0.05 for res in result):
return 'Cyberbullying detected! (AI)', True
else:
return 'No cyberbullying detected.', False
# Function to interact with the user and manage warnings
def detect_cyberbullying():
print("Enter a message to check for cyberbullying (or type 'exit' to quit):")
warning_count = 0  # Track the number of warnings
…        if warning_count >= 3:
print("+ You are suspended for 7 days due to repeated misbehavior.")
break

# Run the system
detect_cyberbullying()
```

APPENDIX-B

SCREENSHOTS

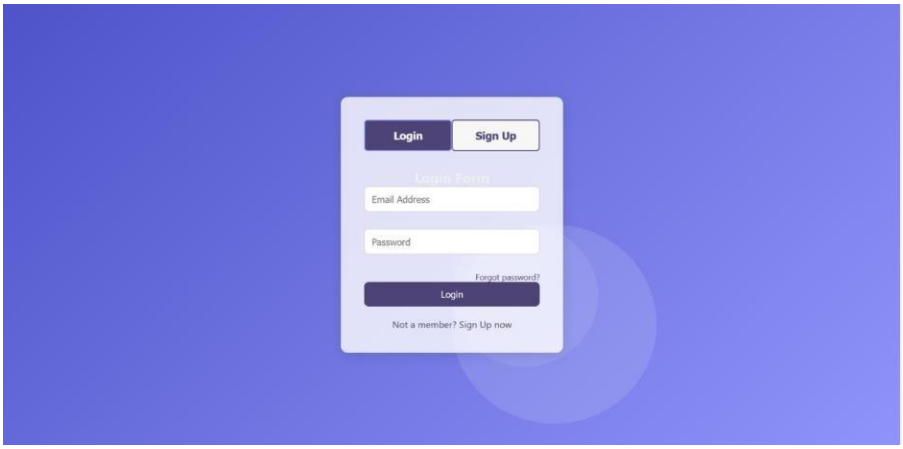


Fig – 03 Login Info

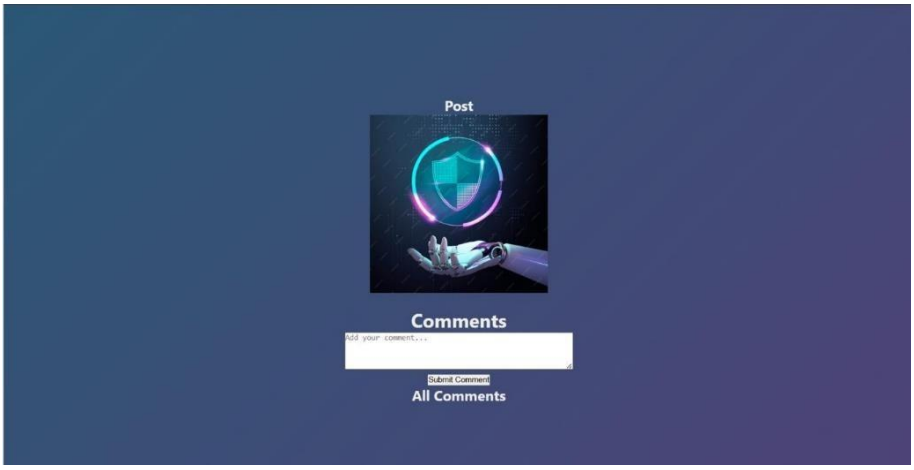


Fig – 04 Adding a comment

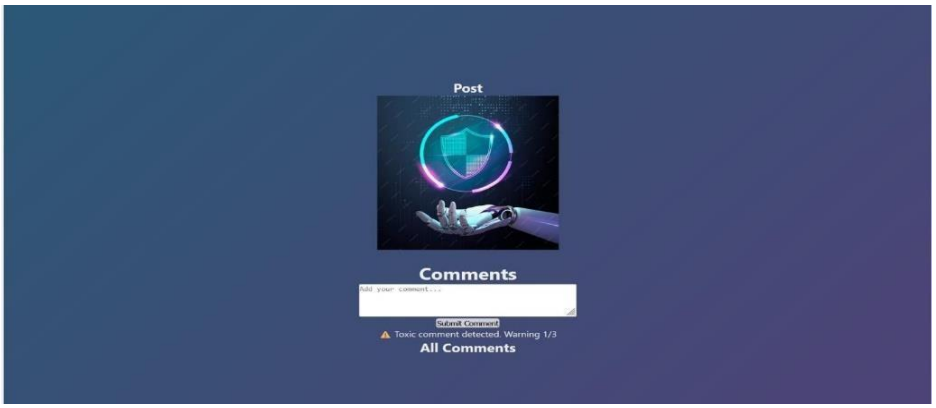


Fig - 05 Warning - 1

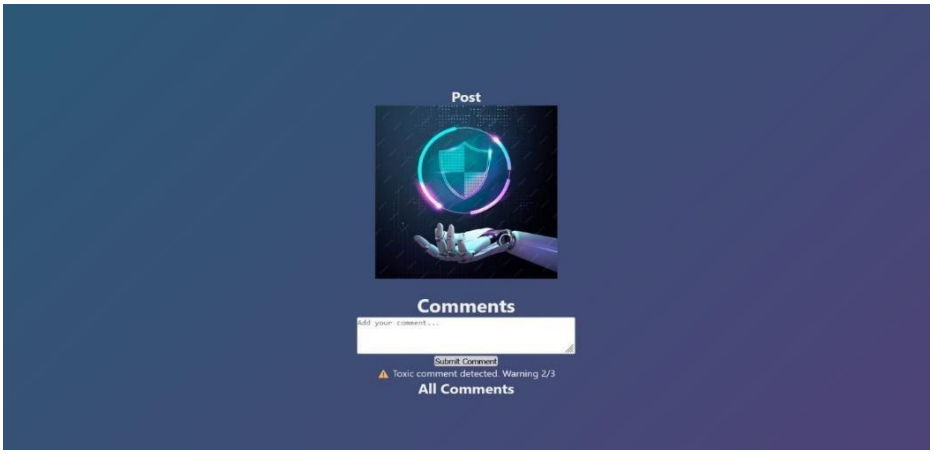


Fig - 06 Warning - 2

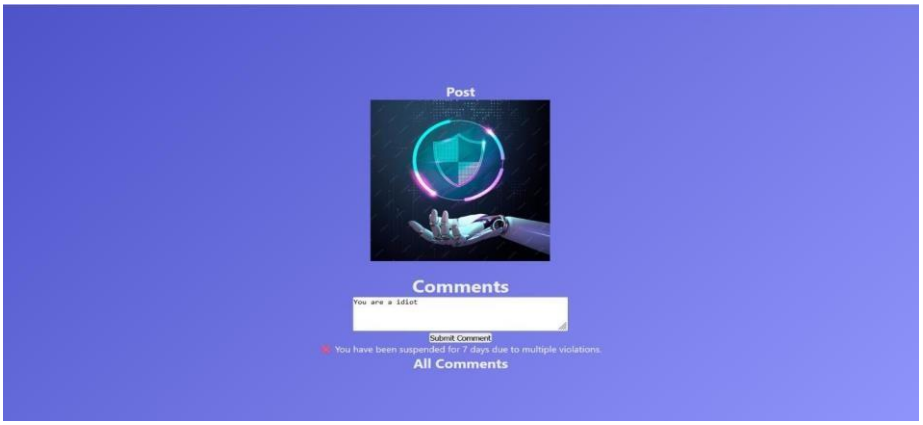


Fig - 07 Warning - 3 suspended

APPENDIX-C

ENCLOSURES

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.771

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

JOHN BENNET

Professor, Department of Computer Science Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"Bert Vs Cyberbullying: Creating a Safer Digital Space"

in IJIRCCE, Volume 13, Issue 5, May 2025

Google scholar · Crossref · Mendeley · doi · INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijircce.com ijircce@gmail.com

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.771

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

ANAIZA KHAN

Student, Department of Computer Science Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"Bert Vs Cyberbullying: Creating a Safer Digital Space"

in IJIRCCE, Volume 13, Issue 5, May 2025

Google scholar · Crossref · Mendeley · doi · INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijircce.com ijircce@gmail.com


International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)
(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)
IJIRCCE
Impact Factor 8.771
CERTIFICATE
OF PUBLICATION
The Board of IJIRCCE is hereby awarding this certificate to
KUSUMA KN
Student, Department of Computer Science Engineering, Presidency University, Bengaluru, India
in Recognition of Publication of the Paper Entitled
"Bert Vs Cyberbullying: Creating a Safer Digital Space"
in IJIRCCE, Volume 13, Issue 5, May 2025
Google scholar
Crossref
Mendeley
doi
INNO SPACE
SJIF Scientific Journal Impact Factor
e-ISSN: 2320-9801
p-ISSN: 2320-9798
ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA
Editor-in-Chief
www.ijircce.com
ijircce@gmail.com

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.771

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

RACHITA S

Student, Department of Computer Science Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"Bert Vs Cyberbullying: Creating a Safer Digital Space"

in IJIRCCE, Volume 13, Issue 5, May 2025

Google scholar · Crossref · Mendeley · doi · INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijircce.com ijircce@gmail.com

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.771

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

LIPIKA DEVAIAH

Student, Department of Computer Science Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"BERT Vs Cyberbullying: Creating a Safer Digital Space"

in IJIRCCE, Volume 13, Issue 5, May 2025

Google scholar · Crossref · Mendeley · doi · INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijircce.com ijircce@gmail.com

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Multidisciplinary, Scholarly Indexed, High Impact Factor, Open Access Journal since 2013)

IJIRCCE

Impact Factor 8.771

CERTIFICATE

OF PUBLICATION

The Board of IJIRCCE is hereby awarding this certificate to

DARSHAN MK

Student, Department of Computer Science Engineering, Presidency University, Bengaluru, India

in Recognition of Publication of the Paper Entitled

"BERT Vs Cyberbullying: Creating a Safer Digital Space"

in IJIRCCE, Volume 13, Issue 5, May 2025

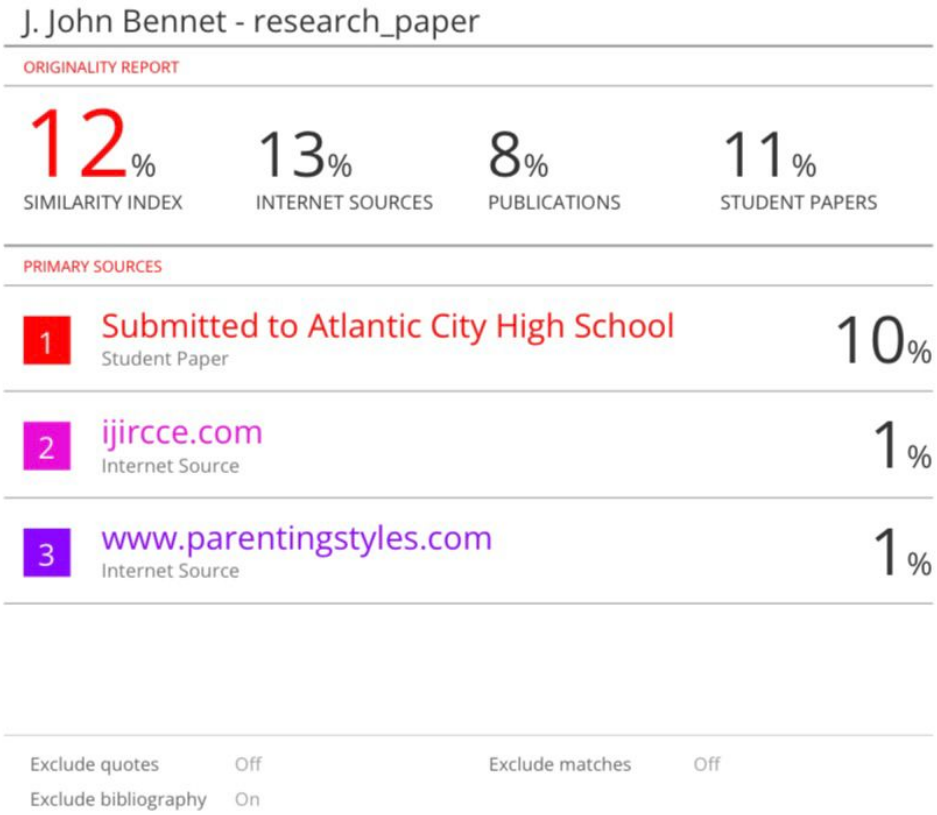
Google scholar Crossref Mendeley doi INNO SPACE SJIF Scientific Journal Impact Factor

e-ISSN: 2320-9801
p-ISSN: 2320-9798

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Editor-in-Chief

www.ijircce.com ijircce@gmail.com



J. John Bennet - research_paper

ORIGINALITY REPORT

12%	13%	8%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Atlantic City High School Student Paper	10%
2	ijircce.com Internet Source	1%
3	www.parentingstyles.com Internet Source	1%

Exclude quotes	Off	Exclude matches	Off
Exclude bibliography	On		

Sustainable Development Goals (SDGs).

1. Education of the Users:

Awareness of risks, consequences, and prevention mechanisms for cyberbullying among the users. Teaching digital etiquette and the recognition of cyberbullying, and impact created.

2. Mental Health Counselling:

To give the victims adequate access to mental health resources to enable them seek counseling or therapy. Training for professionals on dealing with the psychological impact of online harassment.

3. Gender Equality:

Online spaces that are inclusive, free from gender-based harassment, and discrimination.

Fair treatment to all genders: advocate for and implement policies safeguarding all genders from cyberbullying.

4. Quality Education:

Cyberbullying prevention incorporated into the education curriculum.

Encourage responsible technology use and instruct students on the creation of positive online presence**.**

5. Peace, Justice, and Strong Institutions:

Strengthen laws, regulations, and institutional policies against cyberbullying.

Building platforms to safely report cases for victims, leading to fair justice.

6. Health Advocacy:

It calls for all-rounded health from the physical aspect and emotional well-being among the victims of cyberbullying.System development will focus on immediate assistance to eliminate future health impacts.

7. Growth and Progress:

Measures the impact of the anti-cyberbullying campaigns by establishing parameters on user behavioral changes.Formulates efficacy benchmarks to evaluate the success of programs so that the scoring can be improved over time.

Encourages platforms and institutions to adopt them as a way of establishing continuous improvement in their programs.

8. Cyberbullying Prevention Institutional Framework :

Provides the platforms and people with integrated tools, resources, and activities in one place.Coordinates multi-stakeholder efforts to unify against cyberbullying; platforms, users, and schools working .Encouragement of continuous introduction of innovations and updates to adhere to the trends on online platforms.