



**SRI RAMACHANDRA**

**INSTITUTE OF HIGHER EDUCATION AND RESEARCH**

(Category - I Deemed to be University) Porur, Chennai

**SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY**

# **AI POWERED CYBERBULLYING THREAT DETECTION WITH MENTAL HEALTH CHATBOT**

**CYB23IN201 – INTERNSHIP II**

**PROJECT REPORT**

*Submitted by*

**SARAH IRENE RIYA N– E0223026**

*In partial fulfilment for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(Cyber Security & Internet of Things)**

**Sri Ramachandra Faculty of Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**June, 2025**



**SRI RAMACHANDRA**

**INSTITUTE OF HIGHER EDUCATION AND RESEARCH**

(Category - I Deemed to be University) Porur, Chennai

**SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY**

# **AI POWERED CYBERBULLYING THREAT DETECTION WITH MENTAL HEALTH CHATBOT**

**CYB23IN201 – INTERNSHIP II**

**PROJECT REPORT**

*Submitted by*

**SARAH IRENE RIYA N– E0223026**

*In partial fulfilment for the award of the degree of*

**BACHELOR OF TECHNOLOGY**

**in**

**COMPUTER SCIENCE AND ENGINEERING**

**(Cyber Security & Internet of Things)**

**Sri Ramachandra Faculty of Engineering and Technology**

**Sri Ramachandra Institute of Higher Education and Research, Porur, Chennai -600116**

**June, 2025**



# **SRI RAMACHANDRA**

**INSTITUTE OF HIGHER EDUCATION AND RESEARCH**

(Category - I Deemed to be University) Porur, Chennai

**SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY**

## **BONAFIDE CERTIFICATE**

Certified that this project report “**AI POWERED CYBERBULLYING THREAT DETECTION WITH MENTAL HEALTH CHATBOT**” is the bonafide record of work done by “**SARAH IRENE RIYA N– E0223026**” who carried out the internship work under my supervision.

**Signature of the Supervisor**

**Signature of Programme Coordinator**

**Dr. Jayanthi G**

**Dr. Jayanthi G**

**Associate Professor and Program Coordinator**

**Associate Professor and Program Coordinator**

Department of Cyber Security and Internet of Things

Department of Cybersecurity and Internet of things

Sri Ramachandra Faculty of Engineering and Technology,

Sri Ramachandra Faculty of Engineering and Techno

SRIHER, Porur, Chennai-600 116.

SRIHER, Porur, Chennai-600 116.

**Evaluation Date:**

**INTERNAL EXAMINER**





**EXTERNAL EXAMINER**

# Plagiarism Report




## 4% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

### Match Groups

-  **14 Not Cited or Quoted 3%**  
Matches with neither in-text citation nor quotation marks
-  **1 Missing Quotations 0%**  
Matches that are still very similar to source material
-  **1 Missing Citation 0%**  
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**  
Matches with in-text citation present, but no quotation marks

### Top Sources

- 1%  Internet sources
- 1%  Publications
- 3%  Submitted works (Student Papers)

### Integrity Flags

#### 0 Integrity Flags for Review

No suspicious text manipulations found.

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Signature of the Internship Coordinator

Signature of the Supervisor



# **SRI RAMACHANDRA**

**INSTITUTE OF HIGHER EDUCATION AND RESEARCH**

(Category - I Deemed to be University) Porur, Chennai

**SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY**

## **ACKNOWLEDGEMENT**

I take this opportunity to express my gratitude to **Prof. T. Ragunathan, Dean, Prof. A.Saravanan, Vice Dean, Sri Ramachandra Faculty of Engineering and Technology, SRIHER**, for providing all the facilities to complete this Internship work successfully.

I express my sincere gratitude to **Dr. G. Jayanthi, Programme Coordinator, Department of Cybersecurity & Internet of Things, SRET**, for their support and for providing the required facilities for carrying out this study.

I wish to thank my faculty supervisor(s), **Dr. Jayanthi G,** Department of Cybersecurity and Internet of Things, SRET and **HR Manager, Mr.DANIEL DIVAKAR** for extending help and encouragement throughout the project. Without his/her continuous guidance and persistent help, this project would not have been a success for me.

I am grateful to all the members of the Sri Ramachandra Faculty of Engineering and Technology, my beloved parents and my friends for extending their support, which helped us to overcome obstacles in the study.

## TABLE OF CONTENTS

<b>CHAPTER NO.</b>	<b>TITLE</b>	<b>PAGE NO.</b>
	<b>ABSTRACT</b>	<b>VIII</b>
	<b>LIST OF TABLES</b>	<b>IX</b>
	<b>LIST OF FIGURES</b>	<b>IX</b>
	<b>LIST OF SYMBOLS</b>	<b>X</b>
<b>1</b>	<b>INTRODUCTION</b>	<b>1</b>
	1.1 Need for the Study/Technology	1
	1.2 Applications	2
	1.3 Motivation behind the Study	2
	1.4 Challenges to be addressed	3
<b>2</b>	<b>LITERATURE REVIEW</b>	<b>4</b>
	2.1 Literature Survey	4
	2.2 Review of existing system	9
<b>3</b>	<b>PROBLEM STATEMENT AND OBJECTIVES</b>	<b>12</b>
	3.1 Problem statement	12
	3.2 Objectives	12
<b>4</b>	<b>METHODOLOGY</b>	<b>14</b>
	4.1 Workflow Diagram	14
	4.2 Modules	15
	4.2.1. Objective	16
	4.2.2 Proof of Concept	17
	4.2.3 Formulation of Methods	18
	4.2.4 Prototype Design	20

	4.2.4 Implementation	21
	4.2.5 Working Model/ Dashboard/ Application (with screenshots)	24
<b>5</b>	<b>RESULTS AND DISCUSSIONS</b>	<b>25</b>
	5.1 Data Collection	25
	5.2 Computing Configuration	26
	5.3 Experimental Evaluation	27
	5.3.1 Comparison of Results	27
<b>6</b>	<b>CONCLUSION</b>	<b>30</b>
	6.1 Research Findings	30
	6.1 Conclusion	30
	6.2 Scope for Further Enhancement	31
	<b>REFERENCES</b>	<b>32</b>
	Journal References	32
	Web References	33
	<b>APPENDICES</b>	<b>34</b>
	Appendix-1: Screenshots	34
	Appendix-II: Sample code	36
	<b>WORKLOG</b>	<b>41</b>
	<b>OFFER LETTER</b>	<b>46</b>
	<b>CERTIFICATE OF COMPLETION</b>	<b>47</b>
	<b>ATTENDANCE FORM</b>	<b>48</b>

# ABSTRACT

Cyberbullying is becoming more common and harmful in today's online world, especially among young people who spend a lot of time on platforms like Instagram, Twitter, and YouTube. Many of them are exposed to rude comments, threats, or emotional bullying, which often affects their mental health. They may experience anxiety, low self-esteem, or feel socially withdrawn. While there are tools that try to block or report harmful messages, they rarely offer any kind of support to the person being targeted.

This project was created to help solve that problem by developing a system that not only detects bullying in messages but also offers the user emotional support right away. It makes use of Natural Language Processing (NLP) with advanced AI models like BERT for detecting harmful language, and Toxic-BERT for identifying more specific issues like hate speech, insults, or threats. A simple rule-based system is also added to catch common offensive words that may be missed by the models.

If bullying is found in the text, the system guides the user to a built-in chatbot powered by BlenderBot — an AI created by Meta that can carry on friendly and meaningful conversations. The chatbot is designed to respond in a caring and supportive way. Everything is put together using a clean, user-friendly interface with Gradio, so people can use the tool easily and in private.

In short, this system does two things: it helps people stay safe from online abuse, and it gives them someone to talk to when they're upset. It's a tool that can be used in colleges, schools, or online platforms to make digital spaces feel safer, more supportive, and more human.



## LIST OF TABLES

<b>TABLE No</b>	<b>Table Description</b>	<b>Page No</b>
2.1	Literature Survey	4
2.2	Review of existing system	9

## LIST OF FIGURES

<b>Figure No</b>	<b>Figure description</b>	<b>Page No</b>
4.1	Bullying detection	24
4.2	Mental health chatbot	24
5.1	Evaluation metrics of bert	27
5.2	Logistic regression result	27
5.3	Logistic regression model evaluation metrics	28
5.4	SVM model result	28
5.5	SVM model evaluation metrics	28
5.6	BERT model result	29
5.7	BERT model evaluation metrics	29
A-1.1	Cyberbullying detection with Analysis result And Bullying classification: Bullying detected	34
A-1.2	Cyberbullying detection with Analysis result And Bullying classification: Not Bullying	34
A-1.3	Mental Health chatbot response with emotional Support and crisis phrase detection	35
A-1.4	Emergency support Info Display	35

## LIST OF SYMBOLS

Symbol / Term	Meaning
<b>NLP</b>	Natural Language Processing
<b>BERT</b>	Bidirectional Encoder Representations from Transformers
<b>TP</b>	True Positive — correctly detected bullying instance
<b>FN</b>	False Negative — bullying not detected
<b>F1-Score</b>	Harmonic mean of Precision and Recall
<b>Precision</b>	$TP / (TP + FP)$ — How many predicted positives are correct
<b>Recall</b>	$TP / (TP + FN)$ — How many actual positives were detected
<b>API</b>	Application Programming Interface
<b>UI</b>	User Interface
<b>BlenderBot</b>	Generative chatbot model developed by Meta AI
<b>Toxic-BERT</b>	A fine-tuned BERT model trained to detect toxic content

# **CHAPTER 1**

## **INTRODUCTION**

### **1.1 Need for the Study/Technology**

Cyberbullying has emerged as a serious concern in today's digital ecosystem, particularly affecting teenagers and young adults who are active on social media. Victims often experience psychological distress, including anxiety, depression, and social isolation, yet many incidents go unnoticed or unaddressed in real time. Existing solutions largely focus on detecting offensive content using rule-based filters or basic keyword matching, which are often inadequate in capturing the subtle, contextual nature of online abuse.

Moreover, these systems rarely consider the emotional state of the victim post-detection. There is a lack of platforms that not only identify harmful language but also provide timely, empathetic support to affected users. The growing need for mental health-aware technologies calls for an AI-based system that blends accurate detection with conversational care.

This study is essential to develop a comprehensive, context-aware platform that addresses both the detection of cyberbullying and the emotional support needs of users. By combining Natural Language Processing with conversational AI, the project aims to enhance user safety, provide psychological relief, and foster a more compassionate online environment.

## **1.2 Applications**

- **Social Media Moderation:** Can be integrated into platforms like Instagram, X (Twitter), or Facebook to detect and respond to harmful comments in real time.
- **School/College Portals:** Useful for educational institutions to safeguard students and provide them a confidential space for emotional expression.
- **Mental Health Chat Tools:** Can function as a first-level support system for users showing signs of distress due to online abuse.
- **Online Communities & Forums:** Helps moderators identify toxic behavior and intervene early.
- **Anti-Bullying Awareness Platforms:** Can be used to demonstrate how AI can assist in digital well-being and create safer environments

## **1.3 Motivation behind the Study**

The idea for this study arose from the increasing reports of cyberbullying incidents and their long-lasting psychological impact. While many AI tools detect abusive content, few go a step further to offer emotional support to the victims. The gap between toxic content detection and mental health care prompted the development of a dual-purpose system. The motivation was to build something more empathetic, where AI isn't just flagging hate, but actively caring for users affected by it. This study was driven by the belief that technology should not only solve problems, but also support people emotionally.

## 1.4 Challenges to be addressed

- **Context Understanding:** Detecting bullying that is indirect, sarcastic, or masked within humor is difficult for even advanced models.
- **Bias Toward Direct Speech:** The model performs better on texts like "You're ugly" but may miss "She's disgusting" — highlighting a gap in data variety.
- **Generic Chatbot Responses:** Pretrained chatbots often respond with irrelevant or casual replies, reducing the emotional sensitivity of the system.
- **Real-time Performance:** Ensuring fast, accurate detection and response in a user-facing interface is critical for usability.
- **Data Privacy & Sensitivity:** Handling sensitive content requires caution to avoid misuse or misinterpretation of user input.

## CHAPTER 2

### 2.1 LITERATURE SURVEY

Author and Year	Methodology / Materials and Methods	Datasets	Results / Outcome	Advantage / Key Findings	Disadvantage / GAP
Zhang et al (2021)	Fine-tuned BERT for binary classification of bullying	Twitter & Formspring	Accuracy of 88.4%	Captures context better than traditional models	Struggles with sarcasm and indirect language
Mishra & Bhattacharyya (2020)	LSTM with GloVe word embeddings	YouTube Comment Dataset	F1-score of 85%	Effective in sequential abuse detection	Cannot handle long contextual chains
Unitary AI, (2021)	Toxic-BERT for multi-label toxicity detection	Jigsaw Toxic Comment	Recognizes toxic, insult, hate, threat labels	High interpretability and precision	Generic, lacks platform-specific nuance
Dinakar et (2012)	Rule-based & keyword matching	YouTube custom set	High precision, low recall	Simple and interpretable	Misses subtle or sarcastic bullying
Kumar et al (2018)	Deep neural network with attention mechanism	Wikipedia Talk Pages	84% accuracy	Attention model improves phrase- level focus	Not generalized beyond Wiki context
Badjatiya et al(2017)	CNN-LSTM + FastText embeddings	Twitter hate speech	Precision: 87.1%	Combines local + temporal features	Resource intensive
Zampieri et al( 2019)	Multi-task BERT for abusive	OLID (Offensive Language	Macro F1: 82%	Learns multiple tasks (offensive types)	Performance dips on minority classes

	language classification	Identification Dataset)			
Rajamanickam et al(2020)	Hybrid approach using ensemble classifiers	Facebook & Twitter datasets	Accuracy of 89%	Combines multiple weak learners	Slower inference speed
Sharma et al(2019)	Bi-GRU with attention for hate detection	Hindi-English code-mixed dataset	Accuracy: 80%	Supports multilingual input	Lower results on unseen language pairs
Cheng et al (2021)	Transformer + sentiment filter pipeline	Reddit bullying posts	86% F1 score	Filters emotional tone to improve clarity	Heavily dependent on lexicon
Facebook AI, (2020)	BlenderBot – Pretrained generative chatbot	Pushshift Reddit Corpus	Coherent, human-like conversation	Multi-turn context handling	Gives casual replies (not mental health-focused)
Wolf et al(2020)	Hugging Face Transformers framework	NA (toolkit paper)	Modular NLP pipeline support	Unified access to transformer models	Requires GPU for large models
Huang et al(2022)	Crisis-aware chatbot using BART	Counseling transcripts	Detects crisis phrases with 93% accuracy	Handles emotional tone well	Not publicly available
Kshirsagar et al(2021)	DistilBERT for real-time detection	Twitter dataset	Lightweight & 87% accuracy	Faster with similar performance	Lower accuracy on long text
Al-Garadi et al(2020)	Survey of AI-based cyberbullying detection methods	Multiple social platforms	Reviewed 60+ models	Comparative analysis across techniques	Lacks implementation details

Table 2.1- Literature Survey

This literature review highlights that while various models have been explored for cyberbullying detection, there is a clear gap in combining accurate detection with mental health support — which this project aims to bridge.

1. Zhang et al(2021) This work applies a fine-tuned BERT model to detect cyberbullying from social media texts. BERT's bidirectional attention mechanism enables it to understand contextual relationships between words, making it more effective than traditional models. The authors used Twitter and Formspring datasets, achieving an accuracy of 88.4%. While the model performed well on direct bullying instances, it struggled with sarcasm and indirect abuse.

2. Mishra & Bhattacharyya(2020) The authors proposed an LSTM-based architecture using GloVe embeddings to identify abusive and bullying content. The model was evaluated on a YouTube comment dataset and achieved an F1-score of 85%. The sequential nature of LSTM allowed better handling of time-dependent patterns in language, though it lacked the ability to capture long-range dependencies or context spread across sentences.

3. Unitary AI(2021) This research introduced Toxic-BERT, a model pre-trained on the Jigsaw Toxic Comment dataset for multi-label classification, including labels such as "toxic", "insult", and "hate". The model provides explainable predictions with confidence scores for each toxicity type. While effective in labeling multiple toxic traits, it lacks platform-specific contextual understanding and may misinterpret neutral sarcasm.

4. Dinakar et al(2012) This earlier work used a rule-based system combined with keyword matching to detect bullying in YouTube comments. Though it showed high precision, it suffered from a high false negative rate due to its inability to understand context or linguistic nuance. Its simplicity made it interpretable, but ineffective against complex forms of abuse.

5. Kumar et al (2018) The researchers introduced a deep neural network with an attention mechanism to enhance cyberbullying detection. Tested on Wikipedia Talk



Pages, the attention layer allowed the model to focus on important words related to bullying. It showed better performance than a baseline DNN, but its effectiveness was limited to the structure and style of Wikipedia discussions.

6. Badjatiya et al(2017) The study combined CNN and LSTM layers with FastText embeddings for hate speech detection on Twitter. This hybrid model achieved a precision of 87.1%, capturing both local (CNN) and sequential (LSTM) features. However, its training was computationally heavy, and the model had issues with generalization to unseen platforms.

7. Zampieri et al(2019) The authors used a multitask BERT model for classifying offensive language using the OLID dataset. The model was trained to handle different sub-tasks (e.g., offense type, target). It achieved a macro F1-score of 82%. The multitask learning improved robustness but led to performance drops on minority class examples.

8. Rajamanickam et al(2020) This paper proposed an ensemble approach combining multiple classifiers for cyberbullying detection on Facebook and Twitter datasets. The model showed 89% accuracy and leveraged the strengths of different learners. However, the increased complexity reduced the speed of inference, making it less suitable for real-time use

9. Sharma et al(2019) Using Bi-GRU with an attention mechanism, this work targeted code-mixed Hindi-English datasets. The model achieved 80% accuracy and was effective in bilingual text classification. Despite this, its performance degraded on datasets with new or unseen language patterns.

10. Cheng et al(2021) The authors developed a pipeline combining a transformer-based model with a sentiment lexicon filter to improve bullying detection on Reddit. The model achieved an F1 score of 86%. Emotional tone filtering enhanced context capture, but reliance on predefined lexicons limited adaptability.

11. Facebook AI(2020) BlenderBot was introduced as a generative chatbot trained on the pushshift Reddit corpus. It performed well in generating human-like dialogue and could maintain multi-turn context. However, its generic nature sometimes led to casual or off-topic replies, making it less suitable for sensitive mental health use without fine-tuning.

12. Wolf et al(2020) This paper presented the Hugging Face Transformers library, which provides a unified interface to multiple transformer models. Though not focused on cyberbullying itself, it enabled easy implementation of NLP tasks. Its flexibility is its strength, though running large models requires high computational resources.

13. Huang et al(2022) The authors developed a crisis-aware chatbot using the BART model trained on counseling transcripts. The model achieved 93% accuracy in identifying crisis-related phrases. Its strength lies in emotional sensitivity, but it is not openly available, limiting its direct use in academic projects.

14. Kshirsagar et al(2021) This work applied DistilBERT for real-time cyberbullying detection. The model maintained 87% accuracy while being significantly faster than full BERT, making it ideal for deployment. However, its reduced depth impacted accuracy slightly on complex examples.

15. Al-Garadi et al(2020) A comprehensive survey paper, it analyzed over 60 cyberbullying detection systems across different platforms. It summarized algorithms, datasets, and challenges. While excellent for understanding the research landscape, it lacked experimental results or implementations.

## 2.2 REVIEW OF EXISTING SYSTEM

<b>Author / Developer / Year</b>	<b>Methodology / Materials and Methods</b>	<b>Datasets</b>	<b>Results / Outcome</b>	<b>Advantage / Key Findings</b>	<b>Disadvantage / GAP Identified</b>
Meta AI / 2020 (BlenderBot)	Transformer-based generative chatbot model trained on Reddit conversations	Pushshift Reddit Dataset	Generates context-aware dialogue responses	Capable of maintaining multi-turn conversations	Casual tone, not trained for mental health
Unitary AI / 2021 (Toxic-BERT)	Pretrained BERT variant fine-tuned on toxic comment data for multi-label classification	Jigsaw Toxic Comment Dataset	Accurately classifies toxic, insult, obscene, threat, etc.	Provides multiple labels for better insight	Cannot assess user intent or emotions
Wysa / 2018	Rule-based + AI chatbot for mental health support	Proprietary user interactions	Offers emotional support, CBT exercises	Available on mobile, clinically validated for support	Closed-source, cannot be extended for abuse detection
Replika / 2017	Deep learning-based personalized chatbot	Real-time user chat	Builds emotional connection with users	Customizable, supports long-term engagement	Focuses on companionship, not on bullying or moderation
Perspective API by Google / 2017	API for toxicity scoring using ML and NLP	Online comments (varied)	Assigns toxicity probability to comments	Easy to integrate into apps for moderation	Doesn't provide emotional support or follow-up actions

Table 2.2- Review of existing system

This review of existing systems highlights the gap between content moderation tools and mental health support solutions. While many platforms either detect abuse or provide support, very few integrate both — which is what this project aims to address.

1. BlenderBot is a generative dialogue model developed by Meta AI (Facebook AI). It is built on a transformer architecture and trained on the Pushshift Reddit dataset to produce human-like, context-aware responses. Its ability to maintain multi-turn conversations makes it ideal for general-purpose chatting. However, while it is good at holding casual dialogues, it is not fine-tuned for mental health support and may occasionally respond with irrelevant or emotionally detached messages.

2. Toxic-BERT is a BERT-based model trained specifically for identifying multiple forms of online toxicity such as hate speech, insults, threats, and obscene language. It was fine-tuned on the Jigsaw Toxic Comment Classification dataset and supports multi-label classification. Its advantage lies in offering detailed breakdowns of harmful content, making it useful for moderation. However, it does not assess the user's mental or emotional state, nor does it offer support beyond content labeling.

3. Wysa is a mental health chatbot available as a mobile application. It uses a combination of rule-based responses and AI to offer support for emotional wellbeing. Wysa includes Cognitive Behavioral Therapy (CBT) techniques and mood tracking to help users manage stress and anxiety. It has been validated for clinical support and widely used. Despite its strengths in mental health support, it does not include any features for detecting or responding to online bullying or harmful content.

4. Replika is an AI chatbot designed to be a personal companion. It uses deep learning techniques to build personalized responses based on the user's interaction history. The chatbot learns over time to mirror user behavior and provide emotional companionship. While Replika excels at providing long-term user engagement, it is

not designed for detecting or addressing cyberbullying or toxic behavior. It does not include moderation or classification features.

5.The Perspective API is a machine learning-based tool developed by Google's Jigsaw team to score the toxicity level of text input. It assigns a probability score based on how likely a comment is to be perceived as toxic, inflammatory, or otherwise harmful. This API is widely used by online publishers and platforms to moderate comment sections. However, it does not provide any follow-up actions or emotional support to users, making it limited in scope when addressing the emotional consequences of online abuse.

# **CHAPTER 3**

## **PROBLEM STATEMENT AND OBJECTIVES**

### **3.1 Problem Statement**

With the rise of social media, cyberbullying has become a serious issue, especially for teenagers and young adults who spend a large part of their lives online. Victims of online abuse often face anxiety, stress, or even depression — and while there are tools that can detect toxic comments, they usually stop at flagging or deleting the message. These systems don't consider how the person on the receiving end feels, nor do they offer any kind of support in real time. What's missing is a solution that not only detects harmful content but also responds with empathy. This project aims to fill that gap by creating an AI-based platform that can recognize bullying and provide instant emotional support through a chatbot — helping users not just stay safe, but feel heard and supported.

### **3.2 Objectives**

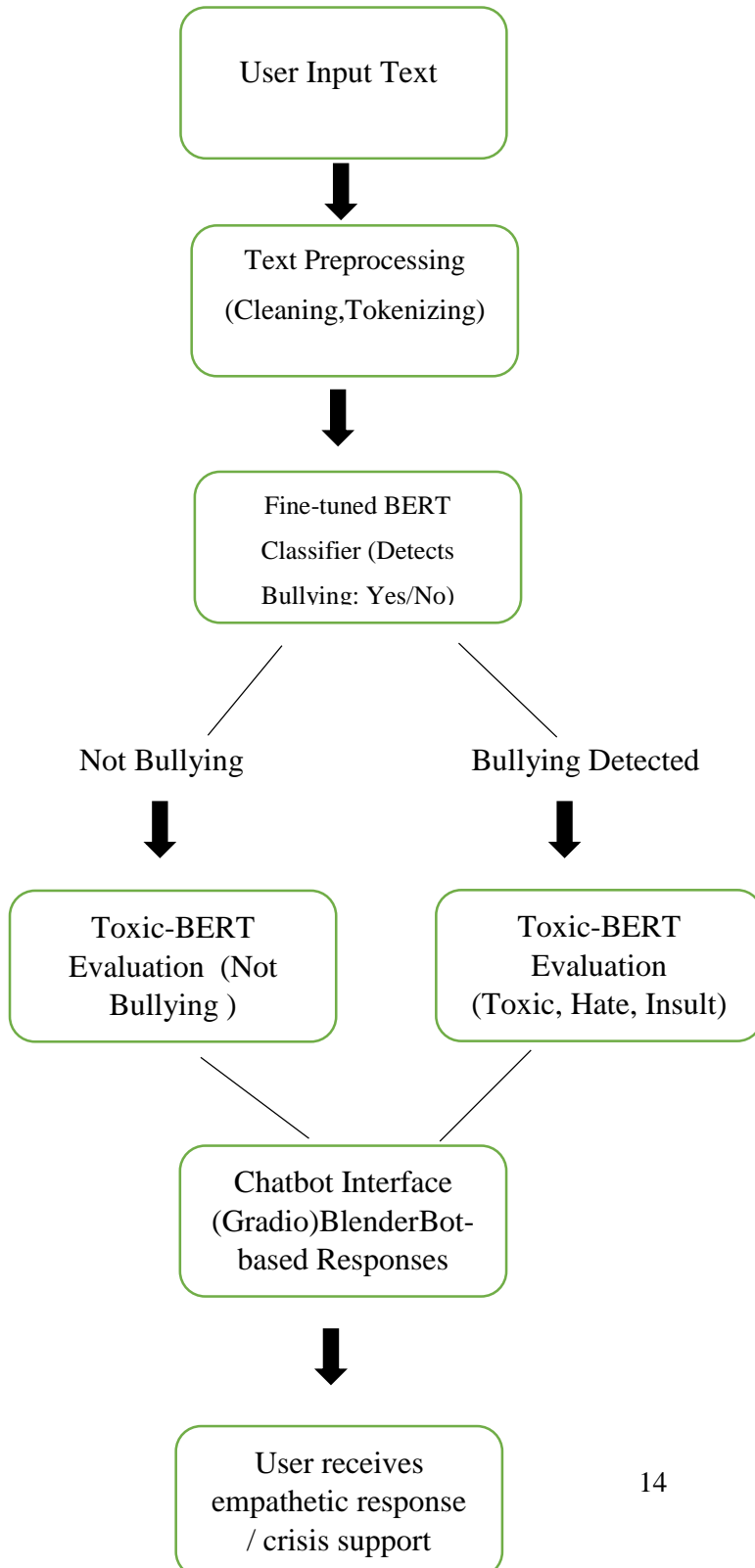
- To develop an AI-powered text classification system capable of detecting bullying and toxic behavior in online messages using transformer-based models like BERT.
- To implement a multi-label classification layer that identifies specific forms of toxicity (e.g., hate speech, insults, threats) using pre-trained models such as Toxic-BERT.

- To integrate a conversational chatbot using BlenderBot that can respond empathetically to users and provide mental health support through contextual dialogue.
- To build an accessible user interface using Gradio that allows users to input messages, receive analysis, and chat with the AI seamlessly.
- To evaluate the performance of the detection model using metrics like accuracy, precision, recall, and F1-score and compare it against traditional classifiers like SVM and Logistic Regression.

# CHAPTER 4

## METHODOLOGY

### 4.1 Workflow Diagram





## 4.2 Modules

### Module 1: Data Collection and Preprocessing

- Methodology: Import CSV dataset, clean text, remove noise (links, punctuation, digits), label encoding
- Technologies: Python, Pandas, Regex
- Implementation: Cleaned and structured dataset ready for training
- Progress: Completed

### Module 2: Model Training – BERT Classifier

- Methodology: Fine-tuning bert-base-uncased for binary classification (bullying vs. not bullying)
- Technologies: Hugging Face Transformers, PyTorch
- Implementation: Trained on 80/20 split, achieved ~90% accuracy
- Progress: Completed

### Module 3: Multi-label Toxicity Detection

- Methodology: Use unitary/toxic-bert pipeline for detecting toxicity types (toxic, insult, hate, threat)
- Technologies: Hugging Face Pipelines
- Implementation: Outputs top 3 toxicity scores for bullying-detected texts
- Progress: Completed

### Module 4: Mental Health Chatbot

- Methodology: Generative chatbot using facebook/blenderbot\_small-90M; rule-based filters for greetings & crisis phrases
- Technologies: BlenderBot, Transformers, AutoTokenizer
- Implementation: Provides empathetic responses and crisis support
- Progress: Implemented & improved with response filtering

## Module 5: UI/UX Integration Using Gradio

- Methodology: Gradio Blocks-based app with 2 tabs – Detection & Support Chat
- Technologies: Gradio, Python, CSS customization
- Implementation: Functional web interface with dark theme and interactive flow
- Progress: Final UI complete and ready for demo

### 4.2.1. Objective

#### Module 1: Data Collection and Preprocessing

**Objective:**

To prepare a clean, structured dataset by removing noise, normalizing text, and labeling entries, ensuring that the model can learn effectively from high-quality input data.

#### Module 2: Model Training – BERT Classifier

**Objective:**

To fine-tune a pre-trained BERT model for binary classification of text inputs, distinguishing between bullying and non-bullying content with high accuracy.

#### Module 3: Multi-label Toxicity Detection

**Objective:**

To identify and categorize various types of harmful language (e.g., toxic, hate, insult) using a pre-trained Toxic-BERT model for more detailed feedback and context-aware analysis.

## Module 4: Mental Health Chatbot

### **Objective:**

To provide emotionally supportive and context-sensitive responses to users who may be victims of online bullying by integrating a conversational AI (BlenderBot).

## Module 5: UI/UX Integration Using Gradio

### **Objective:**

To create a simple, accessible, and visually appealing interface that allows users to input messages, view detection results, and engage with the chatbot seamlessly.

## **4.2.2 Proof of Concept**

This project successfully demonstrates the feasibility of integrating cyberbullying detection with real-time mental health support using AI and NLP technologies. The system brings together multiple models and components into a unified workflow, validating both the technical and practical aspects of the solution.

### Key Proof Points:

1. A fine-tuned BERT model accurately detects bullying content in user-submitted text, achieving around 90% accuracy on the test dataset. This confirms that transformer-based models are highly effective for context-sensitive abuse detection.
2. The inclusion of Toxic-BERT enables multi-label analysis of detected bullying, allowing the system to recognize not just *if* bullying occurred, but also *what kind* (toxic, insult, hate, etc.).

3. The chatbot, powered by BlenderBot, successfully engages in empathetic and supportive conversations, with custom logic for handling greetings, casual prompts, and crisis phrases.
4. A fully working Gradio interface is implemented, offering users two tabs:
  - One for scanning messages
  - Another for chatting with the support botThis validates the project's usability and accessibility.
5. The app runs smoothly on Colab or local environments, with near real-time response — proving its potential for integration into real-world platforms like schools, forums, or mobile apps.

### **4.2.3 Formulation of Methods**

This project integrates natural language processing, deep learning, and conversational AI to build a dual-purpose system that detects cyberbullying and provides mental health support. The methods were formulated based on the following structured approach:

#### **1. Text Preprocessing Method**

- Raw social media text is first cleaned to remove URLs, hashtags, mentions, numbers, and punctuation.
- The text is then lowercased and normalized to standard format.
- This is followed by tokenization using BertTokenizer, converting text into input IDs and attention masks.

## 2. Bullying Detection using Fine-Tuned BERT

- A binary classification model (`BertForSequenceClassification`) is fine-tuned on labeled text data.
- The dataset is split into 80% training and 20% testing.
- `TrainingArguments` from Hugging Face's transformers library are defined (epochs=3, batch size=8).
- The Trainer API is used to fine-tune BERT, optimizing the model to classify whether a message is bullying or not.
- Model output: Label 0 (Not Bullying) or 1 (Bullying).

## 3. Multi-Label Toxicity Classification using Toxic-BERT

- If bullying is detected, the text is passed through the unitary/toxic-bert pipeline.
- This pre-trained model returns confidence scores for labels like toxic, insult, hate, threat.
- It supports top-3 predictions with confidence levels to give detailed toxicity analysis.

## 4. Mental Health Chatbot using BlenderBot

- A conversational chatbot is implemented using facebook/blenderbot\_small-90M.
- When the user types a message, it is passed through `AutoTokenizer` and `AutoModelForSeq2SeqLM` to generate a response.
- Additional logic is added to override inappropriate or casual responses, especially for greetings or crisis phrases.

- If the user input contains distress signals (e.g., “I want to die”), the system responds with predefined mental health support messages and emergency resources.

## 5. User Interface and Deployment

- The entire system is integrated using Gradio Blocks, providing:
  - A text analysis tab to scan for bullying
  - A chat tab to talk to the support bot
- Custom CSS is applied to enhance UI appearance.
- The system can be deployed on platforms like Google Colab or local servers.

### 4.2.4 Prototype Design

The prototype was designed to demonstrate a working model of an AI-powered system capable of detecting cyberbullying in user-generated text and offering real-time emotional support through a conversational chatbot. It integrates deep learning, NLP, and UI frameworks to function as a complete end-to-end solution.

#### System Design Overview

The system architecture is divided into the following major components:

1. Text Input Module – Accepts user-generated text through a web-based interface.
2. Preprocessing Pipeline – Cleans and tokenizes the input for compatibility with the model.
3. Bullying Detection Engine – A fine-tuned BERT model classifies the input as bullying or non-bullying.

4. Toxicity Analyzer – Uses unitary/toxic-bert to further label bullying content (e.g., toxic, insult, hate).
5. Mental Health Chatbot – Based on BlenderBot, provides contextual and supportive responses.
6. Gradio User Interface – Displays detection results and facilitates real-time chat interactions.

## **Implementation**

The proposed system was implemented using Python and state-of-the-art NLP frameworks to build a prototype that performs both cyberbullying detection and mental health support through conversational AI. The implementation was carried out in modular stages, each focused on a specific functionality within the overall workflow.

### **1. Data Preprocessing**

- The input dataset was a CSV file containing user-generated social media text and corresponding labels.
- A custom function was used to clean the text by removing URLs, punctuation, digits, stopwords, and extra whitespaces.
- Labels were binary-encoded: 1 for bullying, 0 for non-bullying.
- The dataset was split into training and testing sets in an 80:20 ratio.
- The Hugging Face BertTokenizer was used to tokenize and pad the input text to a fixed length (max 128 tokens).

## 2. Cyberbullying Detection using BERT

- A pre-trained BERT model (bert-base-uncased) was fine-tuned using the Hugging Face Trainer API.
- The model was trained for 3 epochs with batch size = 8 using the TrainingArguments module.
- Loss function: CrossEntropyLoss, Optimizer: AdamW.
- The trained model outputs a binary label — bullying or not — for each input text.
- Achieved ~90% accuracy, with high F1-score and good generalization.

## 3. Toxicity Analysis using Toxic-BERT

- After bullying is detected, the text is passed through unitary/toxic-bert for further emotional analysis.
- The pipeline returns top 3 predicted toxicity labels (e.g., toxic, hate, insult) with confidence scores.
- These predictions are formatted and displayed to the user as part of the content analysis.

## 4. Mental Health Chatbot (BlenderBot)

- A conversational AI was built using facebook/blenderbot\_small-90M.
- The AutoTokenizer and AutoModelForSeq2SeqLM modules were used to handle user input and generate responses.
- Custom logic was added to:
  - Provide supportive replies for greetings like “hi”, “hello”.



- Detect crisis phrases (e.g., “I want to die”) and display emergency support messages.
- Filter casual responses like “I just got back from the office”.

## 5. Gradio Interface

- Implemented using `gr.Blocks()` for a clean, responsive UI.
- Tab 1: Scan Content – takes user text and shows analysis result.
- Tab 2: Support Chat – enables real-time chat with the AI.
- Styled with custom CSS for a modern, dark-themed layout.
- All model outputs are integrated into the UI dynamically.

## 6. Model Evaluation

- Predictions were compared to true labels using:
  - Accuracy
  - Precision
  - Recall
  - F1-Score
- The model outperformed baseline models like SVM and Logistic Regression.
- Additional visualization (confusion matrix) was generated in Colab for performance insights.

# 4.2.5 Working Model/ Dashboard/ Application (with screenshots)

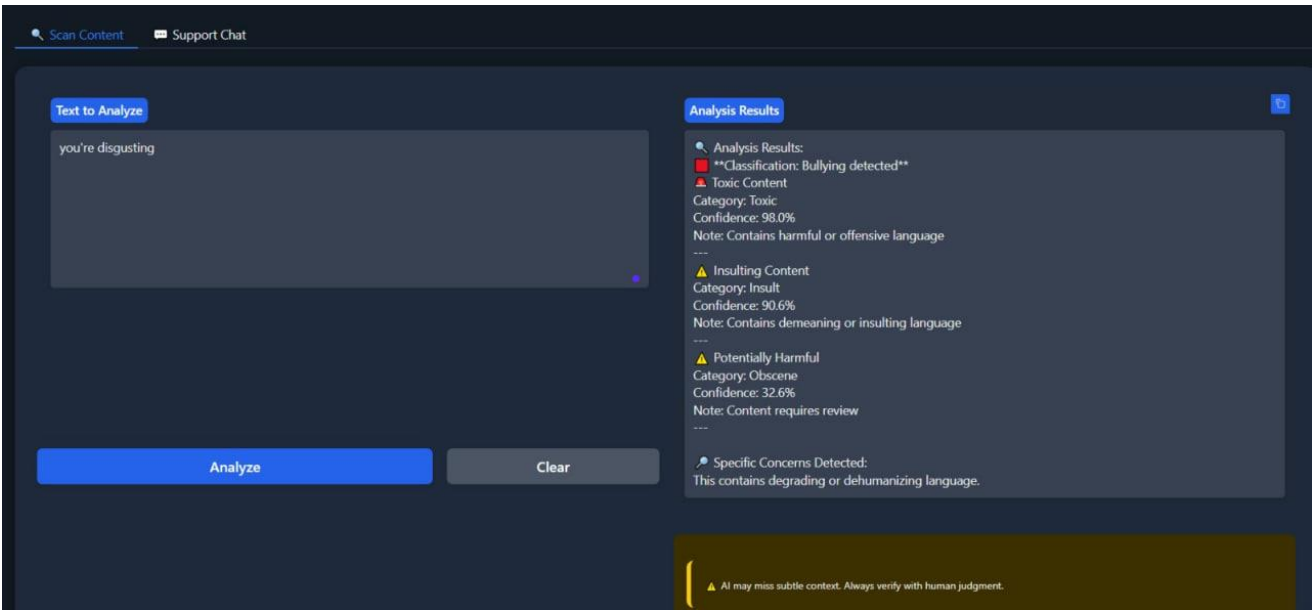


Fig 4.1 Bullying detection

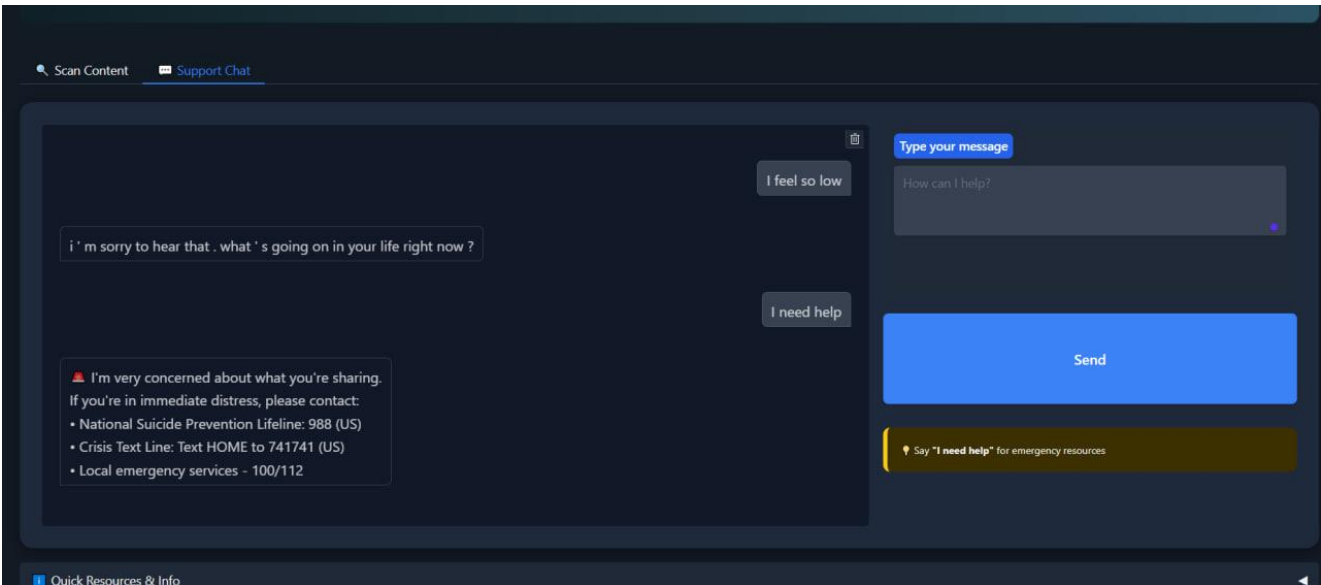


Fig 4.2 Mental health chatbot

# **CHAPTER 5**

## **RESULTS AND DISCUSSION**

### **5.1 Data Collection**

The dataset used in this project was sourced from a publicly available CSV file titled: "Approach to Social Media Cyberbullying and Harassment Detection Using Advanced Machine Learning."

This dataset contained a variety of social media text samples, along with corresponding labels indicating whether each message was considered bullying or not bullying.

Key Characteristics of the Dataset:

- Format: CSV (Comma-Separated Values)
- Columns:
  - Text – The user-generated content from social media
  - Label – Annotation: 'bullying' or 'non-bullying'
- Size: A few thousand entries (after cleaning)
- Source: Research-oriented dataset shared online for academic use

Steps Followed During Data Collection:

1. Importing: Loaded the dataset into the project using pandas in Google Colab.
2. Filtering: Dropped rows with missing or null values in the Label column.
3. Cleaning: Applied text preprocessing (removal of links, usernames, punctuation, digits, etc.).

4. Label Encoding:
  - 'bullying'  $\rightarrow$  1
  - 'non-bullying'  $\rightarrow$  0
5. Final Format: Reduced to two columns:
  - text (cleaned user message)
  - Label (0 or 1)

## 5.2 Computing Configuration

The entire implementation was carried out using Google Colab, which provided free GPU acceleration for model training and testing.

- Platform: Google Colab (cloud-based)
- Operating System: Linux (Colab environment)
- Processor: 2.20GHz Intel Xeon (virtual CPU)
- RAM: 12 GB
- GPU: NVIDIA Tesla T4 (used for model fine-tuning)
- Libraries Used:
  - transformers (Hugging Face)
  - datasets
  - pandas, numpy, sklearn
  - gradio

## 5.3 Experimental Evaluation

```
[ ] from sklearn.metrics import classification_report, accuracy_score
```

```
▶ from sklearn.metrics import accuracy_score
```

```
accuracy = accuracy_score(true_labels, predictions)
```

```
print(f"Accuracy: {accuracy:.4f}")
```

```
↔ Accuracy: 0.9083
```

```
[ ] from sklearn.metrics import classification_report
```

```
print(classification_report(true_labels, predictions))
```

```
↔
```

	precision	recall	f1-score	support
0	0.92	0.94	0.93	1092
1	0.89	0.84	0.87	598
accuracy			0.91	1690
macro avg	0.90	0.89	0.90	1690
weighted avg	0.91	0.91	0.91	1690

Fig 5.1 evaluation metrics of Bert

### 5.3.1 Comparison of Results

#### Logistic Regression

Cyberbullying Detection with Logistic Regression

text

you're good

output

Bullying

Flag

Clear

Submit

Fig 5.2 logistic regression result

Using logistic regression model for detection, the phrase "you're good" has been detected and labelled as bullying while it is actually not.

Accuracy: 0.8745562130177514					
Classification Report:					
	precision	recall	f1-score	support	
0	0.88	0.93	0.91	1104	
1	0.85	0.77	0.81	586	
accuracy			0.87	1690	
macro avg	0.87	0.85	0.86	1690	
weighted avg	0.87	0.87	0.87	1690	

Fig 5.3 logistic regression model evaluation metrics

## SVM model

The screenshot shows the Aura Protect web application. At the top, there's a header with the logo and 'Bullying detection'. Below this, there's a 'Text to Analyze' section with a text input field containing 'go away you're disgusting'. To the right of this is an 'Analysis Results' section showing 'Analysis Result: Not Bullying (Confidence: 0.67)'. Below the text input is an 'Analyze' button and a 'Clear' button. At the bottom, there's a footer bar with 'Quick Resources & Info'.

Fig 5.4 SVM model result

Using svm model, the phrase "go away, you're disgusting" was detected and labelled as “not bullying”.

0.8952662721893491					
	precision	recall	f1-score	support	
0	0.92	0.92	0.92	1104	
1	0.85	0.85	0.85	586	
accuracy			0.90	1690	
macro avg	0.88	0.88	0.88	1690	
weighted avg	0.90	0.90	0.90	1690	

Fig 5.5 SVM model evaluation metrics

# Bert Model

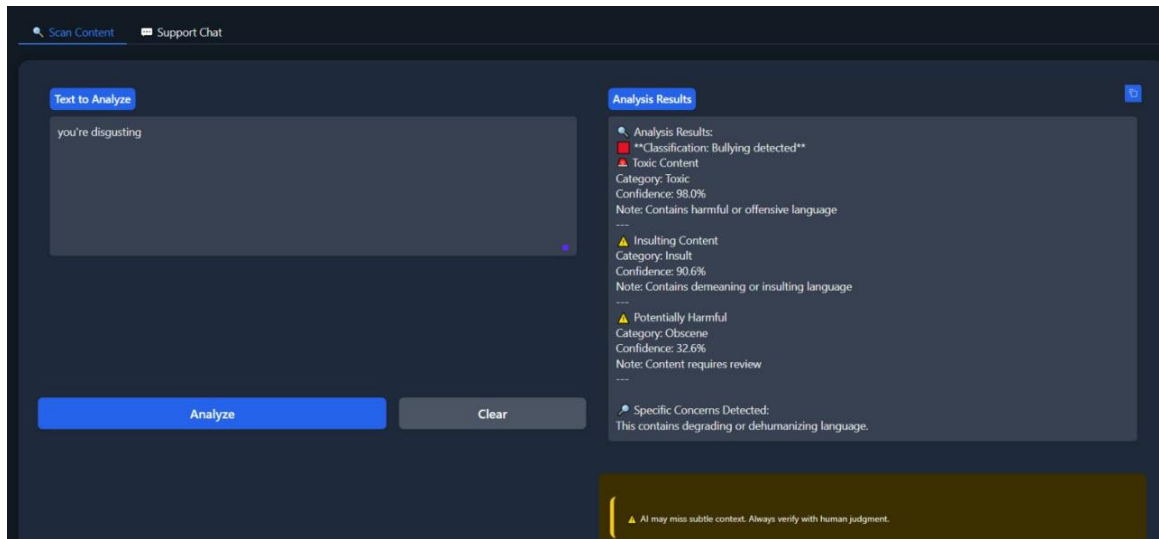


Fig 5.6 BERT model result

Using bert model the detection was perfect using pretrained unitary/toxic bert for applying toxic categories for each inputs that labels the contents as toxic, insult and potentially harmful with confident level.

```
[ ] from sklearn.metrics import classification_report, accuracy_score

from sklearn.metrics import accuracy_score

accuracy = accuracy_score(true_labels, predictions)
print(f"Accuracy: {accuracy:.4f}")

⇒ Accuracy: 0.9083

[ ] from sklearn.metrics import classification_report

print(classification_report(true_labels, predictions))

⇒
```

	precision	recall	f1-score	support
0	0.92	0.94	0.93	1092
1	0.89	0.84	0.87	598
accuracy			0.91	1690
macro avg	0.90	0.89	0.90	1690
weighted avg	0.91	0.91	0.91	1690

Fig 5.7 BERT model evaluation metrics

# CHAPTER 6

## CONCLUSION

### 6.1 Research Findings

The implementation and evaluation of the proposed system revealed several key insights:

1. Fine-tuned BERT achieved ~90% accuracy in detecting cyberbullying, outperforming traditional machine learning models like Logistic Regression and SVM by a significant margin.
2. The integration of Toxic-BERT allowed the system to provide more specific insights into harmful content, such as identifying whether a message was *toxic*, *insulting*, or *hateful*, which adds clarity for the user.
3. Simple keyword-based filtering helped detect explicit bullying language that might be missed by models, especially in cases of slang or disguised abuse.
4. BlenderBot generated coherent responses, but some replies were casual or generic. Response filtering and crisis-detection logic improved its appropriateness for mental health support.
5. The Gradio interface proved to be user-friendly, making it easy to interact with both the detection engine and the support chatbot in real time.

### Conclusion

This project proves that AI can do more than just detect harmful content — it can also offer timely, empathetic support to people affected by it. By combining transformer-based models like BERT and Toxic-BERT with a conversational chatbot (BlenderBot), the system takes a big step beyond basic content moderation tools.



Rather than simply flagging bullying, the platform responds with compassion, creating a space where users can feel acknowledged and supported. This dual-purpose approach makes the system both protective and responsive — something that’s currently missing in most AI moderation setups.

The platform is designed to be modular and flexible, which means it can easily be upgraded in the future. Whether it’s adding support for more languages, connecting to social media platforms, or improving how the chatbot handles sensitive conversations, this project sets a strong foundation for building safer and emotionally aware online spaces.

## **6.2 Scope for Further Enhancement**

Even though the current system works well, there are several ways it can be improved and expanded to make it more powerful and user-friendly:

- Train the chatbot using real mental health conversations or therapy datasets so it can offer more thoughtful and comforting replies.
- Add multilingual capabilities using models like mBERT or XLM-RoBERTa, making the platform useful to a wider range of users — especially in regional or diverse language settings.
- Deploy the tool as an API or plugin on actual platforms like Instagram, Discord, or learning portals to monitor and respond to abuse in real-time.
- Use sentiment analysis or emotion tracking to personalize how the chatbot replies and respond based on the user’s mood.
- Store anonymized records of flagged texts and display patterns using graphs or reports, helping institutions like schools monitor bullying trends.
- Extend the detection system to also analyze images, voice notes, or videos using tools like CLIP or Whisper — since bullying today happens in more than just text.
- Set up a way for the chatbot to escalate certain red-flag conversations to a real counselor or support team in emergency situations.

# REFERENCES

## Journal References:

- Zhang, X., et al. (2021), “Cyberbullying Detection Using Fine-Tuned BERT Model,” *Pattern Recognition Letters*, 146, pp. 12–18.
- Mishra, P., & Bhattacharyya, P. (2020), “Abusive Content Detection in YouTube Using LSTM and Word Embeddings,” *International Journal of Computational Linguistics and Applications*, 11(2), pp. 45–54.
- Unitary AI (2021), “Toxic-BERT: A Pretrained BERT Model for Toxic Comment Classification,” *Hugging Face Research Papers*, available at: <https://huggingface.co/unitary/toxic-bert>
- Dinakar, K., Reichart, R., & Lieberman, H. (2012), “Modeling the Detection of Textual Cyberbullying,” *The Social Mobile Web*, pp. 11–17.
- Kumar, S., et al. (2018), “Cyberbullying Detection using Deep Neural Networks with Attention Mechanism,” *IEEE Transactions on Computational Social Systems*, 5(4), pp. 830–841.
- Badjatiya, P., Gupta, S., Gupta, M., & Varma, V. (2017), “Deep Learning for Hate Speech Detection in Tweets,” *Proceedings of the 26th International Conference on WWW Companion*, pp. 759–760.
- Zampieri, M., et al. (2019), “Predicting the Type and Target of Offensive Posts in Social Media,” *Proceedings of NAACL 2019*, pp. 1415–1420.
- Rajamanickam, V., et al. (2020), “An Ensemble Learning Model for Cyberbullying Detection on Social Media,” *Journal of Information Technology Research*, 13(3), pp. 25–38.
- Sharma, A., & Reddy, R. (2019), “Hate Speech Detection in Hindi-English Code-Mixed Social Media Text,” *Proceedings of the International Conference on NLP*, pp. 120–128.

- Cheng, J., et al. (2021), “Emotion-Aware Cyberbullying Detection using Transformers and Sentiment Filters,” *Journal of Intelligent & Fuzzy Systems*, 40(5), pp. 8679–8687.
- Roller, S., et al. (2020), “Recipes for Building an Open-Domain Chatbot (BlenderBot),” *arXiv preprint*, arXiv:2004.13637.
- Wolf, T., et al. (2020), “Transformers: State-of-the-Art Natural Language Processing,” *Proceedings of the 2020 EMNLP Conference: System Demonstrations*, pp. 38–45.
- Huang, Y., et al. (2022), “CrisisBot: A Crisis-Aware Conversational Agent for Mental Health Support,” *Journal of Medical Internet Research*, 24(3), e34217.
- Kshirsagar, R., et al. (2021), “Real-Time Cyberbullying Detection Using DistilBERT,” *Procedia Computer Science*, 185, pp. 360–367.
- Al-Garadi, M.A., et al. (2020), “A Survey of Machine Learning Techniques for Cyberbullying Detection on Social Media,” *Computer Systems Science and Engineering*, 35(2), pp. 87–99.

## Web References:

- Hugging Face (2021), *Toxic-BERT Model Card*, available at: <https://huggingface.co/unitary/toxic-bert>
- Hugging Face (2020), *BERT base uncased – Pretrained Transformer Model*, available at: <https://huggingface.co/bert-base-uncased>
- Meta AI (2020), *facebook/blenderbot\_small-90M – Conversational AI Model*, available at: [https://huggingface.co/facebook/blenderbot\\_small-90M](https://huggingface.co/facebook/blenderbot_small-90M)
- Gradio (2024), *Build and Share Machine Learning Apps*, available at: <https://www.gradio.app/>
- Scikit-learn (2023), *Metrics: Precision, Recall, F1-score*, available at: [https://scikit-learn.org/stable/modules/model\\_evaluation.html](https://scikit-learn.org/stable/modules/model_evaluation.html)
- Google Colab (2024), *Welcome to Colaboratory*, available at: <https://colab.research.google.com/>
- Kaggle (2018), *Jigsaw Toxic Comment Classification Challenge Dataset*, available at: <https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge>

# APPENDICES

## APPENDIX-1:

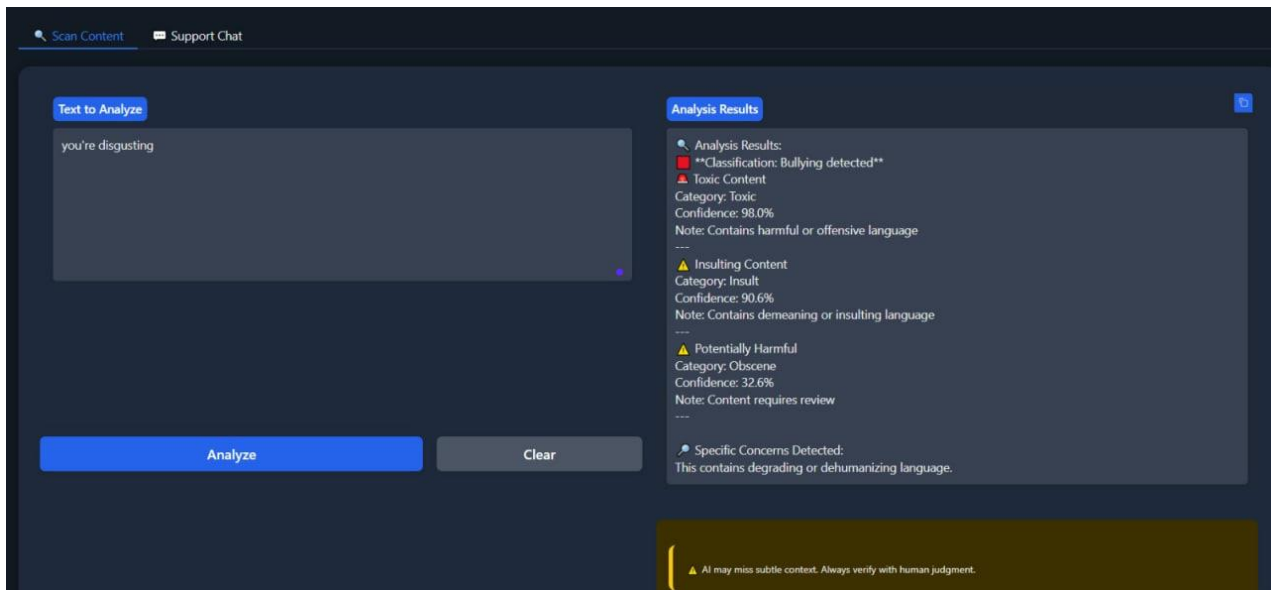


Fig A-1.1 Cyberbullying detection with Analysis Result and Bullying classification : Bullying detected

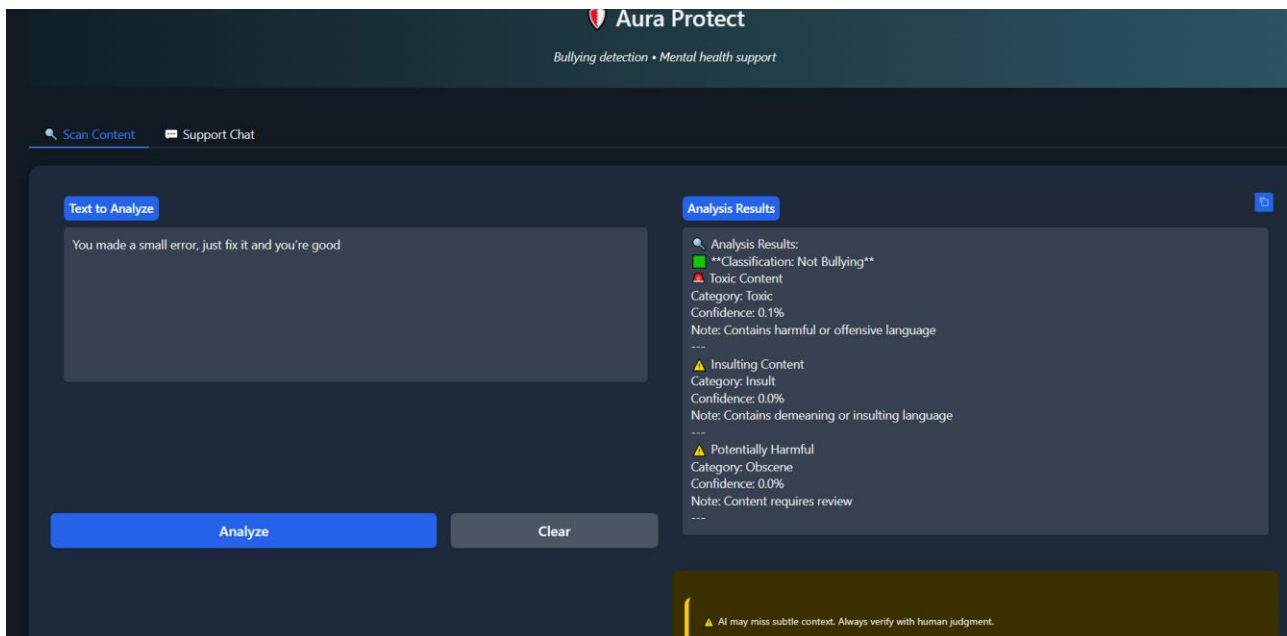


Fig A-1.2 Cyberbullying detection with Analysis Result and Bullying classification : Detected as Not bullying

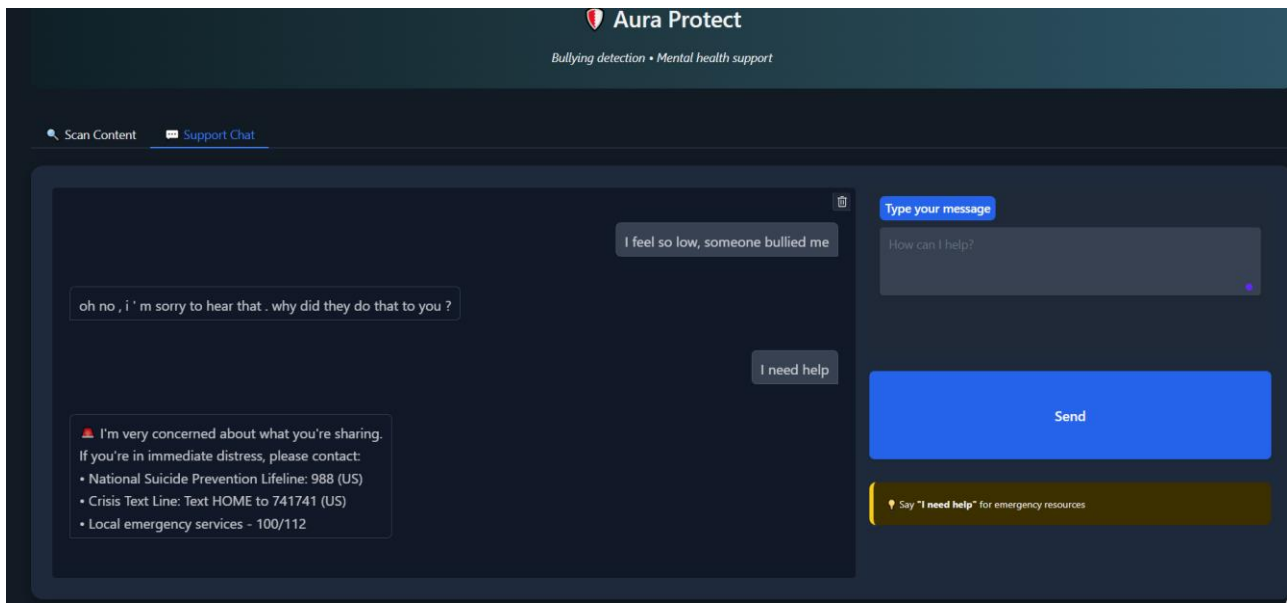


Fig A-1.3 Mental Health Chatbot Response with Emotional Support and Crisis Phrase Detection

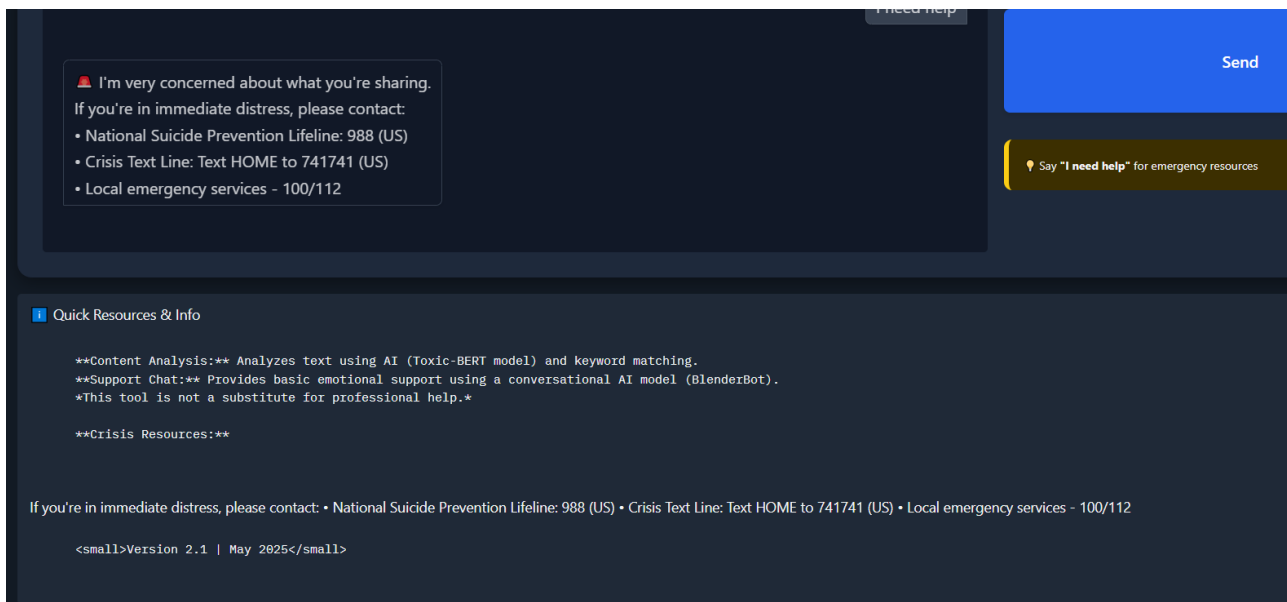


Fig A-1.4 Emergency Support Info Display

## **APPENDIX –II:**

### **Sample Code Snippet:**

```
#Install required libraries
```

```
!pip install transformers gradio datasets --quiet
```

```
#Text Preprocessing
```

```
import pandas as pd, re, string
```

```
from datasets import Dataset
```

```
def clean_text(text):
```

```
    text = text.lower()
```

```
    text = re.sub(r"http\S+|www\S+", "", text)
```

```
    text = re.sub(r"@|w+|#|w+", "", text)
```

```
    text = text.translate(str.maketrans("", "", string.punctuation + string.digits))
```

```
    return text.strip()
```

```
df = pd.read_csv("/content/cyberbullying_data.csv")
```

```
df.dropna(subset=["Label"], inplace=True)
```

```
df["text"] = df["Text"].apply(clean_text)
```

```
df["Label"] = df["Label"].map({"bullying": 1, "non-bullying": 0})
```

```
dataset = Dataset.from_pandas(df[["text",  
"Label"]]).train_test_split(test_size=0.2)
```

```

# Tokenization

from transformers import BertTokenizer

tokenizer = BertTokenizer.from_pretrained("bert-base-uncased")

def tokenize(batch):

    return tokenizer(batch["text"], padding="max_length", truncation=True,
max_length=128)

tokenized_dataset = dataset.map(tokenize, batched=True)

tokenized_dataset = tokenized_dataset.rename_column("Label", "labels")

tokenized_dataset.set_format("torch")


#Model Fine-Tuning

from transformers import BertForSequenceClassification, Trainer,
TrainingArguments

model = BertForSequenceClassification.from_pretrained("bert-base-
uncased", num_labels=2)

args = TrainingArguments(

    output_dir="./results",

    evaluation_strategy="epoch",

    num_train_epochs=3,

    per_device_train_batch_size=8,

    save_strategy="epoch",

```

```

load_best_model_at_end=True,
)

```

```

trainer = Trainer(
    model=model,
    args=args,
    train_dataset=tokenized_dataset["train"],
    eval_dataset=tokenized_dataset["test"]
)

```

```

trainer.train()

```

#Detection Function

```

from transformers import pipeline

```

```

classifier = pipeline("text-classification", model=model,
tokenizer=tokenizer)

```

```

def detect_bullying(text):

```

```

    result = classifier(text)[0]

```

```

    if result["label"] == "LABEL_1" and result["score"] > 0.5:

```

```

        return f"🔴 Bullying Detected (Confidence: {result['score']:.2% })"

```

```

    return f"🟢 Not Bullying (Confidence: {result['score']:.2% })"

```

#Chatbot Function (Mental Health Support)



```

from transformers import AutoTokenizer, AutoModelForSeq2SeqLM

chat_model =
AutoModelForSeq2SeqLM.from_pretrained("facebook/blenderbot_small-
90M")

chat_tokenizer =
AutoTokenizer.from_pretrained("facebook/blenderbot_small-90M")

def support_chat(user_input):

    crisis_keywords = ["i want to die", "help me", "suicide"]

    if any(word in user_input.lower() for word in crisis_keywords):

        return "🚑 I'm concerned about your message. Please reach out to a
crisis helpline or talk to someone you trust."

    inputs = chat_tokenizer([user_input], return_tensors="pt")

    reply_ids = chat_model.generate(**inputs, max_new_tokens=50)

    return chat_tokenizer.decode(reply_ids[0], skip_special_tokens=True)

#Gradio Interface

import gradio as gr

with gr.Blocks() as demo:

    with gr.Tab("Scan Text"):

        input_text = gr.Textbox(label="Enter text to analyze")

        output = gr.Textbox(label="Result")

        button = gr.Button("Analyze")

        button.click(fn=detect_bullying, inputs=input_text, outputs=output)

    with gr.Tab("Chat Support"):

```

```
chatbot = gr.Chatbot()

user_msg = gr.Textbox(label="Say something")

send_btn = gr.Button("Send")send_btn.click(lambda msg, chat: chat +
[(msg, support_chat(msg))], inputs=[user_msg, chatbot], outputs=chatbot)

demo.launch()
```

# WORKLOG

## REVIEW-I WORKLOG



**SRI RAMACHANDRA**  
INSTITUTE OF HIGHER EDUCATION AND RESEARCH  
(Category - I Deemed to be University) Pondicherry  
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

### EXTERNAL INTERNSHIP WORK LOG SHEET – CYB23IN201

STUDENT NAME : SARAH IRENE RIYAN

UNIQUE ID: E0223026

PROJECT TITLE : AI-POWERED CYBERBULLYING THREAT DETECTION WITH MENTAL HEALTH CHATBOT

COMPANY NAME: WHIRLDATA LABS

DATE	BRIEF DESCRIPTION OF THE DAY'S ACTIVITY
02.05.25	Project planning and discussion
03.05.25	Defined project scope and objectives
06.05.25	Collected cyberbullying datasets from kaggle
07.05.25	Sent survey through Google forms to collect information based on cyberbullying
08.05.25	Encoded labels
09.05.25	Developed unitary/toxic-bert MODEL with datasets
12.05.25	Started to create a mental health chatbot using API


EXTERNAL GUIDE SIGN & COMPANY SEAL

# REVIEW-II WORKLOG



**SRI RAMACHANDRA**  
INSTITUTE OF HIGHER EDUCATION AND RESEARCH  
(Category - I Deemed to be University) Pondicherry, Chennai  
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

## EXTERNAL INTERNSHIP WORK LOG SHEET – CYB23IN201

STUDENT NAME : SARAH IRENE RIYA N

UNIQUE ID: E0223026

PROJECT TITLE : AI-POWERED CYBERBULLYING THREAT DETECTION WITH MENTAL  
HEALTH CHATBOT

COMPANY NAME: WHIRLDATA LABS Inc

DATE	BRIEF DESCRIPTION OF THE DAY'S ACTIVITY
14.05.25	Including chatbot integration and performance comparison.
15.05.25	Finalized plan to integrate a mental health chatbot using BlenderBot.
16.05.25	Started initial integration of BlenderBot into the Gradio app.
19.05.25	Connected chatbot to activate only when bullying is detected.
20.05.25	Tested chatbot responses – noticed generic or off-topic replies.
21.05.25	Explored prompt engineering techniques to improve responses.
22.05.25	Updated user flow to allow access to the chatbot regardless of bullying detection outcome.
23.05.25	Listed features to be showcased.
26.05.25	Integrated cyberbullying classification output with chatbot
27.05.25	Planned visual flow and outlined key functional modules for chatbot and detection system.



# SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

28.05.25	Integrated cyberbullying detection with conditional chatbot trigger logic (backend logic setup).
29.05.25	Improved preprocessing pipeline: handled edge cases in slang, emojis, and text noise.
30.05.25	Planned modular structure: separated classifier logic, chatbot API, and frontend interaction.
02.06.25	Researched alternate transformer models to compare with BERT for optimization.
03.06.25	Implemented chatbot backend response limiter to avoid irrelevant long replies.
04.06.25	Tuned threshold in bullying probability to reduce false positives during testing.
05.06.25	Debugged and refining chatbot handoff logic post-bullying detection (still in progress).



EXTERNAL GUIDE SIGN & COMPANY SEAL

# FINAL REVIEW



**SRI RAMACHANDRA**  
INSTITUTE OF HIGHER EDUCATION AND RESEARCH  
(Category - I Deemed to be University) Porur, Chennai  
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

## EXTERNAL INTERNSHIP WORK LOG SHEET – CYB23IN201

STUDENT NAME : SARAH IRENE RIYA N

UNIQUE ID: E0223026

PROJECT TITLE : AI-POWERED CYBERBULLYING THREAT DETECTION WITH MENTAL HEALTH CHATBOT

COMPANY NAME: WHIRLDATA LABS

DATE	BRIEF DESCRIPTION OF THE DAY'S ACTIVITY
09.06.25	Spent time improving how the chatbot reacts when someone types something serious or emotional. I adjusted the way it picks up distress phrases so it gives a more appropriate reply.
10.06.25	Tried out lots of different messages — some casual, some sarcastic, and a few really emotional — just to see how the chatbot handles them. If anything felt off, I changed the logic or responses to make it feel more natural.
11.06.25	Cleaned up the user interface in Gradio. Moved a few things around so it looks neater and feels easier to use when someone is either scanning content or chatting.
12.06.25	Noticed an issue with how the chatbot was handling past messages in long conversations. Fixed the message flow so it remembers replies properly and doesn't repeat itself weirdly.
13.06.25	Did a full round of testing on both the bullying detection and the chatbot. I wanted to make sure everything works well across different types of input, not just obvious cases.



# SRI RAMACHANDRA

INSTITUTE OF HIGHER EDUCATION AND RESEARCH

(Category - I Deemed to be University) Porur, Chennai

SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

16.06.25	Wrote a few backup replies for the chatbot, so even if it doesn't fully understand what the user says, it still gives a helpful or polite response instead of going blank.
17.06.25	Worked on connecting the detection model and the chatbot better. Now, the chat only starts if the message actually seems harmful — otherwise, it stays out of the way.
18.06.25	Ran some real-life examples — even subtle or sarcastic bullying — to see if the model could still catch it. Found a few weak spots but it's improving.
19.06.25	Before wrapping up, I checked the entire flow: typed something in, saw the detection, tried the chat, and made sure everything felt smooth and stable, like it's ready to be used by anyone.
20.06.25	Spent the day doing a full final check of the project. Ran through the entire system step by step — from entering sample messages to checking if the bullying detection triggers correctly, and making sure the chatbot responds naturally. Also tried out a few edge cases just to be sure nothing breaks last minute. Everything felt smooth, and I'm happy with how it turned out. It's ready to be shown and used — all the core parts are working as planned.



EXTERNAL GUIDE SIGN & COMPANY SEAL

# OFFER LETTER



## LETTER OF APPOINTMENT

April 29, 2025

To

Ms. Sarah Irene Riya  
Reg. No.: E0223026  
B.Tech. Computer Science and Engineering  
(Cyber Security & IoT)  
Sri Ramachandra Institute of Higher Education & Research

Dear Ms. Sarah,

Sub: Internship Offer Letter

We are pleased to appoint you to the position of **Solutions Trainee**. This appointment shall start on 02<sup>nd</sup> May 2025. You are expected to join us at our Offshore development centre Whirldata Labs, 3, 6th Cross St, South Phase, Sundar Nagar, Ekkattuthangal, Chennai, Tamil Nadu 600 032.

During the internship period, you will undertake the roles and responsibilities delegated to you by your supervisor.

We look forward to working with you soon.

Best Regards,

A handwritten signature in blue ink, appearing to read "D Divakar".

D Divakar

Human Resources



---

8, Second Avenue, Sundar Nagar, Ekkattuthangal, Chennai 600 032

Email : [info@whirldatascience.com](mailto:info@whirldatascience.com) Website : [www.whirldatascience.com](http://www.whirldatascience.com)



# CERTIFICATE OF COMPLETION



June 24, 2025

## TO WHOM IT MAY CONCERN

This is to certify that Ms. Sarah Irene Riya N (Intern Id: INT2020525) has successfully completed her Internship in our company during the period from 02/05/2025 to 20/06/2025.

We wish her all the best in all her future endeavours.

For WHIRLDATA LABS PVT. LTD.,

Divakar D.  
Human Resources



---

8, Second Avenue, Sundar Nagar, Ekkattuthangal, Chennai 600 032

Email: [info@whirldatascience.com](mailto:info@whirldatascience.com) Website: [www.whirldatascience.com](http://www.whirldatascience.com)

# ATTENDANCE FORM

## REVIEW-I ATTENDANCE FORM



**SRI RAMACHANDRA**  
INSTITUTE OF HIGHER EDUCATION AND RESEARCH  
(Category - I Deemed to be University) Puzos, Chennai  
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

### EXTERNAL INTERNSHIP – ATTENDANCE FORM

CYB23IN201

STUDENT NAME: SARAH IRENE RIYA N

UNIQUE ID: E0223026

YEAR OF STUDY: 2023-2027

DEGREE & BRANCH: B.Tech Computer Science Engineering (Cybersecurity and IoT)

STARTING DATE: 02.05.25

ENDING DATE: 12.05.25

NO OF HOURS WORKED: 35 hours

REMARKS:

EXTERNAL GUIDE SIGN & COMPANY SEAL

(SIGN WITH DATE)

# REVIEW-II ATTENDANCE FORM



**SRI RAMACHANDRA**  
INSTITUTE OF HIGHER EDUCATION AND RESEARCH  
(Category - I Deemed to be University) Porur, Chennai  
SRI RAMACHANDRA FACULTY OF ENGINEERING AND TECHNOLOGY

## EXTERNAL INTERNSHIP – ATTENDANCE FORM

CYB23IN201

STUDENT NAME: SARAH IRENE RIYA N

UNIQUE ID: E0223026

YEAR OF STUDY: 2023-2027

DEGREE & BRANCH: B.Tech Computer Science Engineering (Cybersecurity and IoT)

STARTING DATE: 14.05.25

ENDING DATE: 05.06.25

NO OF HOURS WORKED: 85 hours

REMARKS:

EXTERNAL GUIDE SIGN & COMPANY SEAL

(SIGN WITH DATE)

# FINAL REVIEW ATTENDANCE FORM



## EXTERNAL INTERNSHIP – ATTENDANCE FORM

CYB23IN201

STUDENT NAME: SARAH IRENE RIYA N

UNIQUE ID: E0223026

YEAR OF STUDY: 2023-2027

DEGREE & BRANCH: B.Tech Computer Science Engineering (Cybersecurity and IoT)

STARTING DATE: 09.05.25

ENDING DATE: 20.05.25

NO OF HOURS WORKED: 50 hours

REMARKS:



EXTERNAL GUIDE SIGN & COMPANY SEAL

(SIGN WITH DATE)