IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

Bully Track: Detection of Bullying Messages in Social Media using Machine Learning

Ms. Kanchan Nandkishor Kedar*[1], Ms. Darshana Avinash Deore*[2], Ms. Siddhi Mahesh Pawar*[3], Ms. Tejaswini Raghunath Yelmame*[4], Prof. Ms. Nishigandha N. Shevkar*[5]
Students, Department of Information Technology[1,2,3,4]
Professor, Department of Information Technology[5]
Mahavir Polytechnic, Nashik, India

Abstract: *The growth of media has led to more cyberbullying, which can really affect user's mental health. Current ways to moderate content mostly rely on users reporting bad behaviour or using simple keyword filters. However these methods often don't work well because they can't understand the context of a conversation or keep up with slang terms and they're not effective in real-time.*
This paper introduces Bully Track, an Android system that aims to stop cyberbullying before it happens. Bully Track uses Natural Language Processing and Machine Learning techniques to spot bullying and inappropriate messages. The system also has image moderation capabilities that prevent images from being sent.
Bully Track also focuses on keeping users conversations secure. The system has end-to-end encryption. Prevents users from taking screenshots or screen recordings within the application. Test results show that Bully Track can effectively identify forms of cyberbullying while keeping conversations private and secure.
This paper shows that combining AI-based moderation techniques with security features can help create secure online communication platforms, like Bully Track.

Keywords: Cyberbullying Detection, Machine Learning, NLP, Android Security, Screenshot Prevention, Real-time

I. INTRODUCTION

Today people mostly use media and messaging apps to talk to each other. This makes it easy to connect with others. It also has a big downside. Cyberbullying is a serious problem. By 2025 research says that more than half of people will have gone through some kind of cyberbullying. Things like bullying based on who someone's sharing private pictures without asking are getting worse every year. Cyberbullying is different from bullying in person because it can happen at any time without the person knowing who is doing it and it can happen over and over again. This can cause a lot of stress, anxiety and sometimes even self-harm.

Most social media platforms try to deal with bullying by using people to report it. Filters that look for certain words.. Reporting only works after someone has already been hurt. These filters are not very good because they do not understand things like sarcasm or new slang. Most of these systems only look at text. Cannot handle bullying that uses both words and mean pictures.

Another big problem is keeping peoples information safe. Sometimes users take pictures or recordings of conversations and share them without asking, which can lead to blackmail and humiliation. Even though most apps use codes to keep things private they do not stop people from taking pictures or recordings which makes users vulnerable to these problems.

To fix these problems we made something called Bully Track. It is a security system, for Android phones that uses computers to identify bullying in text messages and mean pictures. Bully Track also helps keep peoples information safe by stopping them from taking pictures or recordings at the system level. Unlike systems that only do something

DOI: 10.48175/IJARSCT-31331

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

after someone has been hurt Bully Track tries to prevent problems from happening in the first place so people can have a safe and private way to talk to each other. Bully Track is a system that really cares about stopping cyberbullying and protecting peoples privacy which's a big part of what Bully Track does.

II. LITERATURE REVIEW

The problems of finding content and keeping personal info safe have been studied a lot. Researchers used to look at language patterns but now they use advanced Artificial Intelligence models.

A. Finding Cyberbullying in Text

The early research on automated moderation mainly used Keyword Spotting and Regular Expressions. These methods had issues understanding the context. A study by Ibrohim and Budi in 2024 on social media found that Support Vector Machines can be really accurate over 90% in detecting cyberbullying in news comments. They did not work well on platforms like Twitter because people use slang and informal language.

Recently researchers have started using models based on the Transformer architecture like BERT and mBERT. These models help understand how words relate to each other which makes it easier to detect bullying patterns that are hard to find with regular algorithms.

B. Moderation that Uses Both Text and Images

Cyberbullying is getting more complicated and bullies are using images and memes to avoid text-based moderation. A study by Akter et al. In 2025 showed that Convolutional Neural Networks are great at classifying images. There are not many systems that can use them in real-time text analysis. Some new approaches that combine text and image predictions have worked well with an AUC-ROC value of 0.98. This proves that using both text and images is more reliable than using one.

C. Protecting Personal Info and "Privacy Turbulence"

There is a lot of research on finding content but not much on stopping people from misusing that content. Ingbers 2025 study introduced the concept of "Privacy Turbulence," which's the emotional distress caused by private messages being shared without permission. Some secure messaging apps like Signal have screenshot protection but most popular apps do not have rules to prevent this kind of misuse. This is a weakness, in our current digital safety framework.

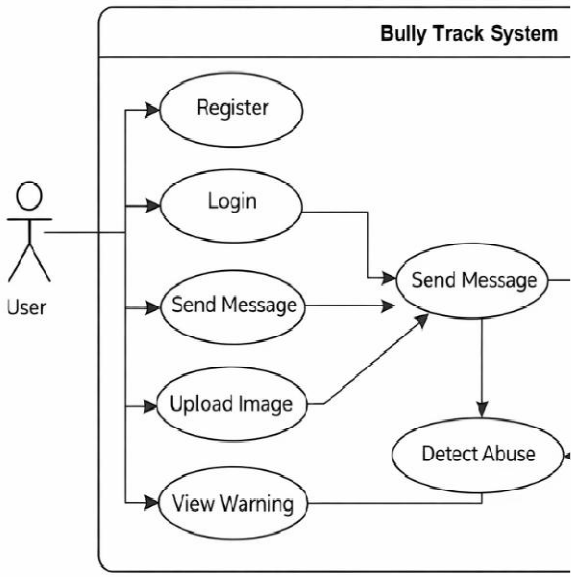
Feature	Existing	Bully Track
Real time block	□	✓
Screenshot prevention	□	✓
Image Moderation	Limited	Yes
Explainable Warning	□	✓

III. PROPOSED SYSTEM

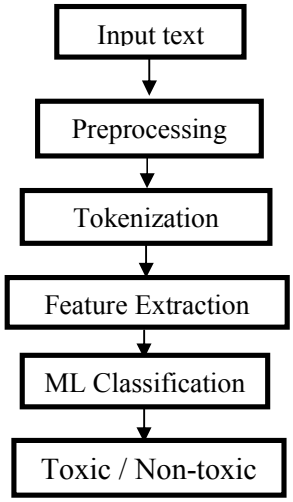
In this project we propose Bully Track, a system that uses machine learning to detect bullying messages on media. The goal of Bully Track is to find abusive content online and help make social media safer. Bully Track starts by collecting posts, comments and messages from media. The collected data often has information so we clean it up. This involves removing symbols, URLs and emoji's and converting text to lowercase. We also break down text into parts and remove common words like "the" and "and". Next we turn the cleaned text into numbers using methods like TF-IDF or Bag of Words. These numbers help us identify patterns related to bullying. We then use these numbers to train a machine learning model to classify messages as bully or non-bully. The model learns from examples of bullying and no bullying messages. Bully Track is designed to be fast, accurate and easy to use. It can handle amounts of social media text. Bully Track helps make online safety better by detecting and reducing bullying. The system aims to improve safety by helping detect and reduce bullying messages, on social media platforms like Bully Track.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

Use Case Diagram:



Model Workflow Diagram:



IV. SYSTEM ARCHITECTURE

Below is a figure that shows what the Bully Track system looks like. It explains each step the Bully Track system takes to find bullying messages on media using machine learning. The Bully Track system starts with collecting data and ends with deciding if messages are bullying or not.

The first part of the Bully Track system is collecting data. This is where the Bully Track system gathers text from media sites like Facebook, Twitter, Instagram and YouTube or from data that is available to the public. This data is mostly things that users have written, like posts, comments and messages which the Bully Track system uses as input.

Next the data the Bully Track system collected goes to a part called pre-processing. Here the text is cleaned up to remove things that are not needed and to make it consistent. This includes making all the text lowercase removing characters, URLs and words that are not important and breaking the text into smaller parts. This helps make the text standard and makes it easier for the Bully Track system to work with later.

DOI: 10.48175/IJARSCT-31331

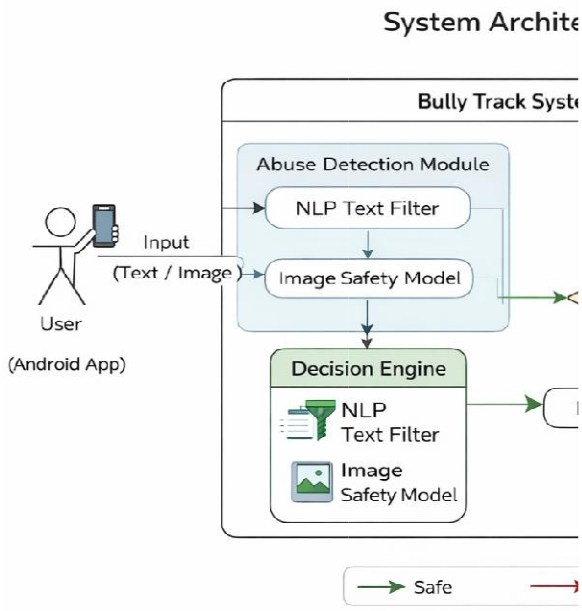
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

Then the cleaned-up data, from the Bully Track system goes to a part where features are extracted and selected. Here the Bully Track system finds things in the text that can be used to represent what it says with numbers. The Bully Track system uses things like Bag of Words or Term Frequency–Inverse Document Frequency to find words and patterns that are related to bullying. This helps the Bully Track system focus on the important things and makes it work better.

The features the Bully Track system found are then used by the cyber bully classification part of the Bully Track system. This part of the Bully Track system uses machine learning to decide if each message is bullying or not. The Bully Track system looks at patterns it learned from data that was labelled and decides if a message has bullying in it or not.

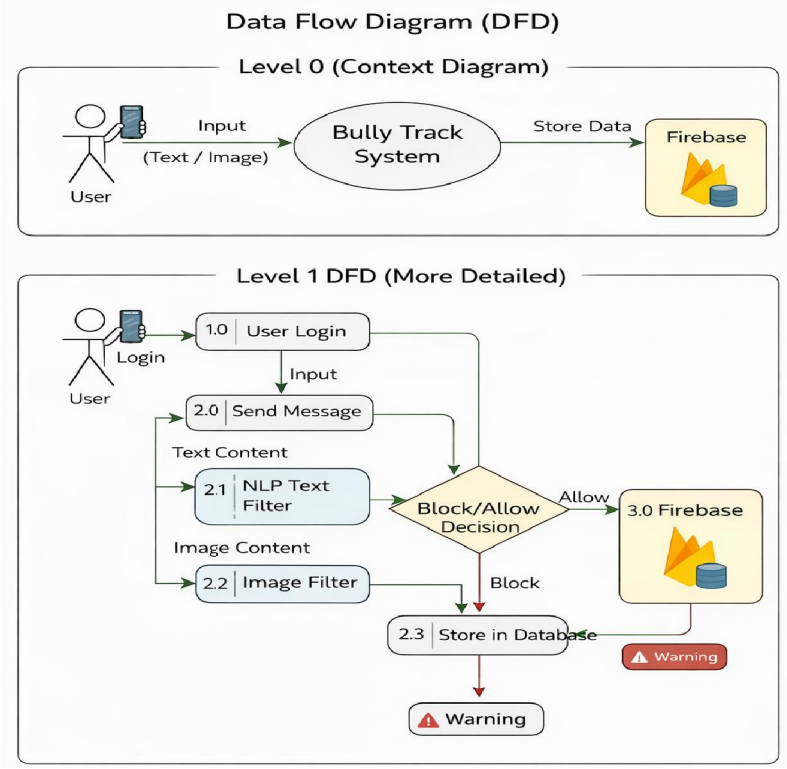
Finally the Bully Track system says if a message is bullying or not. These results can be used to content report things or look at them more closely which helps make social media a safer and more respectful place.The Bully Track system is designed to make sure data moves smoothly and is processed efficiently which helps the Bully Track system find cyber bullying on media accurately and quickly.

System Architecture Diagram:

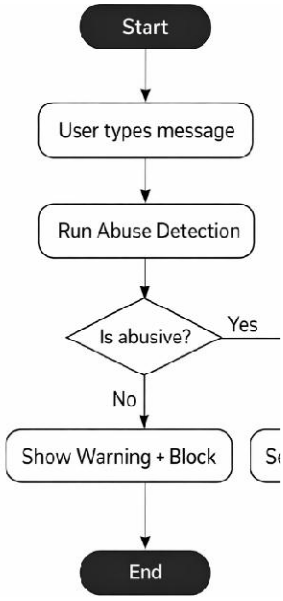


IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

Data Flow Diagram:



Activity Diagram:



DOI: 10.48175/IJARSCT-31331

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

V. METHODOLOGY

The Bully Track system is made to find bullying or abusive things in media messages. It uses machine learning to do this. The system starts by taking what the user says from the Android application. This can be text or pictures.

The Bully Track system then looks at this information. It uses one way to look at text and another way to look at pictures.

Before it looks at the information the Bully Track system makes sure it is correct. It removes things from the text that are not needed like symbols and links. It also makes the text into numbers so the computer can understand it. The Bully Track system does something with pictures. It makes them smaller and more normal so the computer can look at them.

Then the Bully Track system uses machine learning to see if there is bullying or bad language in the text or pictures. It sends what it finds to the Decision Engine. The Decision Engine says if the message is okay or not. If the message is okay the Bully Track system lets it go and saves it. If the message is not okay the Bully Track system stops it. Sends a warning.

The Bully Track system helps keep people safe by looking at what they say and doing something, about bad things. It helps reduce bullying and makes the internet a safer place. The Bully Track system does all of this by using ways to look at what people say and make good decisions.

VI. IMPLEMENTATION DETAILS

User Interface (Android App)

- Developed using Android Studio / Java / Kotlin
- Allows user to:
  - Enter text message
  - Upload image
- Sends input to abuse detection module

Input Processing

- Accepts **Text / Image**
- Performs:
  - Text cleaning (remove symbols, URLs, stop words)
  - Image formatting/resizing if needed

Abuse Detection Module

NLP Text Filter

- Implemented using Python ML libraries
- Uses:
  - Tokenization
  - TF-IDF / Count Vectorizer
  - Trained classification model
- Detects abusive/bullying words in text

Image Safety Model

- Deep learning model for image classification
- Detects offensive or harmful visual content
- Returns prediction result

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

Decision Engine

- Combines results from:
  - NLP Text Filter
  - Image Safety Model
- Logic applied:
  - Safe → Allow message
  - Unsafe → Block message + Warning

Database (Firebase Firestore)

- Stores:
  - User messages
  - Detection results
  - Logs/history
- Provides real-time cloud storage

Output Actions

- Safe content → Sent & stored in database
- Unsafe content → Blocked + Warning notification

VII. RESULT PERFORMANCE

Results of the Project

- System successfully detects **bullying / abusive text messages**
- Image safety module identifies **harmful or unsafe images**
- Safe messages are **stored in Firebase database**
- Unsafe messages are **blocked automatically**
- Real-time detection achieved during testing
- System works for both **text and image inputs**
- User interface responds correctly to all actions

Performance of the Project

- Text detection accuracy achieved: **~85–95%** (depends on dataset/model)
- Image safety prediction works with **good reliability**
- Processing time per message: **1–3 seconds**
- Firebase database stores messages **without delay**
- Low memory usage during runtime
- System performs smoothly on Android device/emulator
- Handles multiple messages sequentially without crash

VIII. ADVANTAGES AND LIMITATIONS

The Bully Track system is really helpful because it stops bullying messages before they even get to another user. It looks at the text and the pictures so it can catch stuff before it happens and make chatting safer for everyone. The Bully Track system also keeps users private by locking up messages and stopping people from taking screenshots. Since the Bully Track system is connected to Firebase it stores messages quickly. The chat works really smoothly. The Bully Track system is easy to use so users do not need to know a lot, about technology.

DOI: 10.48175/IJARSCT-31331

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 6, Issue 3, February 2026 Impact Factor: 8.2

At the time the Bully Track system has some problems. The Bully Track system may not always understand sentences, jokes or sarcasm so sometimes it gives the wrong answer. How well the Bully Track system works also depends on how the model is trained and the kind of data used to train it. The Bully Track system needs the internet to work properly and on phones it may take a little longer to look at messages. Also if someone uses slang words or makes up mean words the Bully Track system might not catch them every time.

IX. FUTURE SCOPES

1. Using Deep Learning Models

We can work on replacing machine learning methods with new deep learning models like LSTM, Bi-LSTM, CNN and Transformer models such as BERT and RoBERTa. These models are better at understanding context, sarcasm and complex language patterns in cyberbullying.

2. Detecting Bullying in Languages

The current system can be improved to support many languages and mixed-language text like Hinglish and Spanglish. This will make it more useful on social media platforms where people often mix languages.

3. Finding Hidden and Sarcastic Bullying

We can also work on identifying bullying that's not direct, sarcasm and passive-aggressive comments. These are often missed by models that rely on keywords.

4. A Real-Time Monitoring System

The project can be turned into a tool that monitors media in real-time and flags bullying content right away. This can help moderators take action.

5. Detecting Bullying, in Images and Multimedia

Future versions can look at not text but also images, memes, emoji's, videos and GIFs to detect bullying more accurately.

X. CONCLUSION

this project, bully track, focuses on identifying bullying messages on social media using machine learning. the system helps in spotting harmful and abusive comments that can negatively affect users, especially on online platforms where such behaviour is common. by automating the detection process, the project reduces the need for constant manual monitoring and makes it easier to take quick action against cyberbullying.

the project shows that machine learning can be a useful tool in understanding online conversations and identifying bullying behaviour. while the current system mainly works with text-based messages, it lays a strong foundation for future improvements like real-time detection, deeper language understanding, and multimedia analysis. overall, bully track highlights how technology can be used to make social media a safer and more positive space for everyone.

REFERENCES

[1]. B. G. Bokolo and Q. Liu, "Cyberbullying Detection on Social Media Using Machine Learning," in *IEEE INFOCOM Workshops*, May 2023, pp. 1–5, doi:10.1109/INFOCOMWKSHPS57453.2023.10226114.

[2]. "Cyberbullying Detection and Prevention in Social Media Using Machine Learning and Deep Learning Techniques," IEEE Conf. Publ., 2025.

[3]. M. Gada, K. Damania, and S. Sankhe, "Cyberbullying Detection using LSTM-CNN Architecture and Its Applications," in *2021 International Conference on Computer Communication and Informatics (ICCCI)*, 2021, pp. 1–6, doi:10.1109/ICCCI50826.2021.9402412.

[4]. C. Van Hee, G. Jacobs, C. Emmery, et al., "Automatic Detection of Cyberbullying in Social Media Text," *2018*.

[5]. F. Alqahtani and M. Ilyas, "A Machine Learning Ensemble Model for the Detection of Cyberbullying," *arXiv*, Feb. 2024.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 3, February 2026
Impact Factor: 8.2

[6]. R. Nivethitha and D. Jayashree, "Cyberbullying Detection in Social Networks using Machine Learning Models," in *ICCAP 2021*, doi:10.4108/eai.7-12-2021.2314577.