

CYBERBULLYING DETECTION ON SOCIAL MEDIA

A project submitted for Machine Learning (UML501)

Submitted by:

Kushali Gupta (102317148)

Manya Jindal (102317253)

BE Third Year

CSE

Submitted to:

Dr. Anjula Mehto

Assistant Professor



THAPAR INSTITUTE
OF ENGINEERING & TECHNOLOGY
(Deemed to be University)

Computer Science and Engineering Department

Thapar Institute of Engineering and Technology, Patiala

November 2025

TABLE OF CONTENTS

S. No	Topic	Page No.
1	Introduction or Project Overview	3
2	Problem Statement	3
3	Overview of the Dataset used	4
4	Project workflow	4
5	Results	6
6	Conclusion	6

1. Introduction:

Cyberbullying has become a major issue on social media platforms, where users can post harmful, abusive, or discriminatory content anonymously. Identifying such harmful text manually is extremely difficult due to the huge volume of online posts.

This project aims to develop a **Cyberbullying detection system** that automatically classifies a given social media post into categories such as *religion*, *gender*, *ethnicity*, *age*, or *other harassment*. To improve accuracy, the model uses **BERT (Bidirectional Encoder Representations from Transformers)** to extract contextual linguistic features from text.

These BERT embeddings are then used as input to two machine-learning classifiers:

- **Support Vector Machine (SVM)**
- **Random Forest Classifier**

This hybrid approach leverages both deep learning (BERT) and classical ML algorithms, producing a robust cyberbullying detection system.

2. Problem Statement:

With the rise of online communication, cyberbullying has increased drastically. Offensive content targeting religious groups, genders, ethnic communities, or individuals is harmful and often goes undetected on social media.

Challenges include:

- Sarcasm or indirect bullying
- Short and informal messages
- Spelling variations
- Context understanding

This project solves these challenges using BERT (for deeper meaning extraction) along with efficient machine-learning classifiers.

3. Overview of the Dataset Used:

The dataset is a **Cyberbullying Classification Dataset** downloaded from Kaggle. It contains labelled tweets with categories:

- **Religion**
- **Gender**
- **Ethnicity**
- **Age**
- **Other harassment**
- **Not Cyberbullying**

Number of Columns

- **tweet text:** Actual tweet text
- **cyberbullying type:** Label / class

Examples

Tweet Text	Label
"Christians are idiots."	religion
"Women can't drive."	gender
"You blacks are stupid."	ethnicity
"You're too old to understand this."	age

Dataset Characteristics

- Real tweets from Twitter
- Contains slangs, informal language, emojis
- Suitable for multi-class classification

4. Project Workflow:

Below is the complete pipeline followed in the project:

Step 1: Data Loading & Cleaning

- Loaded CSV dataset using pandas
- Removed null values
- Cleaned text (lowercasing, removing URLs, symbols, punctuation)

Step 2: Label Encoding

- Categorical labels (religion, gender, etc.) converted to numeric values using `Label Encoder()`

Step 3: BERT Tokenization

BERT tokenizer converts text into:

- Tokens
- Attention masks
- Input IDs

These are required by the BERT model.

Step 4: Extracting BERT Embeddings

Instead of training the entire BERT model , we use a pretrained BERT model (`bert-base-uncased`) to extract 768-dimensional embeddings for each tweet.

These embeddings capture:

- Meaning
- Context
- Sentence structure

Step 5: Training Machine Learning Models

Two ML algorithms are trained on BERT embeddings:

Support Vector Machine (SVM)

- Good for high-dimensional data
- Works well with BERT embeddings
- Provides strong baseline performance

Random Forest Classifier

- Ensemble of multiple decision trees
- Captures non-linear patterns
- Good for multi-class classification

Both models were trained on 80% data and tested on 20%.

Step 6: Prediction System

Created an interactive prediction loop:

Enter tweet: "You are stupid"
Model predicts: ethnicity: 86%, gender: 7%, religion: 4%

The system displays probability for each class from both models.

5. Results:

Both classifiers were evaluated using:

- Accuracy
- Classification Report
- Confusion Matrix

Summary of Results

Model	Accuracy	Strength
BERT + SVM	High	Best for classification, sharp decision boundaries
BERT + Random Forest	Moderate-High	Better interpretability, handles imbalance

Why BERT Improves Results

Traditional models only see words independently.

BERT understands:

- Sentence meaning
- Context
- Word relationships
- Hidden intent

This greatly boosts classification accuracy.

6. Conclusion

The project successfully built a **Cyberbullying Detection System** using the combined power of **BERT** and **Machine Learning**.

Key Achievements

- Achieved high accuracy using BERT embeddings
- Models classify tweets into multiple cyberbullying types
- Real-time prediction loop created for user input

Future Enhancements

- Deploy as a web app using Flask / Streamlit
- Add more datasets for better generalization