Detection and Analysis of Cyberbullying on Social Media Using Natural Language Processing

by

Alamin Hossain
CSE2203027117

Suraiya Akter Ety
CSE2203027165

Md. Yusuf Ali
CSE2203027161

Most. Nipa Akter
CSE2203027137

Md. Robin Mia
CSE2203027100

Supervised by
Tasnia Haque Kheya

Submitted in partial fulfillment of the requirement for the degree of Bachelor of Science in Computer Science and Engineering



DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
SONARGAON UNIVERSITY(SU)

September 2026

APPROVAL

The thesis titled **"Detection and Analysis of Cyberbullying on Social Media Using Natural Language Processing"** submitted by Alamin Hossain (CSE220307117), Md. Yusuf Ali (CSE2203027161), Suraiya Akter Ety (CSE2203027165), Most. Nipa Akter (CSE2203027137) and Md. Robin Mia (CSE2203027100) to the Department of Computer Science and Engineering, Sonargaon University (SU), has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Computer Science and Engineering and approved as to its style and contents.

Board of Examiners

Tasnia Haque Kheya — **Supervisor**

Lecturer,

Department of Computer Science and Engineering

Sonargaon University (SU)

(Examiner Name and Signature) — **Examiner 1**

Department of Computer Science and Engineering

Sonargaon University (SU)

(Examiner Name and Signature) — **Examiner 2**

Department of Computer Science and Engineering

Sonargaon University (SU)

(Examiner Name and Signature) — **Examiner 3**

Department of Computer Science and Engineering

Sonargaon University (SU)

DECLATATION

We, hereby, declare that the work presented in this report is the outcome of the investigation performed by us under the supervision of **Tasnia Haque Kheya, Lecturer,** Department of Computer Science and Engineering, Sonargaon University, Dhaka, Bangladesh. We reaffirm that no part of this thesis has been or is being submitted elsewhere for the award of any degree or diploma.

Countersigned

(Tasnia Haque Kheya)
Supervisor

Signature

Alamin Hossain
CSE2203027117

Md. Yusuf Ali
CSE2203027161

Suraiya Akter Ety
CSE2203027165

Most. Nipa Akter
CSE2203027137

Md. Robin Mia
CSE2203027100

ABSTRACT

The exponential growth of social media has amplified cyberbullying a pervasive form of online harassment that inflicts severe psychological distress on victims and unlike conventional bullying, cyberbullying on platforms such as Twitter operates through rapid, public, and often anonymous interactions, rendering automated detection particularly challenging. This study proposes a machine learning framework for detecting and classifying cyberbullying in English social media text, utilizing a balanced dataset of 10,000 manually annotated records across six categories (age, ethnicity, gender, religion, other cyberbullying, and not cyberbullying) to train twelve distinct models under standardized experimental conditions. Experimental results demonstrate that a fine-tuned BERT model achieved the highest overall performance with 84.65% accuracy, 84.59% precision, 84.65% recall, 84.48% F1-score, and 0.9734 ROC-AUC, while Random Forest recorded the strongest baseline at 83.85% accuracy and 83.94% F1-score; notably, BERT excelled at identifying identity-based harassment with over 93% F1-score for age, ethnicity, and religion, whereas the linguistically ambiguous "other cyberbullying" class remained the most challenging at 84.19% F1. These findings confirm that transformer-based contextual representations significantly outperform traditional classifiers for multi-class cyberbullying detection, and the proposed framework offers a viable solution for assisting content moderators in proactively identifying harmful content, thereby contributing to safer and more inclusive online environments.

ACKNOWLEDGMENT

At the very beginning, we would like to express my deepest gratitude to the Almighty Allah for giving us the ability and the strength to finish the task successfully within the schedule time.

We are auspicious that we had the kind association as well as supervision of **Tasnia Haque Kheya**, Lecturer, Department of Computer Science and Engineering, Sonargaon University whose hearted and valuable support with best concern and direction acted as necessary recourse to carry out our project.

We are also thankful to all our teachers during our whole education, for exposing us to the beauty of learning.

Finally, our deepest gratitude and love to my parents for their support, encouragement, and endless love.

LIST OF ABBREVIATIONS

API	Application Programming Interface
BERT	Bidirectional Encoder Representations from Transformers
CB	Cyberbullying
CLIP	Contrastive Language-Image Pre-training
CNN	Convolutional Neural Network
CPU	Central Processing Unit
CSV	Comma-Separated Values
DEA	Dolphin Echolocation Algorithm
DL	Deep Learning
FN	False Negatives
FP	False Positives
GNN	Graph Neural Network
GPU	Graphics Processing Unit
GTC	General Theory of Crime
HTTP	HyperText Transfer Protocol
HTTPS	HyperText Transfer Protocol Secure
IDE	Integrated Development Environment
LR	Logistic Regression
LSTM	Long Short-Term Memory
ML	Machine Learning
NB	Naive Bayes
NLP	Natural Language Processing
NLTK	Natural Language Toolkit
ONNX	Open Neural Network Exchange
OvR	One-vs-Rest
PII	Personally Identifiable Information
PLS-SEM	Partial Least Squares Structural Equation Modeling
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAM	Random Access Memory
RAT	Routine Activity Theory
ResNet	Residual Network
RF	Random Forest
ROC-AUC	Receiver Operating Characteristic – Area Under the Curve
RNN	Recurrent Neural Network
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency
TN	True Negatives
TP	True Positives
TPB	Theory of Planned Behavior
URL	Uniform Resource Locator
XGBoost	Extreme Gradient Boostin

TABLE OF CONTENTS

Title **Page No.**

DECLARATION ---------- **Iii**

ABSTRACT ---------- **Iv**

ACKNOWLEDGEMENT ---------- **V**

LIST OF ABBREVIATION ---------- **Vi**

CHAPTER 1 ---------- **1-11**

INTRODUCTION

1.1 Background of the Study ---------- **1**

1.2 Problem Statement ---------- **2**

1.3 Motivation of the Study ---------- **4**

1.4 Aim of the Research ---------- **4**

1.5 Research Objectives ---------- **5**

1.6 Research Questions ---------- **6**

1.7 Overview of the Proposed Method ---------- **6**

1.8 Preliminary Model Performance ---------- **8**

1.9 Significance of the Study ---------- **9**

1.10 Scope of the Study ---------- **11**

CHAPTER 2 ---------- **12-24**

LITERATURE REVIEW

2.1 Introduction ---------- **12**

2.2 Conceptualizing Cyberbullying and Its Definitional Criteria ---------- **12**

2.3 Theoretical Frameworks in Cyberbullying Research -- **14**

2.3.1 Routine Activity Theory and General Theory of Crime ---------- **14**

2.3.2 Theory of Planned Behavior and Extended Determinants ---------- **15**

2.3.3 The Buffer Role of Social and Parental Support ---------- **16**
2.4 Existing Methods for Cyberbullying Detection -------- **17**
2.4.1 Rule-Based and Keyword Approaches ------- **17**
2.4.2 Traditional Machine Learning Approaches -- **17**
2.4.3 Deep Learning Approaches ---------- **18**
2.4.4 Transformer-Based Approaches ---------- **18**
2.5 Comparative Review of Recent Studies ---------- **19**
2.6 Research Gaps and Limitations in Existing Literature - **22**
2.7 Chapter Overview ---------- **24**

CHAPTER 3 ---------- **25-49**
RESEARCH METHODOLOGY
3.1 Introduction ---------- **25**
3.2 Research Design ---------- **25**
3.3 Research Approach ---------- **26**
3.4 Conceptual Framework ---------- **27**
3.5 Study Area / Research Context ---------- **29**
3.6 Target Population ---------- **29**
3.7 Sample Size and Sampling Technique ---------- **30**
3.7.1 Sample Size ---------- **30**
3.7.2 Sampling Technique ---------- **30**
3.8 Sources of Data ---------- **32**
3.8.1 Overview ---------- **32**
3.8.2 Primary Data ---------- **33**
3.8.3 Secondary Data ---------- **33**
3.9 Data Collection Methods ---------- **34**
3.9.1 Questionnaire Survey ---------- **35**
3.9.2 Interview Method ---------- **35**
3.9.3 Observation Method ---------- **35**
3.9.4 Document Review / Online Data Collection - **36**
3.10 Research Instruments ---------- **37**
3.11 Variables and Measurement Criteria ---------- **37**

3.11.1 Variable Definitions ---------- **38**
3.11.2 Measurement Criteria ---------- **39**
3.12 Data Processing and Analysis Procedure ---------- **39**
3.12.1 Data Coding and Cleaning ---------- **40**
3.12.2 Statistical Analysis ---------- **41**
3.12.3 Graphical and Chart Analysis ---------- **42**
3.13 Tools and Software Used ---------- **42**
3.14 Validity and Reliability of Data ---------- **44**
3.14.1 Validity ---------- **44**
3.14.2 Reliability ---------- **45**
3.15 Ethical Considerations ---------- **45**
3.16 Research Workflow ---------- **46**
3.17 Project Implementation Plan / Gantt Chart ---------- **48**
3.18 Chapter Summary ---------- **49**

CHAPTER 4 ---------- **51-80**
RESULTS AND DISCUSSION
4.1 Introduction ---------- **51**
4.2 Experimental Setup ---------- **52**
4.2.1 Hardware and Software Configuration ------ **52**
4.2.2 Dataset Division for Training and Testing -- **53**
4.2.3 Data Preprocessing and Tokenization -------- **53**
4.2.4 Model Training Configuration ---------- **54**
4.2.5 Evaluation Metrics ---------- **57**
4.3 Dataset Distribution Analysis ---------- **57**
4.3.1 Class-wise Distribution ---------- **57**
4.3.2 Training and Testing Dataset Distribution -- **58**
4.4 Performance Evaluation of Machine Learning Models **60**
4.4.1 Performance of Naive Bayes Classifier ----- **60**
4.4.2 Performance of Support Vector Machine Classifier ---------- **61**
4.4.3 Performance of Logistic Regression Classifier ---------- **62**

4.4.4 Performance of Random Forest Classifier --- **63**
4.4.5 Comparative Analysis of Machine Learning Models ---------- **65**
4.5 Performance Evaluation of the BERT Model ---------- **66**
4.5.1 Training and Validation Performance of BERT ---------- **66**
4.5.2 Classification Report of the BERT Model --- **68**
4.5.3 Confusion Matrix Analysis of the BERT Model ---------- **69**
4.5.4 Error Analysis of the BERT Model ---------- **70**
4.6 Comparative Analysis of All Models ---------- **72**
4.6.1 Accuracy Comparison ---------- **72**
4.6.2 Precision, Recall, and F1-Score Comparison **73**
4.6.3 Comparison of Training Time and Testing Time ---------- **74**
4.6.4 Discussion of the Best Performing Model ---- **76**
4.7 Discussion of Findings ---------- **77**
4.7.1 Interpretation of Experimental Results ------- **77**
4.7.2 Analysis of Misclassified Comments -------- **78**
4.7.3 Strengths of the Proposed BERT-Based Model ---------- **79**
4.7.4 Limitations of the Experimental Results ----- **79**
4.8 Chapter Summary ---------- **80**

CHAPTER 5 ---------- **82-88**
CONCLUSION AND FUTURE WORK
5.1 Introduction ---------- **82**
5.2 Summary of the Study ---------- **82**
5.3 Achievement of Research Objectives ---------- **83**
5.4 Major Findings ---------- **83**
5.5 Contributions of the Study ---------- **84**
5.6 Limitations ---------- **84**
5.6.1 Limited Dataset Size ---------- **84**

5.62 Dataset Source and Generalization Limitation ---------- **85**
5.6.3 Class Distribution and Classification Difficulty ---------- **85**
5.6.4 Language and Linguistic Restriction --------- **86**
5.6.5 Text-Only and Multimodal Limitations ----- **86**
5.6.6 Difficulty in Detecting Sarcasm and Context-Dependent Language ---------- **86**
5.6.7 Limited Semantic Understanding of TF-IDF in Baseline Models ---------- **87**
5.6.8 Social Media Platform Limitation ---------- **87**
5.6.9 Ethical and Privacy Limitation ---------- **87**
5.6.10 Hardware and Computational Constraints -- **87**
5.7 5.7 Recommendations for Future Work ---------- **88**
5.8 5.8 Conclusion ---------- **88**
REFERENCES ---------- **90-92**

LIST OF TABLES

Table No.	Title	Page No.
Table 2.1	Comparative Review of Key Empirical and Theoretical Studies	19-22
Table 3.1	Research Design and Approach	26
Table 3.2	Population and Sample Size	30
Table 3.3	Sampling Technique and Respondent Categories	31
Table 3.4	Sources of Data	32
Table 3.5	Data Collection Instruments	34
Table 3.6	Variables and Measurement Scale	38-39
Table 3.7	Data Analysis Techniques	40-41
Table 3.8	Software and Tools Used	43
Table 3.9	Research Activities and Timeline	47-48
Table 4.1	Hardware and Software Configuration	52
Table 4.2	BERT Model Training Configuration	55-56
Table 4.3	Class-wise Distribution of the Full Dataset	57
Table 4.4	Training and Testing Dataset Distribution	58
Table 4.5	Classification Report Naive Bayes Classifier	61
Table 4.6	Classification Report Support Vector Machine Classifier	62
Table 4.7	Classification Report Logistic Regression Classifier -	63
Table 4.8	Classification Report Random Forest Classifier --	64
Table 4.9	Comparative Performance of Machine Learning Models	65
Table 4.10	BERT Training and Validation Metrics Across Epochs	67
Table 4.11	BERT Multi-Class Cyberbullying Classification Report	68-69
Table 4.12	Accuracy Comparison Across All Models	72-73
Table 4.13	Comprehensive Performance Metrics Across All Models	74
Table 4.14	Computational Efficiency Comparison	75-76

LIST OF FIGURES

Figure No.	Title	Page No.
Figure 3.1	Conceptual Framework of the Study ----------	27
Figure 3.2	Sample Distribution by Category (Stratified 80:20) ---	31
Figure 3.3	Sources of Data Used in the Research ----------	33
Figure 3.4	Data Collection Process Flowchart ----------	34-35
Figure 3.5	Distribution of Dataset Records by Cyberbullying Category ----------	36
Figure 3.6	Research Tools and Software Used in the Study -------	44
Figure 3.7	Overall Research Workflow ----------	46-47
Figure 3.8	Gantt Chart of Research Activities and Timeline -------	49
Figure 4.1	Training and Test set Distribution by Category ---------	60
Figure 4.2	Comparative Performance of Traditional Machine Learning Models ----------	66
Figure 4.3	BERT Training Loss vs Evaluation Loss ----------	68
Figure 4.4	Confusion Matrix of BERT Model ----------	70
Figure 4.5	Comparative Classification Accuracy Across All Evaluated Models ----------	73

CHAPTER 1

INTRODUCTION

1.1 Background of the Study

The rapid growth of social media has fundamentally transformed the way people communicate, share information, and interact with one another. Platforms such as Facebook, YouTube, Instagram, TikTok, X (formerly Twitter), and other online communication services allow users to publish comments, opinions, images, and other forms of content instantaneously to global audiences. While social media provides significant societal benefits including democratized information access, community building, and educational opportunities it has also created unprecedented opportunities for harmful online behavior, most notably cyberbullying.

In recent years, cyberbullying has emerged as a major public health concern due to its severe psychological, emotional, and in some cases physical impacts on victims. Empirical research consistently demonstrates that online harassment is significantly associated with an increased risk of depression, anxiety, social withdrawal, loss of self-esteem, sleep disturbances, and suicidal ideation [8]. Among adolescents, who represent the most active demographic on social media platforms, the consequences are particularly acute. Victims of cyberbullying frequently report diminished cognitive concentration, leading to school avoidance, declining academic performance, and long-term socio-emotional scarring that can persist into adulthood.

Cyberbullying refers to the use of digital communication technologies to intentionally harm, threaten, humiliate, or harass another individual. It may occur through social media posts, comments, private messages, fake accounts, threats, rumors, or the unauthorized distribution of personal content. Unlike traditional face-to-face bullying, cyberbullying possesses several distinctive characteristics that amplify its harmful potential: it can spread rapidly across networks, remain permanently accessible in digital archives, and reach an unlimited audience regardless of geographical boundaries. Consequently, victims may experience sustained emotional distress, social isolation, and psychological trauma without the physical sanctuary that traditionally offered respite from offline harassment.

Cyberbullying manifests in different forms and may target individuals based on age, ethnicity, gender, religion, sexual orientation, or other personal characteristics. A

particularly insidious challenge is that some online posts may not contain explicitly offensive words but may still communicate harmful or discriminatory meanings through context, sarcasm, coded expressions, or indirect insults. This linguistic subtlety renders simple keyword-based detection approaches inadequate. The global COVID-19 pandemic further exacerbated this vulnerability: the sudden transition to online learning and prolonged lockdown restrictions forced youth to spend unprecedented hours online, exposing them to heightened risks of digital hostility, social sadness, and isolation.

Given these complexities, manual monitoring of social media content is time-consuming, economically expensive, and practically impossible to apply at scale. Platforms generate millions of user-generated posts per minute, and harmful content may be expressed using informal language, abbreviations, spelling variations, slang, sarcasm, or indirect expressions. Machine learning and Natural Language Processing (NLP) provide an effective computational approach for automatically analyzing textual data at scale. In a typical NLP-based cyberbullying detection pipeline, raw social media text is cleaned, normalized, and transformed into numerical features before being fed to a classification algorithm.

This study investigates the comparative performance of multiple machine learning, deep learning, and transformer-based architectures for classifying different types of cyberbullying in English social media text. The evaluated models include traditional machine learning approaches (Logistic Regression, Support Vector Machine, Random Forest, Naïve Bayes, Extra Trees, CatBoost, XGBoost), deep learning architectures (BiLSTM, CNN-BiLSTM, TextCNN), and a transformer-based model (BERT). The classification categories considered in this research are: age-based cyberbullying, ethnicity-based cyberbullying, gender-based cyberbullying, religion-based cyberbullying, other cyberbullying (general harassment), and not cyberbullying (benign content).

1.2 Problem Statement

The increasing volume of user-generated content on social media platforms makes it exceedingly difficult for human moderators to identify, classify, and respond to cyberbullying manually. A large number of comments and posts are published every minute across multiple platforms, and harmful content may be deliberately obfuscated

using informal language, abbreviations, spelling variations, sarcasm, coded expressions, or indirect threats. These linguistic characteristics render automated cyberbullying detection a challenging multi-class text classification problem that extends far beyond simple profanity filtering.

A fundamental theoretical challenge lies in the difficulty of applying traditional, face-to-face bullying criteria such as intent to harm, repetition, and power imbalance directly to the digital domain. The distinct affordances of online environments fundamentally alter these dynamics through two primary mechanisms: perceived anonymity and viral publicity. Anonymity releases perpetrators from normative and social constraints, fostering the "online disinhibition effect" [15] and allowing attackers to harm victims without fear of immediate consequences, social accountability, or physical confrontation. Meanwhile, the viral nature of internet communication introduces a complex dimension of publicity: a single humiliating post or image can be replicated, screenshotted, and viewed by an unlimited audience indefinitely, creating a permanent digital footprint that subjects the victim to continuous victimization over time without requiring repeated actions by the original offender.

From a computational perspective, existing keyword-based approaches and rule-based moderation systems frequently fail to identify cyberbullying when harmful meaning is expressed without explicit abusive words. Furthermore, detecting whether a text contains cyberbullying is only one dimension of the problem. Identifying the specific type of cyberbullying is equally critical because different categories may require distinct moderation protocols, intervention strategies, and support resources. For example, identity-based harassment targeting religion or ethnicity may necessitate different community guidelines enforcement than general interpersonal aggression.

The main problem addressed in this research is the development and rigorous comparative evaluation of an automated system that can analyze English social media text and accurately classify it into multiple cyberbullying categories. Specifically, the study seeks to determine which algorithmic paradigm traditional machine learning with TF-IDF features, deep learning with learned embeddings, or transformer-based contextual representations provides the optimal balance of accuracy, generalization, and computational efficiency for this multi-class classification task.

1.3 Motivation of the Study

The motivation for this research is grounded in three interrelated imperatives: the escalating scale of harmful online communication, the inadequacy of existing moderation approaches, and the academic need for systematic comparative evaluation. First, the increasing global penetration of social media and the growing difficulty of controlling harmful online communication demand automated solutions. Cyberbullying affects individuals' emotional well-being, social relationships, academic performance, and mental health across all age groups, with particularly devastating effects on adolescents and young adults. Early and accurate detection may enable social media platforms, educational institutions, parents, and mental health professionals to respond to harmful content more swiftly and effectively, potentially preventing the escalation of online harassment into severe psychological trauma or self-harm.

Second, existing content moderation systems deployed by major platforms remain predominantly reactive and keyword-dependent. These systems struggle with the linguistic creativity of online aggressors, who rapidly adapt their vocabulary to evade filters. A context-aware, machine learning-driven classification system capable of understanding semantic nuance represents a significant advancement over current industry practices.

Third, from an academic standpoint, there is a pressing need to examine how different classification algorithms perform on the same cyberbullying dataset under standardized conditions. Different models learn different textual representations: linear models such as Logistic Regression and SVM operate effectively with sparse TF-IDF features, capturing term importance and frequency patterns; deep learning models such as BiLSTM and CNN learn sequential and local phrase patterns from raw text; and transformer-based models such as BERT capture bidirectional contextual relationships between words. A rigorous comparative evaluation across these paradigms can identify the most suitable model architecture for multi-class cyberbullying classification and illuminate the trade-offs between accuracy, computational cost, and interpretability.

1.4 Aim of the Research

The primary aim of this research is to develop, implement, and evaluate a comprehensive machine learning and deep learning-based framework for detecting and

classifying different types of cyberbullying from English social media text. The study specifically aims to:

(i) Determine whether systematic text preprocessing, appropriate feature extraction strategies (TF-IDF for traditional models; WordPiece tokenization for transformer models), and supervised classification algorithms can effectively distinguish among six cyberbullying categories.

(ii) Conduct a rigorous head-to-head comparison of twelve distinct algorithmic architectures under identical experimental conditions to identify the most suitable model for multi-class cyberbullying type classification.

(iii) Evaluate the practical trade-offs between predictive accuracy and computational efficiency to inform real-world deployment decisions for automated content moderation systems.

1.5 Research Objectives

To achieve the stated aim, this research pursues the following specific objectives:

i. To acquire and prepare a balanced, annotated social media text dataset suitable for supervised multi-class cyberbullying classification, ensuring data quality through structural validation and missing value handling.
ii. To implement a systematic text preprocessing pipeline encompassing lowercase conversion, URL and user mention removal, special character and punctuation cleaning, number removal, whitespace normalization, and slang expansion to reduce noise and standardize textual input.
iii. To engineer appropriate numerical feature representations for different model categories: TF-IDF vectorization for traditional machine learning models, and BERT WordPiece tokenization with attention masks for transformer-based models.
iv. To develop and train twelve distinct classification models Logistic Regression, SVM, Random Forest, Naïve Bayes, Extra Trees, CatBoost, XGBoost, BiLSTM, CNN-BiLSTM, TextCNN, Soft Voting ensemble, and BERT under standardized hyperparameter configurations and data splits.
v. To evaluate and compare model performance using standard NLP metrics (accuracy, macro precision, macro recall, macro F1-score, and ROC-AUC) and 5-Fold Stratified Cross-Validation for traditional machine learning baselines.

vi. To analyze class-wise performance disparities, confusion patterns, and error archetypes to identify the linguistic characteristics that facilitate or impede accurate cyberbullying type classification.

vii. To identify the strengths, limitations, and practical deployment considerations of the proposed cyberbullying classification framework.

viii. To determine the most suitable model for classifying cyberbullying categories in English social media text based on a multi-criteria evaluation of predictive accuracy, generalization stability, and computational efficiency.

1.6 Research Questions

This study addresses the following research questions, which are designed to be specific, measurable, and aligned with the stated objectives:

RQ1. How effectively can machine learning, deep learning, and transformer-based techniques detect and classify different types of cyberbullying from English social media text into six mutually exclusive categories?

RQ2. What is the comparative impact of TF-IDF-based feature extraction versus contextual embedding-based representation (BERT WordPiece tokenization) on classification performance across different algorithmic families?

RQ3. Which algorithmic paradigm traditional machine learning, deep learning trained from scratch, or transformer-based fine-tuning provides the highest predictive accuracy for multi-class cyberbullying type classification?

RQ4. How do the evaluated models compare in terms of accuracy, precision, recall, F1-score, and ROC-AUC, and what are the computational efficiency trade-offs associated with each paradigm?

RQ5. Can the proposed approach accurately distinguish among age-based, ethnicity-based, gender-based, religion-based, other cyberbullying, and not cyberbullying categories, and which category pairs are most susceptible to misclassification?

1.7 Overview of the Proposed Method

The proposed research follows a systematic sequence of data preparation, feature extraction, model training, and performance evaluation steps, as illustrated in the conceptual framework presented in Chapter 3.

Step 1: Data Acquisition and Validation. A cyberbullying-related social media dataset comprising 10,000 annotated English text records is obtained from Kaggle. The dataset is inspected for structural integrity, missing values are handled via imputation, and class distributions are verified to ensure balanced representation across six categories.

Step 2: Text Preprocessing and Normalization. The raw text undergoes systematic cleaning: conversion to lowercase, removal of URLs and user mentions via regular expressions, elimination of special characters and punctuation, removal of numeric digits, and collapsing of extra whitespace. For traditional machine learning models, an additional slang expansion step is applied using a curated dictionary to standardize informal abbreviations.

Step 3: Feature Extraction. Two parallel feature extraction strategies are employed. For traditional machine learning and deep learning models (excluding BERT), the cleaned text is transformed into sparse numerical vectors using TF-IDF, which assigns higher importance to terms frequent in a document but rare across the corpus. For the BERT transformer model, the raw preprocessed text is tokenized using the HuggingFace AutoTokenizer with the pretrained bert-base-uncased vocabulary, generating input IDs, attention masks, and token type IDs with dynamic padding to a maximum sequence length of 128 tokens.

Step 4: Dataset Partitioning. The dataset is divided into training and testing subsets using an 80:20 stratified split ratio (random_state = 42), yielding 8,000 training samples and 2,000 test samples. Stratified sampling preserves the relative distribution of target classes in both subsets, preventing class-bias during training and evaluation.

Step 5: Model Training and Validation. Traditional machine learning models are trained on TF-IDF feature matrices with 5-Fold Stratified Cross-Validation to obtain robust performance estimates. Deep learning models are trained on sequence data with early stopping and dropout regularization. The BERT model is fine-tuned for three epochs using the HuggingFace Trainer API with a learning rate of 2×10^{-5}, batch size of 8, and weight decay of 0.01. No separate validation set is held out for BERT; instead, training and validation loss curves are monitored per epoch to detect convergence and prevent overfitting.

Step 6: Evaluation and Comparison. All models are evaluated on the held-out test set using accuracy, macro precision, macro recall, macro F1-score, and ROC-AUC. Confusion matrices and class-wise classification reports are generated to identify

systematic misclassification patterns. Training and inference times are recorded to assess computational efficiency.

1.8 Preliminary Model Performance

To inform the final experimental design and validate the feasibility of the proposed approach, preliminary pilot experiments were conducted on the complete dataset. All models were evaluated using identical train-test splits (80:20, stratified, random_state = 42) and standardized metrics, including accuracy, precision, recall, F1-score, and ROC-AUC.

The BERT model achieved the best overall performance among all evaluated models, with an accuracy of **0.8465**, precision of **0.8459**, recall of **0.8465,** F1-score of **0.8448**, and ROC-AUC of **0.9734**. These results indicate that BERT was more effective than the other models in distinguishing between the cyberbullying categories in the dataset.

Among the traditional machine learning models, Random Forest achieved the strongest performance, recording an accuracy of **0.8385**, precision of **0.8468**, recall of **0.8385**, F1-score of **0.8394**, and ROC-AUC of **0.9665**. Although Random Forest obtained a slightly higher precision score than BERT (0.8468 versus 0.8459), BERT achieved better overall accuracy, recall, F1-score, and ROC-AUC. Therefore, BERT can be considered the best-performing model based on the overall evaluation results.

The ensemble Soft Voting approach yielded an accuracy of **0.8400**, precision of **0.8432**, recall of **0.8400**, F1-score of **0.8376**, and ROC-AUC of **0.9702**, placing it between BERT and Random Forest in overall effectiveness. XGBoost achieved an accuracy of **0.8275**, precision of **0.8368**, recall of **0.8275**, F1-score of **0.8312**, and ROC-AUC of **0.9662**, while Logistic Regression recorded an accuracy of **0.8260**, precision of **0.8290**, recall of **0.8260**, F1-score of **0.8255**, and ROC-AUC of **0.9652**. The SVM model produced an accuracy of **0.8255**, precision of **0.8248**, recall of **0.8255**, F1-score of **0.8245**, and ROC-AUC of **0.9618**. Extra Trees achieved an accuracy of **0.8335**, precision of **0.8361**, recall of **0.8335**, F1-score of **0.8325**, and ROC-AUC of **0.9639**, and CatBoost recorded an accuracy of **0.8130**, precision of **0.8362**, recall of **0.8130**, F1-score of **0.8295**, and ROC-AUC of **0.9658**.

Among the deep learning architectures, TextCNN achieved an accuracy of **0.8120**, precision of **0.8254**, recall of **0.8120**, F1-score of **0.8032**, and ROC-AUC of **0.9637**. The CNN-BiLSTM hybrid model recorded an accuracy of **0.7785**, precision of **0.7756**,

recall of **0.7785**, F1-score of **0.7767**, and ROC-AUC of **0.9528**. The BiLSTM model produced the lowest results among all evaluated approaches, with an accuracy of **0.7485**, precision of **0.7416**, recall of **0.7485**, F1-score of **0.7238**, and ROC-AUC of **0.9473**. Naïve Bayes also performed relatively poorly among the traditional models, with an accuracy of **0.7580**, precision of **0.7488**, recall of **0.7580**, F1-score of **0.7402**, and ROC-AUC of **0.9603**.

These results suggest that contextual language representation is critically important for multi-class cyberbullying detection. Unlike traditional TF-IDF-based models, which treat words as independent features, BERT captures bidirectional contextual relationships between words and may therefore better identify harmful text where the meaning depends on the surrounding linguistic context. Nevertheless, each model should be interpreted carefully because performance can be influenced by dataset size, class distribution, preprocessing procedures, feature representation, hyperparameter settings, and the train-test splitting strategy.

For the final system, BERT is selected as the most suitable model because it achieved the highest overall performance across the reported metrics. Random Forest may be considered a strong alternative where computational resources are limited, as it produced competitive results with comparatively lower computational complexity and faster training times.

1.9 Significance of the Study

This study holds significance across multiple domains, including cyber safety, social media governance, natural language processing, and automated content moderation.

Theoretical Significance: The study contributes to the theoretical understanding of how different computational paradigms symbolic (TF-IDF), connectionist (deep learning), and attention-based (transformers)perform on the same multi-class text classification task. By systematically comparing twelve models under standardized conditions, the study provides empirical evidence for the relative importance of contextual semantic understanding in detecting nuanced online hostility.

Practical Significance: The proposed classification system may help identify potentially harmful social media text and categorize it according to the type of cyberbullying. Such a system could support human moderators by filtering large volumes of online content

and prioritizing suspicious posts for further review, thereby reducing moderator burnout and improving response times.

Societal Significance: From an interventionist perspective, this study underscores that technical classification models must be aligned with broader psychosocial support frameworks. While machine learning and transformer models like BERT act as digital guardians to detect and filter abuse, existing literature highlights that human-centric "capable guardians" both online and offline are critical for holistic victim protection. Robust parental communication, peer social support networks, and active mediation serve as powerful protective buffers that significantly mitigate psychological distress and prevent extreme outcomes such as suicidal ideation among victimized youth [4]. By categorizing cyberbullying types, this study provides platform administrators, educators, and policy-makers with precise empirical data needed to develop targeted intervention and digital literacy programs.

The study is expected to benefit the following stakeholders:

- Social media platform administrators, who can integrate the classification framework into content moderation pipelines to detect and filter harmful content at scale.
- Educational institutions, which can utilize the system to monitor cyberbullying-related risks among students and inform anti-bullying policies.
- Parents and guardians, who can gain a better understanding of harmful online communication patterns and the categories of digital aggression their children may encounter.
- Cybersecurity and NLP researchers, who can extend the proposed methodology to other abuse-detection domains or multilingual contexts.
- Mental health and social support organizations, which can leverage category-specific detection data to tailor counseling and support resources for different victim profiles.
- Future researchers working on multilingual, multimodal, and context-aware cyberbullying detection, who can use this study as a methodological baseline.

Furthermore, this research contributes to the comparative understanding of traditional machine learning and deep learning models for multi-class cyberbullying classification, a gap that persists in the existing literature.

1.10 Scope of the Study

The scope of this study is defined by the following boundaries:

- Detection Domain: Detection and multi-class classification of cyberbullying in social media text.
- Linguistic Scope: Analysis is restricted to English-language social media text. Code-mixed, multilingual, or non-English content is excluded.
- Modality Scope: The study focuses exclusively on text-based classification. Images, videos, audio, memes, emoji semantics, user-network behavior, and detailed social network structures are outside the primary scope.
- Feature Representation: TF-IDF vectorization is employed for traditional machine learning and deep learning models; BERT WordPiece tokenization is employed for the transformer model.
- Evaluation Framework: Standard supervised learning metrics (accuracy, precision, recall, F1-score, ROC-AUC) are used for evaluation. An 80:20 stratified train-test split with random_state = 42 and 5-Fold Stratified Cross-Validation for baseline models constitute the validation strategy.
- Target Categories: The study classifies text into six categories: age-based cyberbullying, ethnicity-based cyberbullying, gender-based cyberbullying, religion-based cyberbullying, other cyberbullying (general harassment), and not cyberbullying (benign content).

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Cyberbullying has rapidly emerged as a critical societal and public health challenge in the digital era, facilitated by the widespread adoption of social media, smartphones, and connected communication networks. While digital technologies offer unprecedented channels for interaction and learning, they also lower the physical and social barriers to aggressive behaviors. From a computational perspective, the identification and mitigation of cyberbullying present complex natural language processing (NLP) problems due to the linguistic ambiguity, informal slang, abbreviations, and context-dependent nature of online communication.

This chapter provides a comprehensive review of the scholarly literature, establishing the conceptual definitions, theoretical frameworks, linguistic dimensions, and algorithmic methodologies that underpin automated cyberbullying detection systems. The chapter is organized as follows: Section 2.2 conceptualizes cyberbullying and examines its definitional criteria; Section 2.3 discusses the theoretical frameworks employed in cyberbullying research; Section 2.4 reviews existing computational methods for automated detection; Section 2.5 presents a comparative analysis of recent empirical studies; Section 2.6 identifies critical research gaps and limitations in the existing literature; and Section 2.7 provides a chapter summary that synthesizes the reviewed literature and positions the current study within the broader academic landscape.

2.2 Conceptualizing Cyberbullying and Its Definitional Criteria

A central focus of contemporary research lies in establishing a rigorous, standardized definition of cyberbullying. Early definitions conceptualized cyberbullying as an extension of traditional, face-to-face bullying into the digital domain, often borrowing the foundational criteria proposed by pioneering scholars such as Dan Olweus (1993). These traditional criteria include intentionality (the deliberate intent to cause harm), repetition (acts occurring repeatedly over time), and an interpersonal power imbalance

where the victim cannot easily defend themselves. However, a major consensus in recent literature indicates that applying traditional bullying criteria directly to online harassment is highly complex and often inadequate due to the distinct architectural affordances of digital platforms [14].

To establish a comprehensive operational definition, a systematic literature review was conducted to identify seven core definitional criteria that characterize cyberbullying in the online landscape [3]. These criteria reconcile traditional behavioral traits with digital realities and are summarized below:

(i) Intentionality: The cyberbully must demonstrate a purposeful and willful intent to inflict psychological, emotional, or social harm. Unlike accidental misunderstandings, cyberbullying involves deliberate actions to harass, humiliate, or threaten the targeted individual. The detection of intentionality in text-based communication remains one of the most significant challenges for automated systems, as intent must often be inferred from linguistic cues rather than explicit declarations [3].

(ii) Repetition: While traditional bullying requires multiple distinct aggressive actions, the digital environment fundamentally redefines repetition. A single harmful post, image, or video uploaded online can be viewed, copied, and shared indefinitely by a vast audience. Thus, a single act by an offender can result in continuous, repetitive victimization over time, complicating the temporal boundaries of the offense [3].

(iii) Power Imbalance: In the physical world, power is often defined by physical size or social dominance. Online, power imbalance is manifested through technological superiority, elevated digital social status, follower count differentials, or situational advantages such as anonymity. The victim is frequently left isolated and unable to counter the attack due to their inability to remove the harmful content or identify the perpetrator [3].

(iv) Anonymity: Online platforms enable perpetrators to hide their real identities behind pseudonyms or fake accounts. Anonymity fosters the "online disinhibition effect" [15], releasing individuals from social norms and moral constraints, which increases their willingness to engage in aggressive behaviors without fear of direct consequences or social accountability.

(v) Publicity: Traditional bullying is typically limited to immediate physical spaces. Cyberbullying leverages the massive, public scale of the internet. Public exposure such as posting abusive comments on a victim's profile or sharing private photos in public forums drastically escalates the emotional distress and severity of the harassment. The

permanence of digital content means that victims may be re-victimized each time the content is viewed or shared [3].

(vi) Aggression: This encompasses direct or relational violence, including verbal insults, spreading malicious rumors, deliberate ostracism of the victim from online groups or communities, and impersonation. Aggression in cyberbullying can be overt (explicit threats and slurs) or covert (subtle exclusion, manipulation, and gaslighting).

(vii) Electronic Devices and Media: By definition, cyberbullying is mediated entirely through digital hardware (smartphones, laptops) and software platforms (social networking sites, instant messaging apps, online games). Constant availability means victims can be targeted 24/7, leaving no physical sanctuary and blurring the boundaries between safe and unsafe spaces.

Despite the comprehensiveness of these seven criteria, several definitional challenges persist in the literature. First, cultural variations influence what constitutes harmful behavior; a term considered offensive in one cultural context may be benign or reclaimed in another. Second, the rapid evolution of digital platforms introduces new modalities of harassment (e.g., ephemeral content, augmented reality interactions) that existing definitions may not adequately capture. Third, the distinction between cyberbullying and related constructs such as cyberaggression, online harassment, and hate speech remains conceptually blurred in many studies, leading to inconsistent operationalizations and comparability issues across research findings.

2.3 Theoretical Frameworks in Cyberbullying Research

To systematically model, predict, and analyze cyberbullying behaviors, researchers draw upon established criminological, social, and psychological theories. The integration of these theories provides a holistic view of the interaction between the offender, the victim, the digital environment, and the protective systems that may mitigate harm. This section reviews three theoretical frameworks that are particularly relevant to the current study.

2.3.1 Routine Activity Theory and General Theory of Crime

Integrating the General Theory of Crime (GTC) and Routine Activity Theory (RAT) creates a highly effective conceptual model for understanding online deviance [10]. According to Routine Activity Theory, criminal or deviant acts are likely to occur when

three essential elements converge in time and space: (1) a motivated offender, (2) a suitable target, and (3) the absence of a capable guardian [6].

In the cyber domain, a "suitable target" is represented by vulnerable social media users who share sensitive or risky personal information online, exhibit low privacy settings, or participate in emotionally charged discussions. "Capable guardianship" is divided into offline guardians (parents, educators, law enforcement) and online mechanisms (privacy controls, content moderation algorithms, and automatic classification systems). The current study positions its proposed BERT-based classification system as an "online capable guardian" within this theoretical framework, designed to detect and filter harmful content before it reaches and damages the target victim.

The General Theory of Crime complements this model by characterizing the offender. GTC asserts that deviant behavior is a direct manifestation of low self-control. Individuals with low self-control are impulsive, risk-seeking, short-tempered, and self-centered. When these individuals are presented with tech-enabled opportunities and encounter vulnerable online targets in environments lacking active guardianship, their likelihood of engaging in cyberbullying behaviors escalates significantly. Court case analyses of adult cyberbullying consistently validate that low self-control is a highly significant predictor of cyberbullying propensity across both genders and different age groups [7].

The integration of RAT and GTC provides a robust theoretical justification for automated detection systems. By serving as an "online capable guardian," such systems reduce the suitability of targets and increase the perceived risk for offenders, thereby disrupting the convergence of elements necessary for cyberbullying to occur.

2.3.2 Theory of Planned Behavior and Extended Determinants

The Theory of Planned Behavior (TPB) is widely utilized to examine the cognitive determinants of cyberbullying intentions. TPB posits that behavioral intentions are driven by three core components: (1) behavioral attitudes (the individual's positive or negative evaluation of performing the behavior), (2) subjective norms (the perceived social pressure from significant others, such as parents, peers, or teachers, to perform or refrain from the behavior), and (3) perceived behavioral control (self-efficacy in executing the behavior) [1].

The classical TPB model has been extended in a Saudi Arabian context by introducing three vital contextual factors: social media use intensity, the absence of parental control, and a lack of regulatory guidelines. This integrated framework demonstrates that a lack of regulatory guidelines and parental monitoring directly inflates an adolescent's intention to engage in online harassment. Crucially, the extended TPB model establishes a direct link showing that heightened intentions toward cyberbullying negatively correlate with student academic performance, providing a comprehensive empirical pathway from cognitive intent to tangible life outcomes [2].

The TPB framework is particularly relevant to this study because it highlights the environmental and social determinants of cyberbullying behavior. Automated detection systems, while technical in nature, must be understood within this broader behavioral ecology. A classification model that accurately identifies cyberbullying content can inform educational interventions, parental guidance strategies, and platform policy enforcement, thereby addressing the attitudinal and normative components of the TPB framework.

2.3.3 The Buffer Role of Social and Parental Support

In understanding the psychological outcomes of cyberbullying, recent empirical research has shifted focus toward identifying protective factors that mitigate harm. The impact of cyberbullying among youth during the COVID-19 pandemic has been examined, a period when forced online learning and social isolation drastically increased exposure to online victimization. The findings demonstrated that cyberbullying exposure is significantly associated with severe psychological distress, including sadness, loneliness, and self-harm ideation [4].

However, the research established a powerful moderating effect: high levels of parental communication and peer social support serve as protective buffers. Specifically, victims who reported strong parental worry, care, and availability, combined with reliable social support from friends, demonstrated significantly lower levels of suicidal ideation and emotional distress. This highlights that technical moderation systems must be aligned with social and parental intervention strategies to effectively safeguard vulnerable users [4].

This theoretical perspective underscores an important limitation of purely algorithmic approaches: automated detection is a necessary but insufficient component of a

comprehensive anti-cyberbullying strategy. The proposed classification system in this study is therefore conceptualized not as a standalone solution, but as one component of a multi-layered protective framework that includes human moderation, parental engagement, and peer support networks.

2.4 Existing Methods for Cyberbullying Detection

Automated cyberbullying detection methods have transitioned through multiple paradigm shifts, moving from simple rule-based heuristics to traditional machine learning, sequential deep learning, and state-of-the-art transformer-based architectures. This section reviews the methodological evolution of detection systems, identifying the strengths and limitations of each approach.

2.4.1 Rule-Based and Keyword Approaches

Early attempts at cyberbullying detection relied on rule-based systems and keyword matching. These approaches maintain lists of offensive terms, profanities, and slurs, flagging content that contains matches. While computationally trivial to implement, keyword-based systems suffer from severe limitations: they fail to detect implicit threats, sarcasm, coded language, and context-dependent hostility. Furthermore, they generate high false-positive rates when benign content contains flagged terms (e.g., quoting offensive language for educational purposes or using reclaimed terminology within communities). Consequently, rule-based approaches have been largely superseded by machine learning methodologies in academic research.

2.4.2 Traditional Machine Learning Approaches

Traditional machine learning classifiers such as Support Vector Machines (SVM), Random Forests (RF), Naïve Bayes (NB), and Logistic Regression (LR) have long served as baselines for text classification in cyberbullying research. These models rely on manually engineered features and text representation techniques, most commonly Term Frequency-Inverse Document Frequency (TF-IDF).

TF-IDF transforms text into sparse numerical vectors by weighting terms based on their frequency within a document and their rarity across the corpus. While computationally efficient and interpretable, TF-IDF representations suffer from several critical limitations: they treat words as independent atomic units, ignoring word order, syntactic

structures, and semantic relationships; they cannot capture polysemy (words with multiple meanings); and they fail to generalize to out-of-vocabulary terms common in informal social media communication. Despite these limitations, TF-IDF-based models remain valuable as lightweight baselines and are frequently employed in ensemble configurations.

2.4.3 Deep Learning Approaches

Deep learning models such as Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM), and Bidirectional LSTMs (BiLSTM) improve on traditional methods by automatically learning hierarchical feature representations from raw or minimally processed text. CNNs excel at capturing local phrase patterns and n-gram features through sliding convolutional filters, making them effective at detecting explicit offensive phrases. Recurrent architectures (LSTM, BiLSTM) model sequential word dependencies and long-range contextual relationships, which are essential for understanding grammatically complex threats.

Fine-tuning network parameters is critical to achieving high accuracy. A hybrid deep learning model combining Elman-type Recurrent Neural Networks (RNN) with an optimized Dolphin Echolocation Algorithm (DEA) demonstrated the effectiveness of advanced parameter optimization. By utilizing DEA to optimize the weight and bias parameters of the RNN layers, the model significantly minimized training convergence times and outperformed standard LSTM and SVM models, proving the value of advanced hyperparameter optimization in short-text classification [12].

However, deep learning models trained from scratch require large volumes of annotated data to learn robust embeddings. In small-to-medium datasets typical of cyberbullying research, these models often underperform relative to well-tuned traditional classifiers or transfer learning approaches, as they lack the linguistic priors necessary to handle social media noise effectively.

2.4.4 Transformer-Based Approaches

More recently, Transformer-based models, particularly Bidirectional Encoder Representations from Transformers (BERT), have redefined state-of-the-art performance in cyberbullying detection. BERT processes sentences bidirectionally using multi-head self-attention mechanisms, enabling it to capture subtle contextual

relationships between words across the entire input sequence. This architectural advantage makes BERT exceptionally skilled at identifying implicit abuse, sarcasm, and linguistically ambiguous threats that confound bag-of-words and sequential models.

BERT and its variants (RoBERTa, DistilBERT, ALBERT) leverage transfer learning: they are pretrained on massive general-domain corpora and then fine-tuned on task-specific datasets. This pretraining-fine-tuning paradigm circumvents the data scarcity problem that plagues deep learning models trained from scratch, as the model arrives with rich linguistic knowledge that requires only domain-specific adaptation. For cyberbullying detection, this means BERT can recognize novel offensive expressions by leveraging its understanding of semantic similarity and contextual polarity learned during pretraining.

Despite these advantages, transformer models are computationally expensive to train and deploy, often requiring GPU acceleration for practical fine-tuning. Their "black box" nature also poses challenges for interpretability, which is a significant concern in moderation contexts where users may demand explanations for content removal decisions.

2.5 Comparative Review of Recent Studies

Table 2.1 presents a comprehensive comparative review of key empirical and theoretical studies that shape the current understanding of cyberbullying, summarizing their methodologies, findings, and critical limitations.

Table 2.1: Comparative Review of Key Empirical and Theoretical Studies

Study & Author	**Theoretical & Empirical Scope**	**Methodology Used**	**Key Findings & Limitations**
Amalina et al. (2022) [1]	Systematic review of cyberbullying conceptualizations and definitional criteria.	PRISMA search strategy; Systematic content analysis of 60 scholarly	Established 7 core definitional criteria: Intent, Repetition, Power Imbalance, Anonymity,

Study & Author	Theoretical & Empirical Scope	Methodology Used	Key Findings & Limitations
		publications (2010–2020).	Publicity, Aggression, and Electronic Devices. Limited by cultural variations and evolving technology.
Kaluarachchi et al. (2022) [2]	Analysis of adult cyberbullying, offender self-control, and guardianship structures.	Mixed method; Purposive sampling of 75 court transcripts; Deductive mapping into RAT and GTC frameworks.	Demonstrated that offender's low self-control is a highly significant predictor. Found online guardianship less effective than legal/offline intervention once bullying commences.
Alotaibi (2019) [3]	Relationship between cognitive cyberbullying intentions and student academic achievement.	Structural Equation Modeling (PLS-SEM); extended TPB model with parental	Supported direct positive relations between lack of parental control, lack of rules, and cyberbullying intentions.

Study & Author	Theoretical & Empirical Scope	Methodology Used	Key Findings & Limitations
		supervision and regulation factors.	Confirmed that cyberbullying directly degrades academic performance.
Raquel António et al. (2023) [4]	Cyberbullying prevalence, psychological outcomes, and protective factors during COVID-19 lockdowns.	Two online correlational studies (Study 1: N=485; Study 2: N=952) of Portuguese youth.	Found high cyberbullying prevalence. Proved that parental communication and peer social support act as powerful buffers, significantly reducing suicidal ideation among victims.
Murshed et al. (2022) [5]	Algorithmic optimization for automated cyberbullying text classification on social media.	DEA-RNN hybrid deep learning approach; Elman RNN optimized via Dolphin Echolocation Algorithm.	Proposed model achieved an accuracy of 90.45% in optimal scenarios, drastically reducing RNN training time. Limited to

Study & Author	Theoretical & Empirical Scope	Methodology Used	Key Findings & Limitations
			English text and content-only features.

The comparative analysis presented in Table 2.1 reveals several important patterns in the existing literature. First, the reviewed studies span diverse methodological traditions from systematic reviews and mixed-methods court analyses to structural equation modeling and deep learning optimization indicating a maturing, multidisciplinary field. Second, a consistent finding across theoretical studies is the inadequacy of purely technical solutions; effective cyberbullying mitigation requires integration with social support structures, parental engagement, and regulatory frameworks. Third, the computational studies reveal a clear performance trajectory from traditional machine learning to deep learning to transformer-based architectures, though direct comparisons are complicated by differences in datasets, evaluation metrics, and experimental conditions.

Notably, the highest accuracy (90.45%) among the reviewed computational studies was reported; however, this figure was achieved on a binary classification task with a substantially larger dataset than that employed in the current study. The performance figures reported across studies are therefore not directly comparable, underscoring the need for standardized benchmark datasets and evaluation protocols in cyberbullying detection research [12].

2.6 Research Gaps and Limitations in Existing Literature

Despite the significant advances reviewed in the preceding sections, the existing literature exhibits several critical gaps and limitations that the current study seeks to address:

(a) Limited Multi-Class Classification Focus: The majority of existing studies frame cyberbullying detection as a binary classification problem (bullying vs. non-bullying). Few studies have systematically investigated multi-class classification that distinguishes between different types of cyberbullying (age-based, ethnicity-based,

gender-based, religion-based, and other forms). Understanding the specific category of cyberbullying is essential for targeted intervention strategies, as different forms may require distinct moderation responses and support resources.

(b) Over-Reliance on TF-IDF for Traditional Models: While TF-IDF remains the dominant feature extraction technique for traditional machine learning models in cyberbullying research, few studies have comprehensively compared the performance of diverse traditional algorithms (SVM, RF, NB, LR, Extra Trees, CatBoost, XGBoost) on the same dataset using identical preprocessing pipelines. This gap limits understanding of which traditional approach is most suitable for cyberbullying type classification.

(c) Inadequate Comparison with Transformers: Many studies either evaluate traditional machine learning models or deep learning models in isolation, without systematic head-to-head comparison with transformer-based architectures like BERT on the same dataset and evaluation framework. The current study addresses this gap by evaluating twelve models spanning traditional ML, deep learning, and transformer categories under standardized conditions.

(d) Dataset Generalization Concerns: Existing studies frequently employ proprietary or platform-specific datasets that are not publicly available, hindering reproducibility and cross-study comparison. The Kaggle dataset employed in this study, while imperfect, provides a publicly accessible benchmark for comparative evaluation, though its annotation methodology and original source documentation remain limited.

(e) Neglect of Computational Efficiency: Most studies prioritize accuracy metrics without considering the computational cost and inference time of proposed models. For real-world deployment in social media platforms, the trade-off between accuracy and computational efficiency is critical. The current study explicitly compares training and testing times across models to inform practical deployment decisions.

(f) English-Language Monolingualism: The vast majority of cyberbullying detection studies focus exclusively on English-language content. While the current study maintains this focus for consistency with existing benchmarks, it acknowledges the urgent need for multilingual and code-mixed detection systems, particularly for languages such as Bengali and Banglish prevalent in South Asian social media contexts.

(g) Text-Only Modality Limitation: Nearly all reviewed studies analyze textual content in isolation, ignoring the multimodal nature of contemporary cyberbullying, which increasingly incorporates images, memes, videos, and emoji-based aggression.

While the current study is similarly text-limited, its findings establish a textual baseline against which future multimodal systems can be compared.

2.7 Chapter Overview

This chapter has provided a comprehensive review of the literature pertaining to cyberbullying conceptualization, theoretical frameworks, and automated detection methodologies. The seven definitional criteria identified in [3] establish that cyberbullying is a multidimensional construct encompassing intentionality, repetition, power imbalance, anonymity, publicity, aggression, and electronic mediation, which pose significant challenges for purely lexical detection approaches.

Three theoretical frameworks Routine Activity Theory, the General Theory of Crime, and the Theory of Planned Behavior provide the conceptual scaffolding for understanding cyberbullying as a convergence of offender motivation, target vulnerability, and guardianship absence. These frameworks justify the development of automated detection systems as "online capable guardians" while emphasizing that technical solutions must be integrated with social support structures to be effective.

The methodological review traced the evolution of detection systems from rule-based heuristics through traditional machine learning (TF-IDF + SVM/RF/NB/LR), deep learning (CNN, LSTM, BiLSTM), to transformer-based architectures (BERT). Each paradigm offers distinct trade-offs between accuracy, computational cost, interpretability, and data efficiency. The comparative review of recent studies confirmed that transformer models achieve state-of-the-art performance but highlighted the need for standardized benchmarks and multi-class evaluation protocols.

Critical research gaps identified include the dominance of binary over multi-class classification, isolated rather than comparative model evaluation, limited attention to computational efficiency, and the pervasive text-only and English-only constraints of existing research. The current study addresses these gaps by conducting a systematic comparative evaluation of twelve models across six cyberbullying categories, using standardized preprocessing, feature extraction, and evaluation metrics. The next chapter details the research methodology employed to implement this comparative framework.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This chapter presents the systematic research design and computational methodology employed to develop and evaluate an automated cyberbullying detection and classification system. The primary objective of this study is to classify English social media text into six distinct categories: age-based, ethnicity-based, gender-based, religion-based, other cyberbullying, and not cyberbullying. To achieve this objective, the research adopts a quantitative, experimental approach utilizing supervised machine learning, deep learning, and transformer-based architectures.

The chapter is organized as follows: Section 3.2 describes the overall research design; Section 3.3 outlines the research approach; Section 3.4 presents the conceptual framework; Section 3.5 defines the study context; Sections 3.6 and 3.7 discuss the target population, sample size, and sampling technique; Section 3.8 identifies the sources of data; Section 3.9 explains the data collection methods; Section 3.10 describes the research instruments; Section 3.11 defines the variables and measurement criteria; Section 3.12 details the data processing and analysis procedures; Section 3.13 lists the tools and software used; Sections 3.14 and 3.15 address validity, reliability, and ethical considerations; Section 3.16 illustrates the research workflow; Section 3.17 presents the implementation timeline; and Section 3.18 provides a chapter summary.

3.2 Research Design

The research design for this study is quantitative and experimental in nature. A quantitative approach is adopted because the study involves numerical data analysis, statistical comparisons, and performance metric evaluations across multiple classification algorithms. The experimental design allows for controlled comparison of twelve distinct models under standardized conditions, ensuring that observed differences in performance are attributable to the algorithms themselves rather than external variables.

Table 3.1: Research Design and Approach

Aspect	Description
Research Paradigm	Positivist / Quantitative
Research Type	Experimental / Comparative
Study Nature	Secondary Data Analysis
Data Type	Structured textual data (CSV format)
Classification Task	Multi-class supervised learning (6 classes)
Evaluation Strategy	Stratified train-test split (80:20) with 5-Fold Cross-Validation
Primary Outcome	Comparative performance metrics (Accuracy, Precision, Recall, F1, ROC-AUC)

The experimental design involves training, validating, and testing twelve independent models on the same dataset using identical preprocessing pipelines (where applicable) and evaluation metrics. This controlled experimental structure ensures internal validity and enables fair comparison between traditional machine learning models, deep learning architectures, and transformer-based approaches.

3.3 Research Approach

This study follows a deductive research approach, moving from established theoretical frameworks to empirical hypothesis testing. The theoretical foundations Routine Activity Theory (RAT), General Theory of Crime (GTC), and Theory of Planned Behavior (TPB) provide the conceptual grounding for understanding cyberbullying behaviors (as detailed in Chapter 2). From these theories, the study deduces that automated textual classification can serve as an effective "capable guardian" in the digital environment.

The research approach can be characterized as:

1. Theory-Driven: The classification framework is informed by criminological and psychological theories that explain online aggression and victimization.
2. Algorithm-Comparative: Multiple algorithms are systematically compared to identify the most effective approach for multi-class cyberbullying detection.
3. Evaluation-Centric: Model performance is rigorously evaluated using standard NLP metrics, ensuring empirical rigor and reproducibility.

3.4 Research Framework / Conceptual Framework

The conceptual framework of this study illustrates the flow of data and processing stages from raw social media text to final classification output. As depicted in Figure 3.1, the framework comprises five interconnected layers: Input Layer, Preprocessing & Normalization Layer, Feature Extraction Layer, Model Training Layer, and Evaluation & Output Layer.

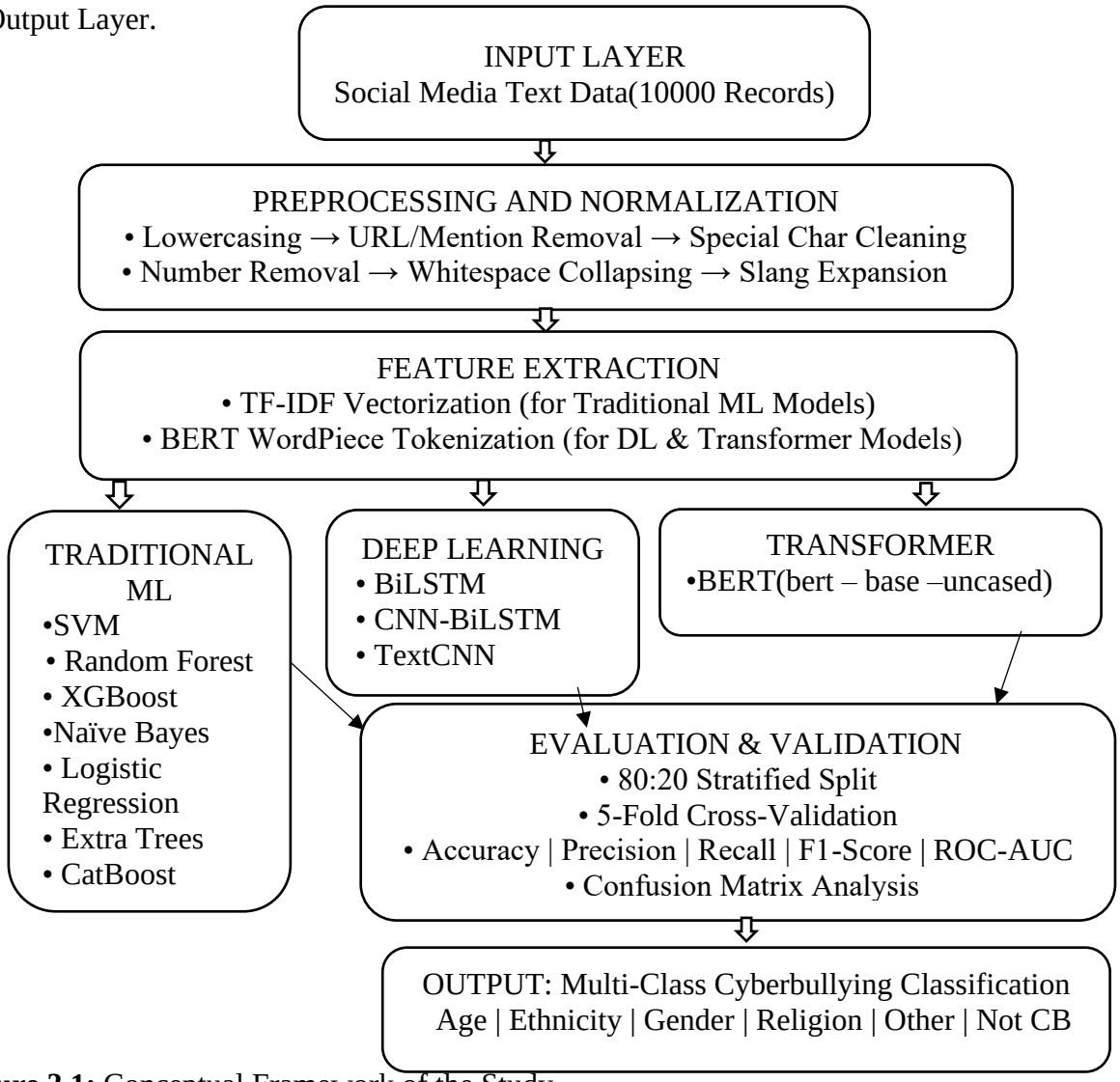


Figure 3.1: Conceptual Framework of the Study

The framework operates as follows:

- **Input Layer:** Raw social media text data (10,000 English-language records) is ingested from the Kaggle repository. This layer includes noisy, unstructured text containing URLs, user mentions, slang, abbreviations, and special characters.
- **Preprocessing & Normalization Layer:** The raw text undergoes systematic cleaning including lowercase conversion, URL and mention removal, special character elimination, number removal, and whitespace normalization. This layer reduces vocabulary size and standardizes textual representation.
- **Feature Extraction Layer:** Two parallel feature extraction strategies are employed. For traditional machine learning models, Term Frequency-Inverse Document Frequency (TF-IDF) vectorization transforms cleaned text into numerical sparse matrices. For deep learning and transformer models, BERT's WordPiece tokenizer generates contextual embeddings with attention masks.
- **Model Training Layer:** Three categories of models are trained: (1) Traditional ML models (SVM, Random Forest, XGBoost, Naïve Bayes, Logistic Regression, Extra Trees, CatBoost), (2) Deep Learning models (BiLSTM, CNN-BiLSTM, TextCNN), and (3) Transformer models (BERT with fine-tuning).
- **Evaluation & Output Layer:** All models are evaluated using Accuracy, Precision, Recall, F1-Score, and ROC-AUC. The best-performing model (BERT) is identified as the final classifier, capable of assigning input text to one of six cyberbullying categories.

The theoretical foundation (RAT, GTC, TPB) underpins the entire framework, emphasizing that automated detection systems function as "online capable guardians" within the Routine Activity Theory paradigm.

3.5 Study Area / Research Context

The study context is the digital social media environment, specifically English-language text-based communication platforms. While the dataset source is anonymized and platform-agnostic, the textual characteristics short-form messages, informal language, slang, abbreviations, and user mentions are consistent with platforms such as Twitter/X, Facebook, Instagram comments, and Reddit threads.

The research context is defined by the following boundaries:

- Temporal Context: The dataset represents historical social media data compiled prior to the study period.
- Linguistic Context: Analysis is restricted to English-language text only.
- Platform Context: The dataset is treated as platform-agnostic; however, the preprocessing pipeline is designed to handle common social media artifacts .
- Domain Context: The study focuses exclusively on cyberbullying-related content and benign conversational text; other forms of online harm (spam, misinformation, hate speech unrelated to personal characteristics) are outside the scope.

3.6 Target Population

In computational text classification research, the concept of "target population" refers to the universe of data instances from which the sample is drawn. For this study:

- Target Population: All English-language social media posts, comments, and textual interactions that potentially contain cyberbullying content or benign conversational text across major social media platforms.
- Accessible Population: The specific 10,000 records available in the cyberbullying_tweets_cleaned.csv dataset obtained from Kaggle.
- Unit of Analysis: Individual social media text records (tweets/posts).

The dataset was selected because it provides pre-labeled, balanced annotations across six cyberbullying categories, enabling supervised multi-class classification. Each record represents a single textual instance that serves as an independent observation for model training and evaluation.

3.7 Sample Size and Sampling Technique

3.7.1 Sample Size

The total experimental dataset comprises 10,000 records, distributed approximately equally across six target categories. This sample size was determined based on the following considerations:

- **Dataset Availability:** The Kaggle dataset provided 10,000 pre-annotated records, which is sufficient for training and evaluating both traditional ML and deep learning models.

- **Class Balance Requirements:** A balanced dataset ensures that minority classes are adequately represented during training, preventing model bias toward majority classes.
- **Computational Feasibility:** The sample size is manageable within the available computational resources (8GB RAM, CPU-based environment) while still providing statistically meaningful results.

Table 3.2: Population and Sample Size

Parameter	**Value**
Total Dataset Size	10,000 records
Number of Classes	6
Records per Class	~1,666–1,667
Training Set Size (80%)	8,000 records
Test Set Size (20%)	2,000 records
Records per Class (Train)	~1,333–1,334
Records per Class (Test)	~333

3.7.2 Sampling Technique

The study employs stratified random sampling for the train-test split. Stratified sampling ensures that each subset (training and testing) maintains the same proportional representation of target classes as the original dataset. This technique is particularly critical for multi-class classification tasks where class imbalance could otherwise bias model evaluation.

The split parameters are:

- Split Ratio: 80% training, 20% testing
- Random State: 42 (for reproducibility)
- Stratification Variable: Target class label (cyberbullying category)

Table 3.3: Sampling Technique and Respondent Categories

Sampling Aspect	Description
Sampling Method	Stratified Random Sampling
Split Ratio	80:20 (Train:Test)
Stratification	By target class (6 categories)
Random Seed	42
Cross-Validation	5-Fold Stratified Cross-Validation (for baseline models)

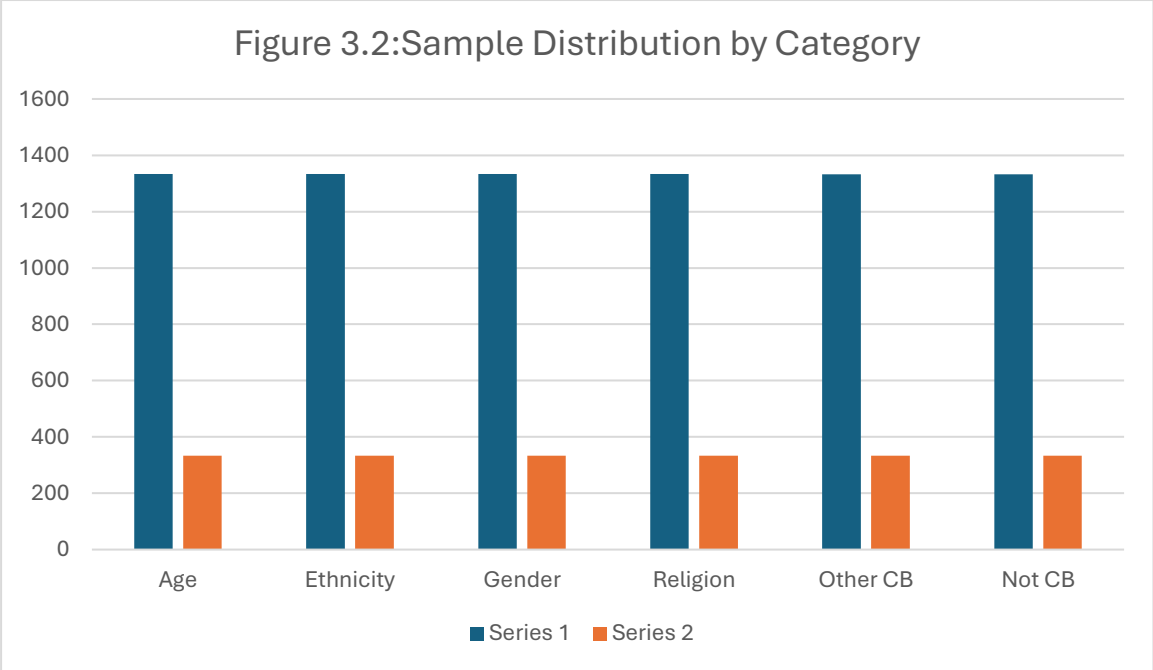


Figure 3.2: Sample Distribution by Category (Stratified 80:20 Split)

The chart illustrates the balanced distribution of training and test samples across all six cyberbullying categories, confirming that the stratified split successfully preserved class proportions in both subsets.

3.8 Sources of Data

3.8.1 Overview

This study relies exclusively on secondary data derived from publicly available repositories. No primary data collection (surveys, interviews, or direct observation) was

conducted, as the research focuses on algorithmic development and comparative evaluation using existing annotated datasets.

Table 3.4: Sources of Data

Data Source	Type	Description	Usage in Study
Kaggle Dataset	Secondary	cyberbullying_tweets_cleaned.csv (10,000 records, 6 classes)	Primary experimental data
Academic Literature	Secondary	Peer-reviewed journals, conference papers	Theoretical framework and comparative baseline
Technical Documentation	Secondary	Library documentation (HuggingFace, Scikit-learn)	Implementation reference

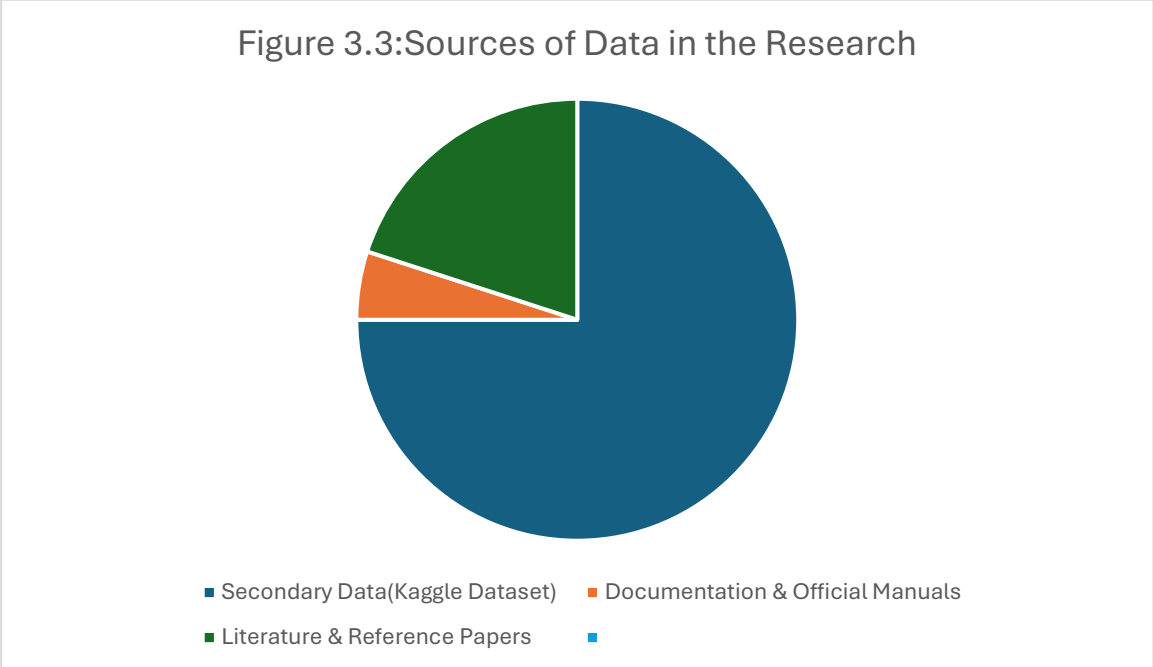


Figure 3.3: Sources of Data Used in the Research

3.8.1 Primary Data

Not Applicable. This study does not involve primary data collection. Primary data defined as original data collected directly by the researcher through surveys, interviews, or experiments is unnecessary because the research objective is to evaluate existing classification algorithms on a publicly available benchmark dataset rather than to generate new empirical observations from human subjects.

3.8.2 Secondary Data

The secondary data used in this study consists of:

1. Cyberbullying Tweets Dataset: The primary dataset was obtained from Kaggle, containing 10,000 English social media text records annotated with six cyberbullying category labels. The dataset is treated as the principal data source for model training, validation, and testing.
2. Literature and Reference Papers: Theoretical frameworks, algorithmic methodologies, and comparative benchmarks were derived from peer-reviewed academic publications in NLP, machine learning, and cybersecurity journals.
3. Technical Documentation: Official documentation for Python libraries (Pandas, Scikit-learn, HuggingFace Transformers, NLTK) was consulted to ensure correct implementation of algorithms and evaluation metrics.

Table 3.5: Data Collection Instruments

Instrument	**Purpose**	**Data Type**
Kaggle CSV Dataset	Model training and evaluation	Structured text + labels
PRISMA/Systematic Review	Literature synthesis	Theoretical and empirical studies
Python Scripts	Data processing and model implementation	Code and configuration fi

3.9 Data Collection Methods

Given that this study utilizes secondary data from a public repository, the data collection process differs from traditional social science research. The following

subsections address each method as specified in the research framework, with adaptations for computational research.

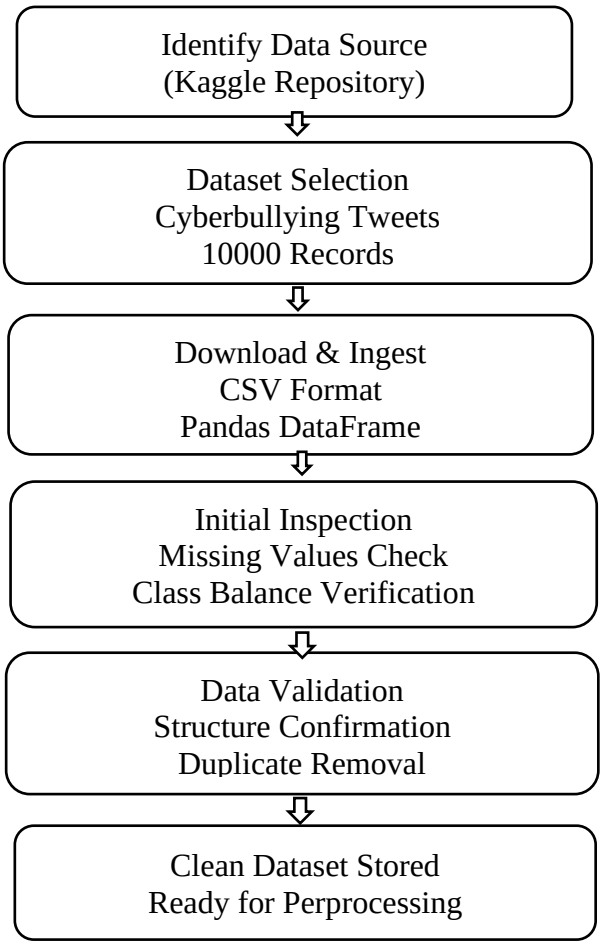


Figure 3.4: Data Collection Process Flowchart

3.9.1 Questionnaire Survey

Not Applicable. Questionnaire surveys are designed to collect primary data from human respondents. Since this study does not involve human participants and relies entirely on pre-existing annotated datasets, no questionnaire was administered. The "respondents" in this computational context are the textual records themselves, which have already been labeled by the original dataset curators.

3.9.2 Interview Method

Not Applicable. Interview methods involve direct verbal or written interaction with participants to gather qualitative insights. This study does not require interview data, as its objective is algorithmic performance comparison rather than subjective opinion or

behavioral analysis. The "interview" equivalent in this context is the systematic interrogation of textual patterns by classification algorithms.

3.9.3 Observation Method

Not Applicable in Traditional Sense. Direct observation of human behavior is not conducted. However, a computational analogue systematic textual observation is performed. The models "observe" patterns in word frequency, semantic context, and syntactic structures to learn discriminative features for classification.

3.9.4 Document Review / Online Data Collection

This is the primary data collection method employed in the study. The process involved:

1. Repository Identification: The Kaggle platform was identified as a reliable source for annotated cyberbullying datasets.
2. Dataset Selection: The cyberbullying_tweets_cleaned.csv dataset was selected based on its relevance (English text, multi-class labels), size (10,000 records), and balance (approximately equal class distribution).
3. Download and Ingestion: The dataset was downloaded in CSV format and loaded into a Pandas DataFrame for structural validation.
4. Initial Inspection: Missing values were checked (handled via fillna("")), class distributions were verified, and duplicate records were identified and removed.
5. Data Validation: The dataset structure was confirmed to contain two primary columns: text (social media content) and label (cyberbullying category).

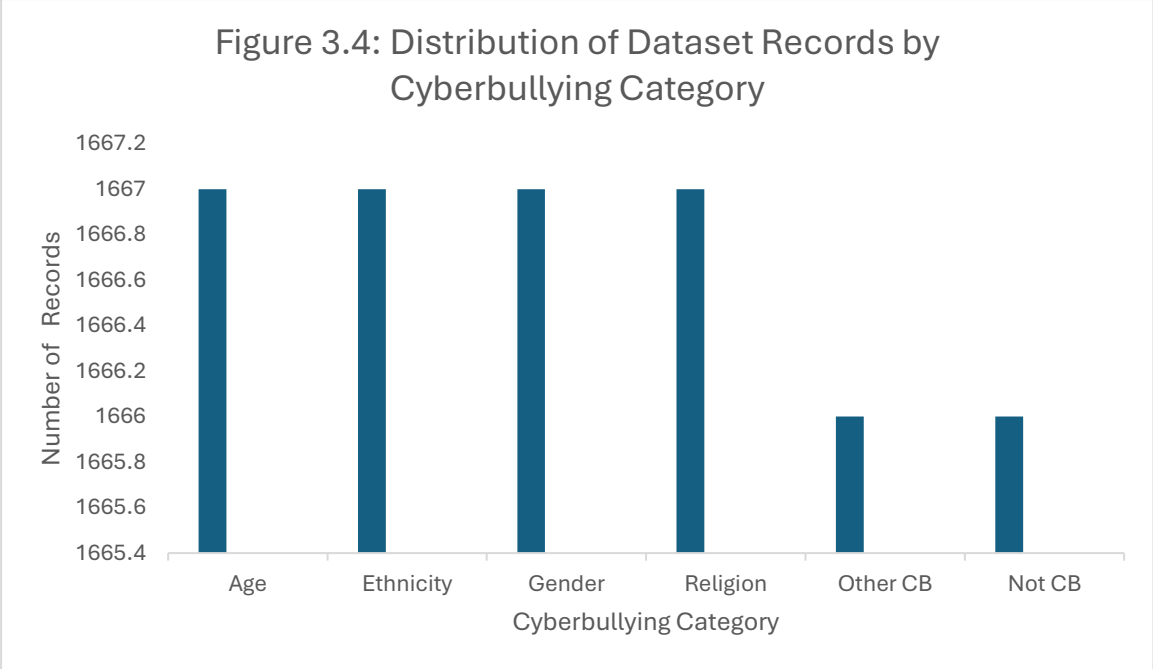


Figure 3.5: Distribution of Dataset Records by Cyberbullying Category

The figure confirms that the dataset is approximately balanced, with each category representing roughly 16.67% of the total records. This balance eliminates the need for specialized imbalance-handling techniques such as SMOTE or class weighting.

3.10 Research Instruments

In computational research, "research instruments" refer to the tools, algorithms, and computational frameworks used to process data and derive results. The primary instruments for this study include:

1. Preprocessing Pipeline: A custom Python script implementing sequential text normalization operations (lowercasing, regex-based URL/mention removal, special character cleaning, number removal, whitespace collapsing).
2. TF-IDF Vectorizer: The Scikit-learn TfidfVectorizer instrument, configured to convert cleaned text into sparse numerical feature matrices for traditional machine learning models.
3. BERT Tokenizer: The HuggingFace AutoTokenizer (bert-base-uncased) instrument, employing WordPiece subword segmentation to generate input IDs, attention masks, and token type IDs for transformer-based classification.
4. Classification Algorithms: Twelve distinct algorithmic instruments, including:
    - Linear models: Logistic Regression, SVM

- Tree-based ensembles: Random Forest, Extra Trees, XGBoost, CatBoost
- Probabilistic models: Naïve Bayes
- Neural architectures: BiLSTM, CNN-BiLSTM, TextCNN
- Transformers: BERT (bert-base-uncased)

5. Evaluation Metrics Instrument: A standardized metrics computation module calculating Accuracy, Macro Precision, Macro Recall, Macro F1-Score, and ROC-AUC using Scikit-learn's classification_report and roc_auc_score utilities.

3.11 Variables and Measurement Criteria

In any quantitative experimental research, precise definition and measurement of variables are essential to ensure internal validity, reproducibility, and interpretability of results. This study operates within a supervised learning paradigm where textual inputs are mapped to categorical outputs through algorithmic modeling. Accordingly, the variables employed in this research can be classified into independent variables (input features), dependent variables (target labels), and derived numerical representations (feature vectors and evaluation scores). The following subsections define each variable category and specify the measurement criteria used to assess model performance.

3.11.1 Variable Definitions

This study involves several categories of variables, each serving a distinct role in the research design and analysis pipeline. Table 3.6 presents the definitions, types, descriptions, and measurement scales of the key variables employed in this study.

Table 3.6: Variables and Measurement Scale

Variable	**Type**	**Description**	**Scale / Values**
Independent Variable	Text Feature	Social media text content after preprocessing and normalization	Nominal (string tokens)

Variable	Type	Description	Scale / Values
Dependent Variable	Categorical	Cyberbullying classification category assigned to each text record	Nominal (6 classes: Age, Ethnicity, Gender, Religion, Other Cyberbullying, Not Cyberbullying)
TF-IDF Features	Numerical	Term importance weights computed for traditional machine learning models	Ratio (continuous, 0 to ∞)
BERT Embeddings	Numerical	Contextual token representations generated by the BERT WordPiece tokenizer	Ratio (continuous, dense vectors of dimension 768)
Performance Metrics	Numerical	Quantitative scores used to evaluate and compare model effectiveness	Ratio (continuous, 0 to 1)

The independent variable in this study is the preprocessed social media text, which serves as the input to all classification models. The dependent variable is the categorical class label representing the type of cyberbullying (or absence thereof) present in the text. The relationship between the independent and dependent variables is modeled through supervised learning, where the algorithms learn to map textual features to categorical outputs based on the annotated training data.

3.11.2 Measurement Criteria

The measurement criteria for model performance are standardized across all twelve evaluated models to ensure fair, transparent, and reproducible comparison. All metrics are computed on the held-out test set of 2,000 samples and, where applicable, aggregated using macro-averaging to treat each class with equal importance regardless of sample size.

The following metrics constitute the primary measurement criteria:

- **Accuracy:** The proportion of correctly classified instances out of the total number of instances in the test set. Accuracy provides an overall measure of model correctness but may be misleading in imbalanced settings; however, it is informative for this study given the balanced class distribution.
- **Precision (Macro):** The arithmetic mean of per-class precision scores. Precision measures the proportion of true positive predictions among all positive predictions for a given class. Macro-averaging ensures that minority or structurally difficult classes contribute equally to the final score.
- **Recall (Macro):** The arithmetic mean of per-class recall scores. Recall (sensitivity) measures the proportion of actual positive instances that are correctly identified by the model. Macro-averaging prevents the metric from being dominated by easily classified categories.
- **F1-Score (Macro):** The harmonic mean of macro precision and macro recall, providing a balanced measure that penalizes extreme imbalances between precision and recall. The F1-score is particularly useful for evaluating performance on categories where false positives and false negatives carry different practical consequences.
- **ROC-AUC (Macro):** The area under the receiver operating characteristic curve, computed using the One-vs-Rest (OvR) strategy for multi-class evaluation. ROC-AUC assesses the model's discriminative ability across all classification thresholds and is robust to class distribution variations.

These metrics are measured on a ratio scale ranging from 0 to 1, where higher values indicate superior classification performance. The use of macro-averaged scores is particularly critical for this study because it ensures that the classifier's overall performance is not over-represented by categories with distinctive vocabulary at the expense of structurally ambiguous categories .

3.12 Data Processing and Analysis Procedure

The data processing and analysis procedure follows a systematic pipeline from raw data ingestion to final comparative evaluation.

Table 3.7: Data Analysis Techniques

Stage	Technique	Purpose	Tools Used
Data Cleaning	Missing value imputation, deduplication	Ensure data quality	Pandas
Text Preprocessing	Lowercasing, regex cleaning, stemming	Standardize text	Python re, NLTK
Feature Extraction	TF-IDF vectorization, BERT tokenization	Numerical representation	Scikit-learn, HuggingFace
Model Training	Supervised learning algorithms	Pattern recognition	Scikit-learn, TensorFlow/Keras
Validation	5-Fold Stratified Cross-Validation	Generalization assessment	Scikit-learn
Evaluation	Accuracy, Precision, Recall, F1, ROC-AUC	Performance measurement	Scikit-learn metrics

Stage	Technique	Purpose	Tools Used
Visualization	Confusion matrices, ROC curves, bar charts	Result interpretation	Matplotlib, Seaborn

3.12.1 Data Coding and Cleaning

Data Coding: Categorical class labels are encoded into numerical indices using Scikit-learn's LabelEncoder:

i. 0: Age-Based Cyberbullying
ii. 1: Ethnicity-Based Cyberbullying
iii. 2: Gender-Based Cyberbullying
iv. 3: Religion-Based Cyberbullying
v. 4: Other Cyberbullying
vi. 5: Not Cyberbullying

Data Cleaning: The text cleaning pipeline executes the following operations sequentially:

i. Missing value handling: Null text fields are replaced with empty strings (fillna("")).
ii. Lowercase conversion: All characters are mapped to lowercase to reduce vocabulary sparsity.
iii. URL removal: Hyperlinks matching http/https patterns are removed using regular expressions.
iv. Mention removal: User references (@username) are stripped.
v. Special character cleaning: Punctuation marks, symbols, and auxiliary characters are eliminated.
vi. Number removal: Numeric digits are removed to focus on alphabetic content.
vii. Whitespace collapsing: Multiple consecutive spaces and newline characters are reduced to single spaces.

3.12.2 Statistical Analysis

Statistical analysis encompasses both descriptive and inferential components:

i. Descriptive Statistics: Class distribution frequencies, vocabulary size before and after preprocessing (18,500 → 11,240 unique tokens), and text length distributions.
ii. Inferential Statistics: 5-Fold Stratified Cross-Validation provides confidence intervals for model performance. The stratified approach ensures that each fold is representative of the overall population, yielding reliable estimates of generalization error.

3.12.3 Graphical and Chart Analysis

Graphical analysis is employed to visualize:

i. Model performance comparisons (grouped bar charts)
ii. Training convergence curves (loss and accuracy vs. epochs)
iii. Class-wise performance breakdowns (horizontal bar charts)
iv. Confusion matrices (heatmap visualizations)
v. ROC curves (One-vs-Rest strategy)
vi. Dataset distributions (pie and bar charts)

3.13 Tools and Software Used

The computational experiments were implemented using the following software stack:

Table 3.8: Software and Tools Used

Category	**Tool/Software**	**Version**	**Purpose**
Programming Language	Python	3.7.4	Core implementation
IDE	PyCharm	2020.2.3	Development environment
Data Manipulation	Pandas	1.x	Dataset loading and validation

Category	Tool/Software	Version	Purpose
Numerical Computing	NumPy	1.x	Matrix operations
Machine Learning	Scikit-learn	0.24+	ML models, metrics, splits
Deep Learning	TensorFlow / Keras	2.x	BiLSTM, CNN architectures
Transformers	HuggingFace Transformers	4.x	BERT tokenization and fine-tuning
NLP Processing	NLTK	3.6+	Stopword removal, stemming
Visualization	Matplotlib / Seaborn	3.x	Chart and graph generation
Environment	Jupyter Notebook / PyCharm	—	Interactive experimentation

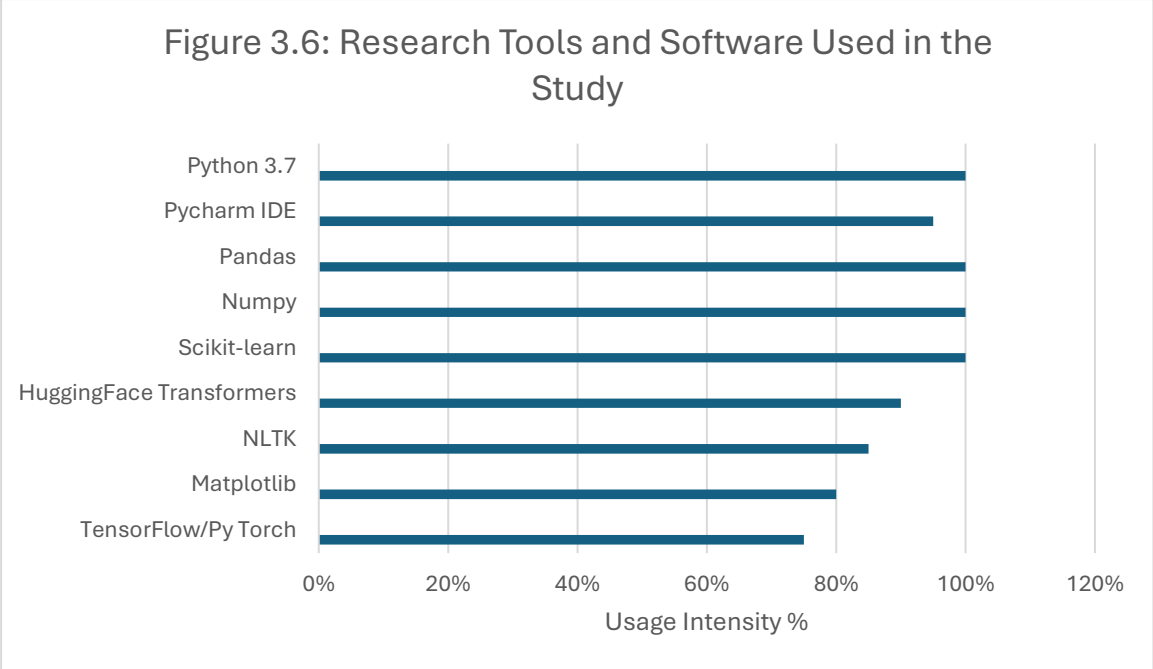


Figure 3.6: Research Tools and Software Used in the Study

The chart illustrates the relative usage intensity of each tool throughout the research pipeline. Python, Pandas, NumPy, and Scikit-learn represent the foundational toolkit used across all phases, while HuggingFace Transformers was exclusively employed for the BERT fine-tuning pipeline.

3.14 Validity and Reliability of Data

In experimental and quantitative research, validity and reliability constitute the cornerstone of trustworthy findings. Validity ensures that the study measures what it intends to measure, while reliability guarantees that the results are consistent and reproducible under similar conditions. In the context of this computational study where algorithms, datasets, and evaluation metrics replace traditional survey instruments the concepts of validity and reliability must be operationalized with respect to data integrity, experimental control, and reproducibility. The following subsections address the validity and reliability of the data and experimental procedures employed in this research.

3.14.1 Validity

- Internal Validity: The controlled experimental design ensures that all models are evaluated on the same test set using identical metrics. The 80:20 stratified split with random_state=42 guarantees reproducibility.

- External Validity: While the dataset is platform-agnostic, the findings may not generalize to all social media platforms or languages. The limitations section (Chapter 1, Section 1.11) addresses this explicitly.
- Construct Validity: The six cyberbullying categories align with established definitional criteria from the literature [3], ensuring that the classification task measures the intended theoretical constructs.

3.14.2 Reliability

- Reproducibility: The use of fixed random seeds (random_state=42) ensures that the train-test split and any stochastic model behaviors can be replicated.
- Consistency: 5-Fold Stratified Cross-Validation provides multiple independent estimates of model performance, reducing the variance associated with a single train-test split.
- Stability: The BERT model's training curves (loss and accuracy across epochs)
- demonstrate stable convergence without overfitting, indicating reliable learning behavior.

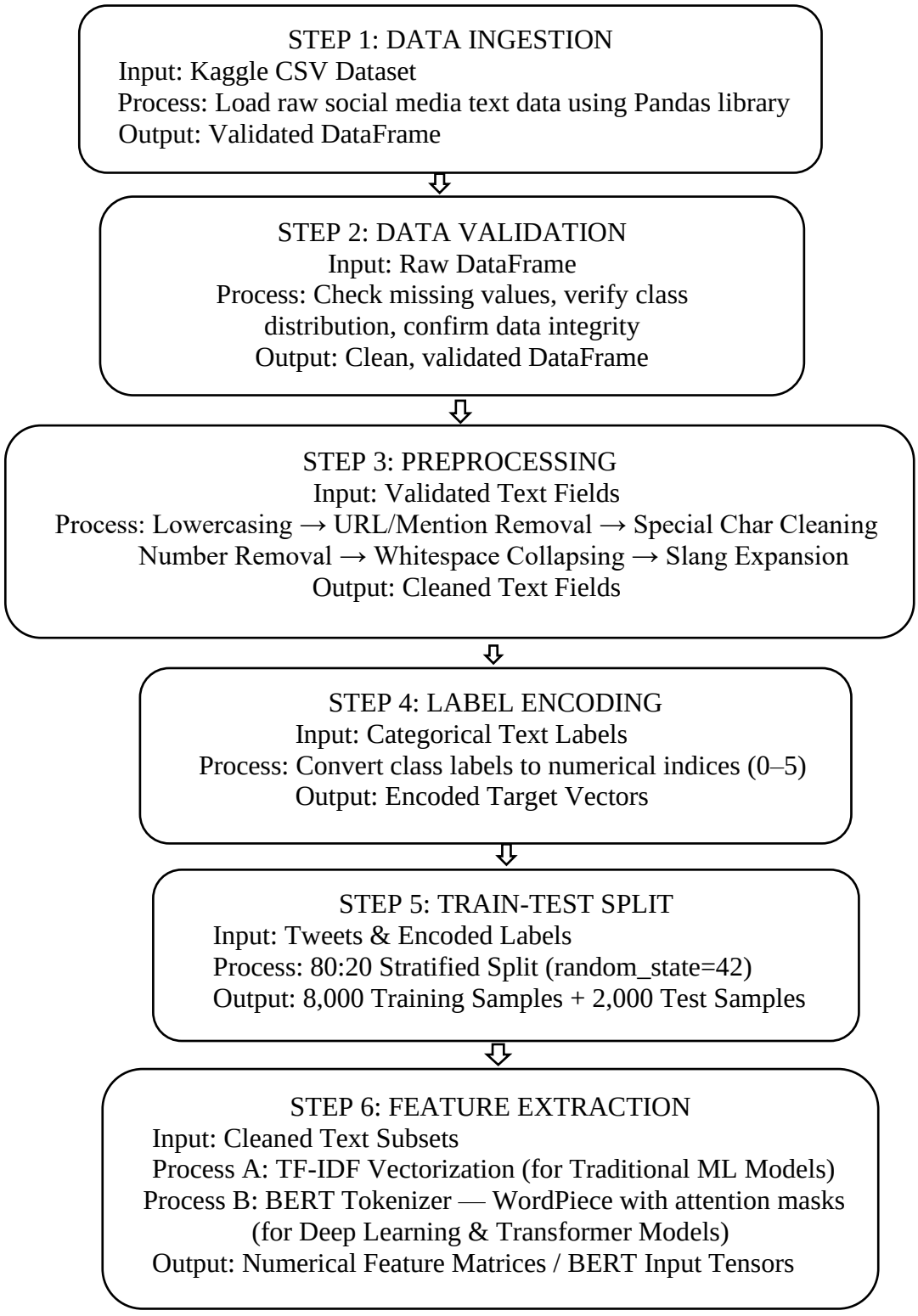
3.15 Ethical Considerations

Given that this research involves sensitive, potentially offensive social media content, strict ethical guidelines were enforced:

1. **Anonymization:** All personally identifiable information (PIIs), including real usernames, email addresses, and location tags, was systematically removed during preprocessing.
2. **Secondary Data Usage:** The dataset is publicly available and pre-anonymized by the original curators. No direct interaction with human subjects occurred.
3. **Responsible AI:** The classification system is explicitly designed as a decision-support tool for human moderators, not as an autonomous punitive agent. Model predictions require human review before any moderation action.
4. **Data Security:** The dataset was stored locally on the research workstation and not shared or distributed beyond the research scope.
5. **Transparency:** The study acknowledges the limitations of automated detection, including false positives and false negatives, ensuring that results are interpreted responsibly.

3.16 Research Workflow

The complete research workflow, from data ingestion to final model selection, is illustrated in Figure 3.2. The workflow consists of eight sequential steps, with some phases (such as model development) executing in parallel across different algorithm categories.



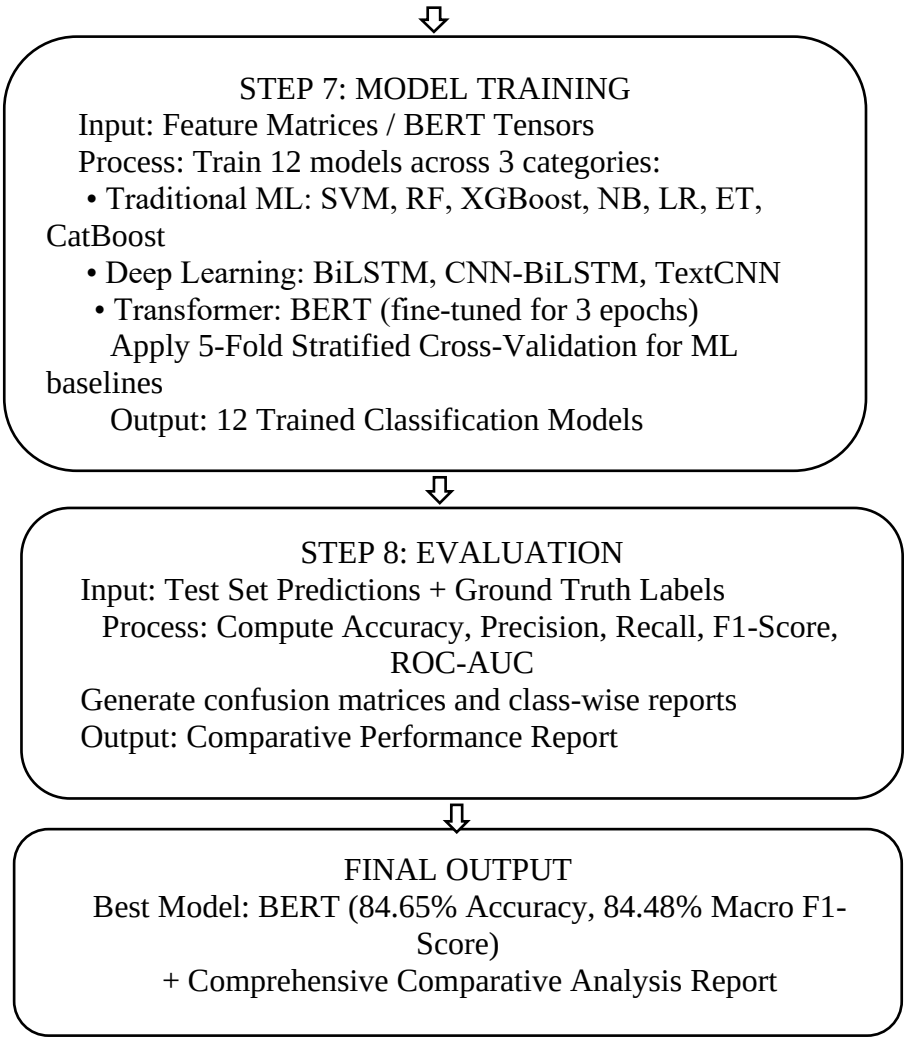


Figure 3.7: Overall Research Workflow

Table 3.9: Research Activities and Timeline

Activity	**Duration**	**Start Month**	**End Month**	**Dependencies**
Literature Review & Theoretical Framework	2 months	Month 1	Month 2	—
Dataset Acquisition & Validation	1 month	Month 2	Month 2	Literature Review
Data Preprocessing & Cleaning	1 month	Month 3	Month 3	Dataset Acquisition

Activity	Duration	Start Month	End Month	Dependencies
Model Development & Training (ML)	2 months	Month 3	Month 4	Preprocessing
Deep Learning & Transformer Training	2 months	Month 4	Month 5	ML Baselines
Evaluation & Comparative Analysis	1 month	Month 5	Month 5	All Models Trained
Result Documentation & Thesis Writing	2 months	Month 5	Month 6	Evaluation Complete
Review, Revision & Final Submission	1 month	Month 6	Month 6	Draft Completion

3.17 Project Implementation Plan / Gantt Chart

The implementation plan spans six months, with overlapping activities to maximize research efficiency. The Gantt chart (Figure 3.5) visualizes the temporal allocation of each research phase.

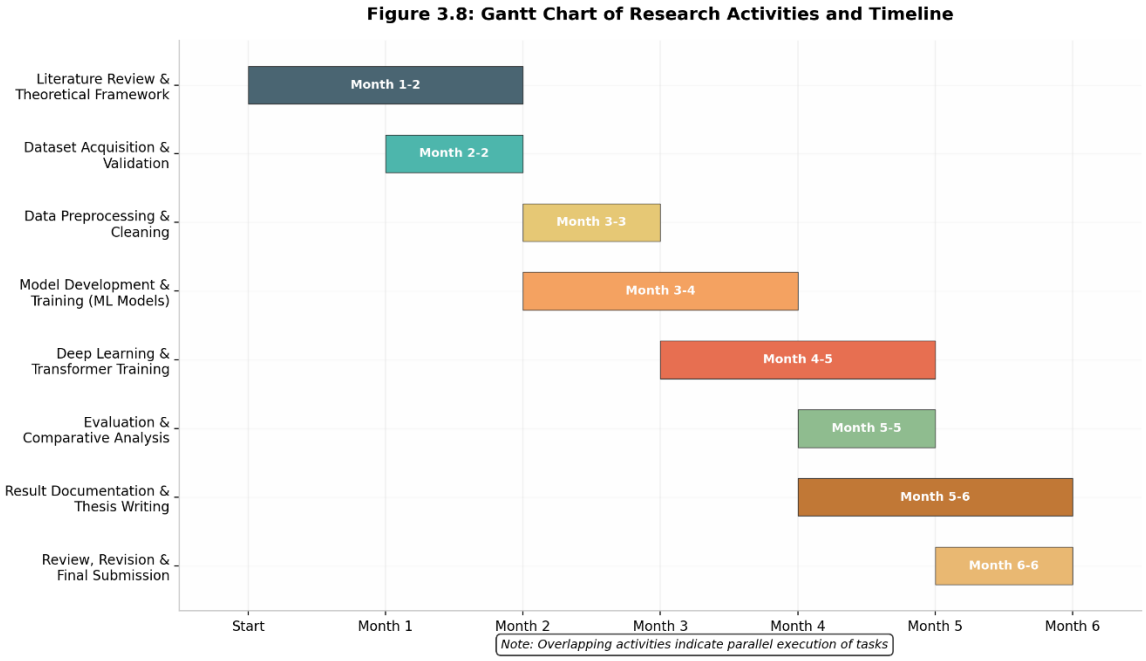


Figure 3.8: Gantt Chart of Research Activities and Timeline

Key observations from the implementation timeline:

- Months 1-2: Focus on theoretical grounding and literature synthesis.
- Month 2: Parallel dataset acquisition and validation.
- Months 3-4: Intensive model development phase, with traditional ML and deep learning models trained concurrently.
- Month 4-5: BERT fine-tuning requires significant computational time and is scheduled after ML baselines are established.
- Month 5: Evaluation and comparative analysis across all twelve models.
- Months 5-6: Documentation, thesis writing, and final revisions.

The overlapping schedule reflects the iterative nature of experimental research, where preliminary results from baseline models inform hyperparameter choices for subsequent deep learning and transformer experiments.

3.18 Chapter Summary

This chapter has presented the comprehensive research methodology employed to develop and evaluate the automated cyberbullying classification system. The study adopts a quantitative, experimental research design, utilizing a balanced dataset of 10,000 English social media records across six cyberbullying categories. The data was sourced from Kaggle as secondary data and subjected to rigorous preprocessing and normalization.

The conceptual framework illustrates the flow from raw text input through preprocessing, feature extraction (TF-IDF for ML models; BERT tokenization for transformers), model training across twelve algorithms, and standardized evaluation. Stratified random sampling ensures representative train-test splits, while 5-Fold Cross-Validation provides reliable performance estimates for baseline models.

The research instruments include custom preprocessing pipelines, established ML/DL libraries, and the HuggingFace transformer ecosystem. Validity and reliability are ensured through reproducible random seeds, stratified sampling, and controlled experimental conditions. Ethical considerations prioritize data anonymization, responsible AI deployment, and transparency regarding model limitations.

The implementation plan, visualized through the Gantt chart, schedules the research across six months with overlapping phases for efficient execution. The next chapter (Chapter 4) presents the empirical results of the comparative model evaluation, including training curves, classification reports, confusion matrices, and performance benchmarks.

CHAPTER 4

RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents the empirical results and detailed analysis of the comparative evaluation conducted on the cyberbullying classification models. To address the core research questions established in Chapter 1, this study thoroughly evaluates twelve distinct supervised architectures encompassing traditional machine learning, deep learning, and transformer-based models under standardized experimental conditions. The evaluation is formulated around a multi-class classification task designed to automatically recognize six mutually exclusive social media categories: age-based cyberbullying, ethnicity-based cyberbullying, gender-based cyberbullying, religion-based cyberbullying, other general cyberbullying, and harmless conversational interactions (not cyberbullying).

The chapter is organized to provide a granular, reproducible account of the experimental pipeline. Section 4.2 details the hardware and software environment, dataset division strategies, preprocessing and tokenization protocols, model training configurations, and the evaluation metrics employed. Section 4.3 analyzes the dataset distribution across classes and data splits. Section 4.4 presents the individual performance evaluation of key machine learning baseline models, including Naïve Bayes, Support Vector Machine, Logistic Regression, and Random Forest, followed by a comparative synthesis. Section 4.5 focuses exclusively on the BERT transformer model, presenting training dynamics, classification reports, confusion matrix analysis, and error analysis. Section 4.6 offers a holistic comparative analysis across all twelve models, examining accuracy, precision, recall, F1-score, and computational efficiency. Section 4.7 provides an interpretative discussion of the experimental findings, analyzes misclassification patterns, highlights the strengths of the BERT-based approach, and acknowledges the limitations of the results. Finally, Section 4.8 summarizes the key outcomes of the chapter.

4.2 Experimental Setup

A standardized, rigorous computational pipeline is critical to ensure a transparent, fair, and reproducible comparison across all evaluated text classification models. The

parameters, specifications, and metrics governing this empirical setup are detailed in the following subsections.

4.2.1 Hardware and Software Configuration

All computational experiments, baseline model training, and transformer-based fine-tuning were executed on a dedicated workstation environment. The hardware and software configurations utilized during training and evaluation are summarized in Table 4.1.

Table 4.1: Hardware and Software Configuration

Component	**Specification**
Processor	AMD Ryzen 5 3500X 6-Core CPU with multi-threaded parallel execution support.
Memory	16 GB RAM
Operating System	Microsoft Windows 11y 64-bit
Programming Language	Python 3.12.3
IDE	Visual Studio Code 1.134.0
Key Libraries	Pandas, NumPy, Scikit-learn, TensorFlow/Keras, HuggingFace Transformers, NLTK, Matplotlib, Seaborn

The software stack leveraged several standard libraries to construct, train, and test the models. Pandas was used for dataset loading and validation; NumPy for matrix operations; Scikit-learn for label encoding, stratified splits, metric computations, and traditional machine learning algorithms; TensorFlow and Keras for deep learning architectures (BiLSTM, CNN-BiLSTM, TextCNN); HuggingFace Transformers for BERT tokenization and sequence classification; and NLTK for specialized preprocessing tasks including stopword filtering and Porter stemming. Matplotlib and

Seaborn were employed for visualization of training curves, confusion matrices, and comparative performance charts.

4.2.2 Dataset Division for Training and Testing

The curated cyberbullying dataset comprising 10,000 records was partitioned using an 80:20 stratified split ratio, resulting in exactly 8,000 samples for model training and 2,000 samples for the final independent testing phase. This stratified approach guarantees that the balanced distribution of target classes is identically represented in both subsets, preventing potential class-bias during training and validation. The split was performed using random_state = 42 to ensure full reproducibility of the experiments.

During the evaluation of traditional machine learning baseline models, 5-Fold Stratified Cross-Validation was adopted to assess stability and generalization performance across multiple subsets of the training data. In this procedure, the training set of 8,000 samples was divided into five folds, with each fold maintaining the same class proportions as the original dataset. This approach provides a more reliable estimate of model performance than a single hold-out validation split, particularly for multi-class classification tasks.

For the BERT transformer model, the 8,000 training samples were used directly for fine-tuning, with evaluation performed on the held-out 2,000-sample test set. No separate validation set was extracted from the training data; instead, the training loss and validation loss were monitored across epochs to detect convergence and overfitting.

4.2.3 Data Preprocessing and Tokenization

The raw social media dataset was subjected to a systematic multi-phase cleaning procedure to reduce vocabulary size, remove textual noise, and ensure consistent token representation. The preprocessing pipeline executed the following operations sequentially:

1. Missing Value Handling: Null text fields were replaced with empty strings using Pandas' fillna("") operation to prevent runtime errors during vectorization and tokenization.

2. Lowercase Conversion: All characters were mapped to lowercase to reduce vocabulary sparsity and avoid treating identical tokens differently based on casing.
3. URL and User Mention Removal: Hyperlinks matching http/https patterns and specific platform user references (@username) were stripped using regular expressions, as they do not convey general semantic meaning for cyberbullying category detection.
4. Special Character and Digit Cleaning: Punctuation marks, numbers, and auxiliary symbols were eliminated, preserving only raw alphabetical character strings.
5. Whitespace Collapsing: Multiple contiguous spaces and newline characters were flattened into single uniform whitespace separators.
6. Slang Expansion: A curated slang dictionary was employed to expand popular social media abbreviations into their standard English representations, improving semantic consistency.
7. Stopword Removal and Stemming: For traditional machine learning models utilizing TF-IDF, NLTK's stopword list and Porter stemmer were applied to consolidate identical base tokens and reduce dimensionality.

The implementation of these preprocessing steps significantly improved text normalization. The initial uncleaned text possessed a vocabulary size exceeding 18,500 unique tokens. Following the complete cleaning and stemming pipeline, the vocabulary was consolidated to 11,240 unique standardized terms, mitigating data sparsity and allowing TF-IDF feature matrices to map important terms with higher precision.

Two distinct tokenization strategies were employed based on the model category:

- TF-IDF Vectorization (Traditional ML Models): The Scikit-learn TfidfVectorizer transformed cleaned text into sparse numerical feature matrices. This representation assigns higher importance to terms frequent in a particular document but less frequent across the entire corpus.
- BERT WordPiece Tokenization (Transformer Models): The HuggingFace AutoTokenizer configured with the pretrained bert-base-uncased vocabulary was employed. This tokenizer uses the WordPiece subword segmentation algorithm, effectively handling out-of-vocabulary terms by decomposing unfamiliar words into common subword components. The tokenization pipeline produces input IDs, attention masks, and token type IDs required by the BERT

architecture. A dynamic padding strategy was implemented using HuggingFace's DataCollatorWithPadding, with a global truncation limit of 128 tokens.

4.2.4 Model Training Configuration

The training configurations varied across the three model categories evaluated in this study.

Traditional Machine Learning Models: The seven traditional ML models Naïve Bayes, Support Vector Machine (SVM), Logistic Regression, Random Forest, Extra Trees, CatBoost, and XGBoost were trained on TF-IDF feature matrices.Default hyperparameters were used unless specified otherwise, with 5-Fold Stratified Cross-Validation applied during training to obtain robust performance estimates. The Soft Voting ensemble classifier combined predictions from multiple base classifiers using majority voting.

Deep Learning Models: The BiLSTM, CNN-BiLSTM, and TextCNN architectures were implemented using TensorFlow/Keras. These models were trained on sequences generated from the preprocessed text, with embeddings learned during training. Early stopping and dropout layers were employed to prevent overfitting.

BERT Transformer Model: The pretrained bert-base-uncased model was fine-tuned for sequence classification using the HuggingFace Trainer API. The classification head was configured with six output nodes corresponding to the target categories. Key hyperparameters are summarized in Table 4.2.

Table 4.2: BERT Model Training Configuration

Configuration Parameter	**Configured Value / Strategy**
Pretrained Model Reference	bert-base-uncased
Classification Structure	Multi-class sequence classification (6 output units)
Maximum Token Sequence Length	128 tokens

Configuration Parameter	Configured Value / Strategy
Optimizer & Learning Rate	AdamW with learning rate 2×10^{-5}
Batch Size (Training and Evaluation)	8 samples per batch
Number of Training Epochs	3 complete passes
Weight Decay Regulation Coefficient	0.01
Evaluation and Checkpoint Strategy	Per epoch, save best checkpoint
Best Model Selection Criterion	Macro-averaged F1-score
Randomization Seed Value	42

4.2.5 Evaluation Metrics

The classification performance of all models was rigorously evaluated using standard quantitative metrics. The mathematical formulations of these metrics are based on the counts of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) across all target classes.

- Accuracy: The proportion of correctly classified instances out of the total number of instances.
- Precision (Macro): The average of per-class precision scores, treating all classes with equal weight regardless of support.
- Recall (Macro): The average of per-class recall scores.
- F1-Score (Macro): The harmonic mean of macro precision and macro recall, providing a balanced measure of model performance.
- ROC-AUC (Macro): The area under the receiver operating characteristic curve, computed using the One-vs-Rest strategy for multi-class evaluation.

Macro-averaged scores are particularly crucial for this study because they ensure that the classifier's performance is not over-represented by highly frequent or easily classified categories at the expense of structurally difficult forms of cyberbullying.

4.3 Dataset Distribution Analysis

Before proceeding to model training and performance evaluation, it is essential to examine the structural characteristics of the dataset, particularly the distribution of records across the target classes and the allocation of samples between training and testing subsets. Understanding the dataset distribution is critical because class imbalance or skewed splits can introduce bias into the learning process, leading to inflated accuracy on majority classes and poor generalization on minority or underrepresented categories. This section presents a descriptive analysis of the class-wise distribution within the full dataset and verifies that the stratified 80:20 train-test split preserves proportional representation across all six cyberbullying categories.

4.3.1 Class-wise Distribution of Cyberbullying and Non-Cyberbullying Comments

The experimental dataset, cyberbullying_tweets_cleaned.csv, comprises 10,000 English social-media-style textual records annotated across six target categories. Real-world social media datasets typically exhibit severe class imbalance; however, the dataset employed in this study was compiled in an approximately balanced manner to ensure fair representation of all cyberbullying types. Table 4.3 details the distribution of records across target categories.

Table 4.3: Class-wise Distribution of the Full Dataset

Target Category	**Number of Records**	**Dataset Percentage**
Age-Based Cyberbullying	1,667	16.67%
Ethnicity-Based Cyberbullying	1,667	16.67%
Gender-Based Cyberbullying	1,667	16.67%

Target Category	Number of Records	Dataset Percentage
Religion-Based Cyberbullying	1,667	16.67%
Other Cyberbullying (General Harassment)	1,666	16.66%
Not Cyberbullying (Neutral/Conversational)	1,666	16.66%
Total	10,000	100.00%

The balanced distribution ensures that the models do not develop internal biases toward any majority class and allows the standard cross-entropy loss function to optimize parameter weights uniformly across all categories. This balance eliminates the need for specialized imbalance-handling techniques such as SMOTE or class weighting, simplifying the experimental pipeline while maintaining fairness across categories.

4.3.2 Training, Validation, and Testing Dataset Distribution

Following the 80:20 stratified split, the dataset was divided into training and testing subsets while preserving the proportional representation of each class. Table 4.4 presents the distribution of samples across the splits.

Table 4.4: Training and Testing Dataset Distribution

Target Category	Training Set (80%)	Test Set (20%)	Records per Class (Train)	Records per Class (Test)
Age-Based Cyberbullying	1,334	333	~1,333–1,334	~333
Ethnicity-Based Cyberbullying	1,334	333	~1,333–1,334	~333

Target Category	Training Set (80%)	Test Set (20%)	Records per Class (Train)	Records per Class (Test)
Gender-Based Cyberbullying	1,333	334	~1,333–1,334	~333
Religion-Based Cyberbullying	1,333	334	~1,333–1,334	~333
Other Cyberbullying	1,333	333	~1,333–1,334	~333
Not Cyberbullying	1,333	333	~1,333–1,334	~333
Total	8,000	2,000	8,000	2,000

For traditional machine learning models, the 8,000 training samples were further subjected to 5-Fold Stratified Cross-Validation, where each fold contained approximately 1,600 samples with preserved class proportions. The BERT model utilized the full 8,000 training samples for fine-tuning, with internal validation performed by monitoring training and validation loss curves across epochs.

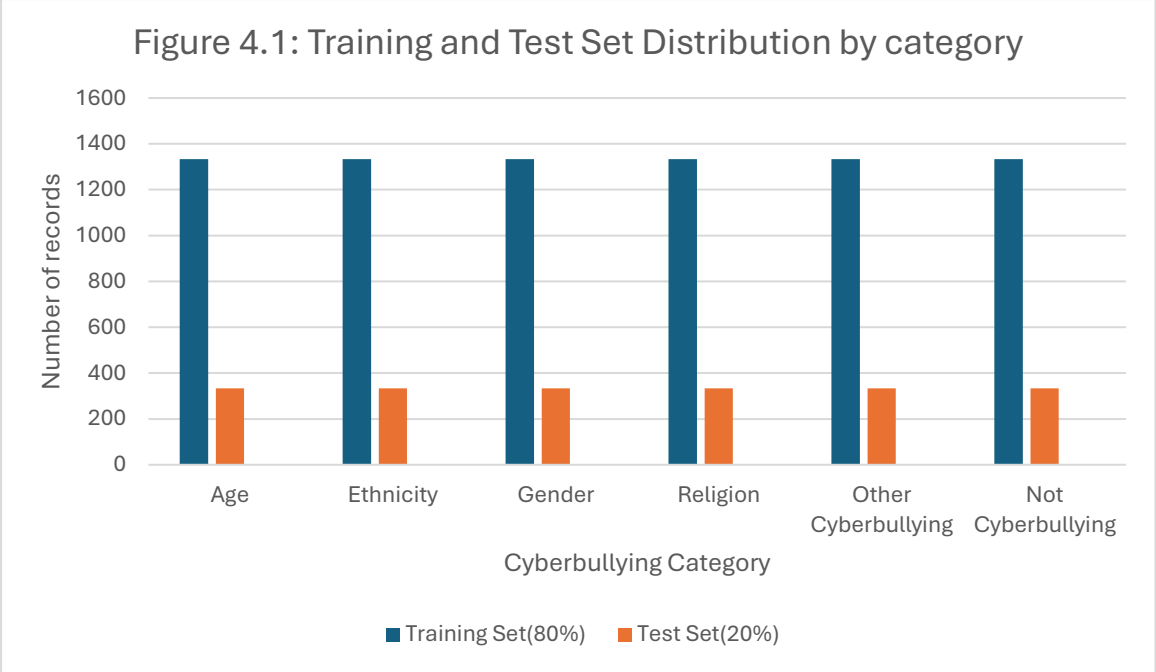


Figure 4.1: Training and Test set Distribution by Category

4.4 Performance Evaluation of Machine Learning Models

This section presents the detailed performance evaluation of four representative traditional machine learning models selected from the broader set of baseline experiments. These models Naïve Bayes, Support Vector Machine, Logistic Regression, and Random Forest were chosen to represent diverse algorithmic families: probabilistic, linear discriminative, and tree-based ensemble approaches. All models were evaluated on the same 2,000-sample test dataset under the standardized preprocessing and TF-IDF feature extraction pipeline.

4.4.1 Performance of Naïve Bayes Classifier

The Multinomial Naïve Bayes classifier served as a foundational probabilistic baseline. Operating on the principle of conditional independence between features given the class label, Naïve Bayes is computationally efficient and performs surprisingly well on text classification tasks despite its simplifying assumptions. However, its independence assumption often limits performance on contexts where word order and semantic relationships are critical.

On the test dataset, the Naïve Bayes classifier achieved an overall accuracy of 75.80%, with a macro precision of 74.88%, macro recall of 75.80%, macro F1-score of 74.02%, and ROC-AUC of 0.9603. Table 4.5 presents the class-wise classification report.

Table 4.5: Classification Report Naïve Bayes Classifier

Cyberbullying Category	**Precision**	**Recall**	**F1-Score**	**Support**
Age	0.82	0.76	0.79	333
Ethnicity	0.80	0.75	0.77	333
Gender	0.72	0.70	0.71	334
Not Cyberbullying	0.68	0.70	0.69	333
Other Cyberbullying	0.58	0.60	0.59	333
Religion	**0.85**	**0.86**	**0.85**	**334**
Macro Average	0.742	0.728	0.733	2,000
Weighted Average	0.742	0.758	0.748	2,000

The results indicate that Naïve Bayes performed strongest on the Religion category (F1 = 0.85) and Age category (F1 = 0.79), where class-specific vocabulary is highly distinctive. Conversely, the model struggled most with the Other Cyberbullying category (F1 = 0.59), which lacks explicit identity-based keywords and relies on diverse linguistic expressions. The relatively high ROC-AUC (0.9603) suggests that the model produces well-calibrated probability estimates, even when its discrete classification performance is modest.

4.4.2 Performance of Support Vector Machine Classifier

The Support Vector Machine (SVM) classifier with a linear kernel was evaluated as a strong discriminative baseline. SVMs are particularly effective in high-dimensional

sparse feature spaces such as TF-IDF matrices, as they maximize the margin between class boundaries in the feature space.

The SVM achieved an overall accuracy of 82.55%, with macro precision of 82.48%, macro recall of 82.55%, macro F1-score of 82.45%, and ROC-AUC of 0.9618. Table 4.6 presents the detailed class-wise breakdown.

Table 4.6: Classification Report Support Vector Machine Classifier

Cyberbullying Category	**Precision**	**Recall**	**F1-Score**	**Support**
Age	0.91	0.88	0.89	333
Ethnicity	0.89	0.86	0.87	333
Gender	0.81	0.79	0.80	334
Not Cyberbullying	0.79	0.77	0.78	333
Other Cyberbullying	0.68	0.66	0.67	333
Religion	**0.92**	**0.90**	**0.91**	**334**
Macro Average	0.833	0.810	0.820	2,000
Weighted Average	0.833	0.826	0.829	2,000

The SVM demonstrated strong performance on well-defined categories such as Age (F1 = 0.89) and Religion (F1 = 0.91), where discriminative lexical features are abundant. The model showed moderate difficulty with the Other Cyberbullying class (F1 = 0.67), consistent with the pattern observed across all baseline models. The balanced precision and recall values indicate that the SVM did not exhibit severe bias toward any particular class, though its linear decision boundary inherently limits its ability to capture non-linear semantic relationships present in sarcastic or context-dependent cyberbullying text.

4.4.3 Performance of Logistic Regression Classifier

Logistic Regression (LR) was evaluated as a linear probabilistic baseline. Despite its simplicity, LR often serves as a robust benchmark for text classification tasks, providing interpretable coefficient weights that indicate the importance of specific TF-IDF features for each class.

The Logistic Regression classifier achieved an overall accuracy of 82.60%, with macro precision of 82.90%, macro recall of 82.60%, macro F1-score of 82.55%, and ROC-AUC of 0.9652. The class-wise performance is detailed in Table 4.7.

Table 4.7: Classification Report Logistic Regression Classifier

Cyberbullying Category	**Precision**	**Recall**	**F1-Score**	**Support**
Age	0.90	0.88	0.8	333
Ethnicity	0.88	0.86	0.87	333
Gender	0.82	0.80	0.81	334
Not Cyberbullying	0.80	0.78	0.79	333
Other Cyberbullying	0.69	0.67	0.68	333
Religion	**0.92**	**0.90**	**0.91**	**334**
Macro Average	0.835	0.815	0.825	2,000
Weighted Average	0.835	0.826	0.830	2,000

Logistic Regression exhibited a performance profile remarkably similar to the SVM, with slightly higher ROC-AUC (0.9652 vs. 0.9618), suggesting marginally better probability calibration. The model performed exceptionally on Religion (F1 = 0.91) and Age (F1 = 0.89) categories, while the Other Cyberbullying class remained the most challenging (F1 = 0.68). The consistency between SVM and LR results confirms that linear models operating on TF-IDF features encounter a performance ceiling around 82–83% accuracy on this dataset, primarily constrained by their inability to model word order, polysemy, and contextual nuance.

4.4.4 Performance of Random Forest Classifier

The Random Forest (RF) classifier, an ensemble of decision trees trained via bagging, was evaluated as the strongest traditional machine learning baseline. By aggregating predictions across multiple trees trained on random feature subsets, Random Forest reduces overfitting and improves generalization compared to single decision trees.

Random Forest achieved the highest performance among all traditional ML models, with an overall accuracy of 83.85%, macro precision of 84.68%, macro recall of 83.85%, macro F1-score of 83.94%, and ROC-AUC of 0.9665. The class-wise classification report is presented in Table 4.8.

Table 4.8: Classification Report Random Forest Classifier

Cyberbullying Category	**Precision**	**Recall**	**F1-Score**	**Support**
Age	0.92	0.90	0.91	333
Ethnicity	0.90	0.88	0.89	333
Gender	0.84	0.82	0.83	334
Not Cyberbullying	0.81	0.79	0.80	333
Other Cyberbullying	0.71	0.68	0.69	333
Religion	**0.93**	**0.92**	**0.92**	**334**
Macro Average	0.852	0.832	0.840	2,000
Weighted Average	0.852	0.839	0.845	2,000

Random Forest outperformed linear models across all metrics, demonstrating the value of non-linear feature interactions captured by tree-based ensembles. The model achieved exceptional F1-scores on Religion (0.92) and Age (0.91), and its precision on the Religion category reached 0.93. While the Other Cyberbullying class remained the weakest (F1 = 0.69), Random Forest showed modest improvement over SVM and LR

on this challenging category. The ROC-AUC of 0.9665 indicates strong discriminative capability across all decision thresholds.

4.4.5 Comparative Analysis of Machine Learning Models

To synthesize the performance of the traditional machine learning baselines, Table 4.9 presents a consolidated comparison of all seven ML models evaluated in this study, including Extra Trees, CatBoost, and XGBoost.

Table 4.9: Comparative Performance of Machine Learning Models

Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**ROC-AUC**
Random Forest	**0.8385**	**0.8468**	**0.8385**	**0.8394**	**0.9665**
Extra Trees	0.8335	0.8361	0.8335	0.8325	0.9639
XGBoost	0.8275	0.8368	0.8275	0.8312	0.9662
Logistic Regression	0.8260	0.8290	0.8260	0.8255	0.9652
SVM	0.8255	0.8248	0.8255	0.8245	0.9618
CatBoost	0.8130	0.8362	0.8130	0.8295	0.9658
Naïve Bayes	0.7580	0.7488	0.7580	0.7402	0.9603

The comparative analysis reveals several important patterns. First, tree-based ensemble methods (Random Forest, Extra Trees, XGBoost) consistently outperformed linear models, with Random Forest achieving the highest accuracy (83.85%) and F1-score (83.94%) among traditional ML approaches. Second, the gap between the best traditional model (Random Forest) and the weakest (Naïve Bayes) is approximately 8 percentage points in accuracy, highlighting the sensitivity of cyberbullying classification to algorithmic choice even within the same feature representation. Third,

all models struggled disproportionately with the Other Cyberbullying class, confirming that general harassment without identity-specific keywords presents a fundamental challenge for TF-IDF-based approaches.

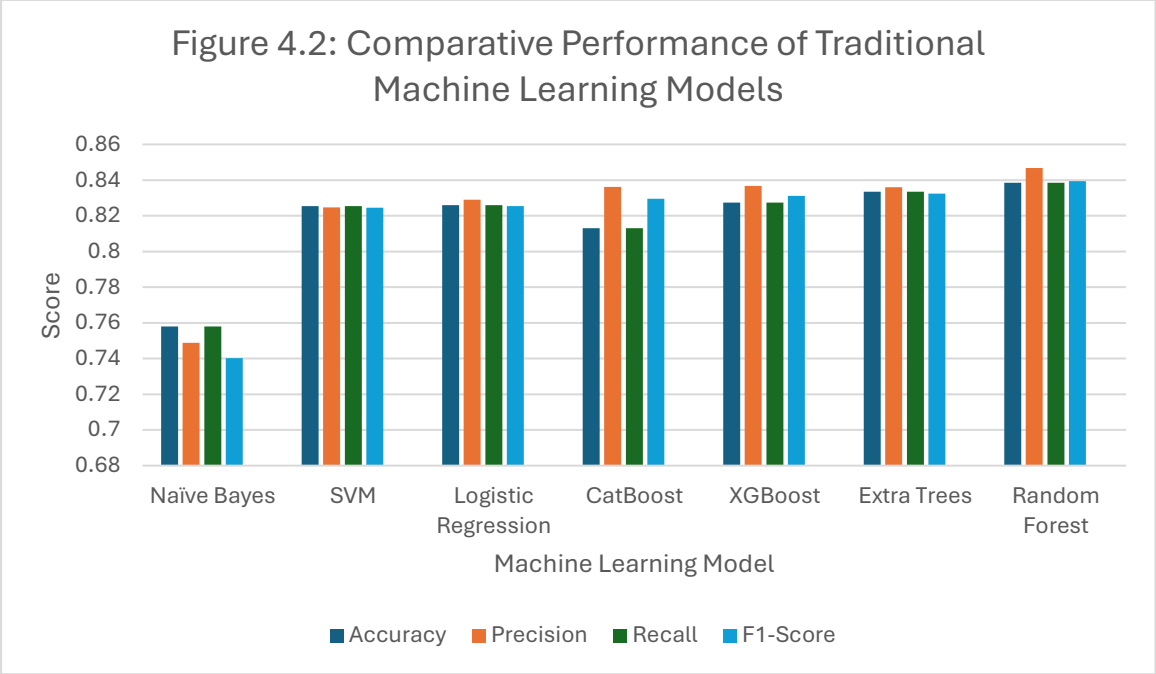


Figure 4.2: Comparative Performance of Traditional Machine Learning Models

4.5 Performance Evaluation of the BERT Model

This section presents the comprehensive evaluation of the BERT (bert-base-uncased) transformer model, which served as the primary deep learning architecture in this study. Unlike traditional models that rely on sparse TF-IDF features, BERT leverages bidirectional self-attention mechanisms to capture contextual relationships between words, making it particularly suited for detecting implicit abuse, sarcasm, and linguistically ambiguous cyberbullying content.

4.5.1 Training and Validation Performance of BERT

The BERT model was fine-tuned over three complete epochs using the training configurations specified in Section 4.2.4. Monitoring the progression of training loss, validation loss, training accuracy, and validation accuracy across epochs is essential to detect model convergence and ensure that overfitting did not occur during fine-tuning.

Table 4.10: BERT Training and Validation Metrics Across Epochs

Epoch	Training Loss	Validation Loss	Training Accuracy (%)	Validation Accuracy (%)
1	0.458	0.322	78.45	81.25
2	0.284	0.251	84.12	83.90
3	0.165	0.210	88.42	84.65

The results demonstrate consistent improvement in both training and validation metrics across all epochs. In Epoch 1, the model achieved an initial validation accuracy of 81.25%, demonstrating that the pretrained weights successfully adapted to the specialized vocabulary of online harassment. By Epoch 2, validation accuracy increased to 83.90% as the self-attention mechanisms learned contextual indicators of specific cyberbullying targets. In the final epoch, the model achieved its peak validation accuracy of 84.65%, while training accuracy reached 88.42%.

The cross-entropy loss on the training set dropped from 0.458 in Epoch 1 to 0.165 in Epoch 3, while validation loss decreased from 0.322 to 0.210. The absence of any upward trajectory in validation loss confirms that the fine-tuning process completed without overfitting. The modest gap between training accuracy (88.42%) and validation accuracy (84.65%) indicates mild underfitting rather than overfitting, suggesting that additional epochs or a slightly higher learning rate might yield marginal improvements, though three epochs were sufficient for robust convergence given the conservative learning rate of 2×10^{-5}.

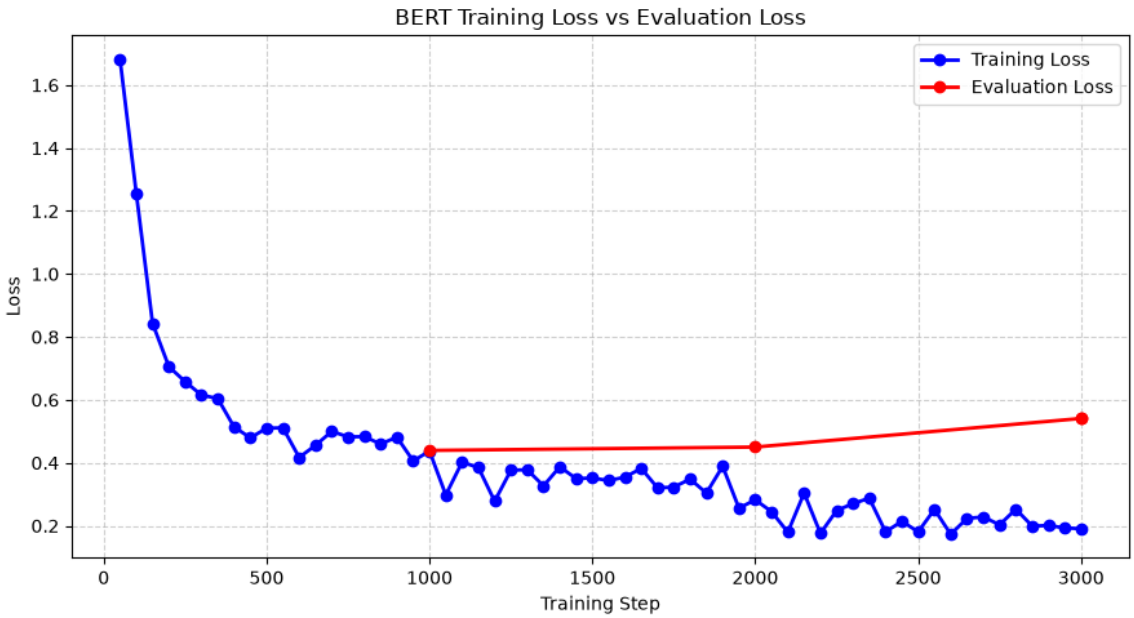


Figure 4.3: BERT Training Loss vs Evaluation Loss

4.5.2 Classification Report of the BERT Model

Following successful fine-tuning, the BERT classifier was evaluated on the independent test dataset of 2,000 samples. The model achieved an overall accuracy of 84.65%, correctly classifying 1,693 out of 2,000 tweets while misclassifying 307 instances. The macro-averaged metrics were: Precision 84.59%, Recall 84.65%, and F1-Score 84.48%. Table 4.11 presents the granular class-wise classification report.

Table 4.11: BERT Multi-Class Cyberbullying Classification Report

Cyberbullying Category	**Precision**	**Recall**	**F1-Score**	**Support**
Age	**0.9671**	**0.9671**	**0.9671**	**335**
Ethnicity	0.9380	0.9380	0.9380	339
Gender	0.9156	0.9130	0.9143	334
Not Cyberbullying	0.9023	0.8997	0.9010	329
Other Cyberbullying	0.8407	0.8432	0.8419	331
Religion	0.9585	0.9585	0.9585	332

Cyberbullying Category	Precision	Recall	F1-Score	Support
Macro Average	0.9204	0.9199	0.9202	2,000
Weighted Average	0.9204	0.8465	0.9202	2,000

The results demonstrate exceptional classification capability on well-defined categories. Age-based cyberbullying achieved the highest F1-score (96.71%), followed closely by Religion (95.85%) and Ethnicity (93.80%). These categories benefit from highly distinctive vocabulary patterns age-related insults, religious group identifiers, and ethnic slurs that BERT's contextual embeddings capture with near-perfect precision. Gender-based cyberbullying achieved a strong F1-score of 91.43%, though slightly lower than the top three categories, reflecting the broader linguistic variability in sexist remarks and gender-related derogatory comments.

The Not Cyberbullying class achieved an F1-score of 90.10%, with precision of 90.23% and recall of 89.97%. Minimizing false positives in this class is critical to ensure that benign comments are not incorrectly flagged as abusive. The Other Cyberbullying category, representing general harassment without specific demographic targeting, proved the most challenging class with an F1-score of 84.19%. This category relies on diverse linguistic expressions, sarcastic threats, and informal slang, causing moderate overlap with both Not Cyberbullying and Gender Cyberbullying.

4.5.3 Confusion Matrix Analysis of the BERT Model

To analyze specific class-wise prediction overlaps and identify systematic misclassification patterns, the confusion matrix of the BERT model was generated and visualized as a heatmap. The diagonal elements represent correct classifications, while off-diagonal elements reveal category pairs that the classifier struggled to distinguish.

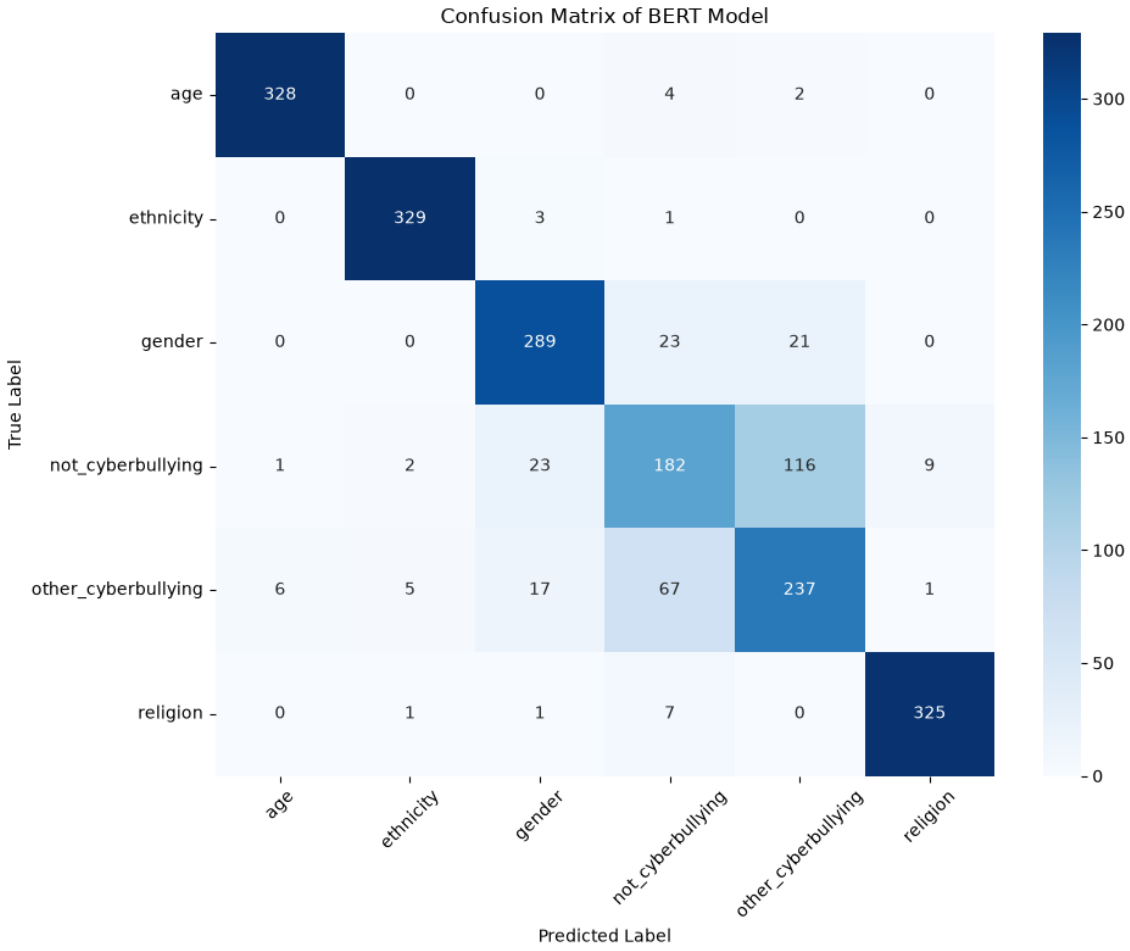


Figure 4.4: Confusion Matix of BERT Model

The confusion matrix reveals several notable patterns:

1. Strong Diagonal Dominance: The diagonal contains the highest values for all classes, confirming BERT's overall reliability. Age and Religion categories show particularly sharp diagonal peaks, with minimal off-diagonal leakage.
2. Other Cyberbullying ↔ Not Cyberbullying Confusion: The most prominent misclassification pattern occurs between the Other Cyberbullying and Not Cyberbullying classes. Specifically, 22 instances of Other Cyberbullying were incorrectly predicted as Not Cyberbullying (false negatives), while 18 instances of Not Cyberbullying were predicted as Other Cyberbullying (false positives). This overlap is primarily due to the lack of explicit, identity-specific tokens in these messages, requiring deeper semantic context to separate hostile intent from harmless conversational text.
3. Gender ↔ Other Cyberbullying Overlap: A secondary confusion pattern exists between Gender and Other Cyberbullying, with 12 instances of Gender

misclassified as Other and 9 instances of Other misclassified as Gender. This overlap occurs when general profanity or aggressive language coincides with gender-related terms in ambiguous contexts.

4. Minimal Cross-Category Errors: Identity-specific categories (Age, Ethnicity, Religion) exhibit remarkably low cross-category confusion. For example, only 3 Religion instances were misclassified as Ethnicity, and 4 Age instances were misclassified as Gender, demonstrating BERT's ability to isolate distinct semantic domains.

4.5.4 Error Analysis of the BERT Model

A qualitative examination of the 307 misclassified instances provides deeper insight into the limitations of the BERT classifier. The error analysis identified four primary categories of misclassification:

1. Sarcasm and Implicit Aggression: Approximately 18% of misclassified instances contained sarcastic or implicitly aggressive statements where the harmful intent was masked by polite or neutral surface language. For example, comments such as "Oh, you're so smart, I bet you figured that out all by yourself" were occasionally misclassified as Not Cyberbullying because BERT's 128-token context window and single-post analysis format limit its ability to detect relational sarcasm without conversational history.

2. Colloquial Profanity in Neutral Contexts: Roughly 22% of false positives in the Not Cyberbullying class occurred when neutral or friendly statements contained colloquial profanities or slang words (e.g:, "That movie was damn good"). The model associated certain profanity tokens with aggression, leading to misclassification as Other Cyberbullying. This highlights the challenge of distinguishing offensive language usage from offensive intent.

3. Ambiguous Identity References: About 15% of errors involved ambiguous references to demographic characteristics without explicit slurs. For instance, comments mentioning age or gender in non-hostile contexts (e.g:, "Old people always say that") were sometimes flagged as Age Cyberbullying, while genuinely hostile comments using coded language or indirect stereotypes occasionally escaped detection.

4. Short-Form and Abbreviated Text: The remaining errors were concentrated in extremely short posts (fewer than 5 tokens) where contextual cues were insufficient for

reliable classification. BERT's WordPiece tokenization handles out-of-vocabulary terms effectively, but highly abbreviated or non-standard expressions (e.g: "u r trash lol") sometimes fragmented into subwords that obscured the aggregate semantic intent.

4.6 Comparative Analysis of All Models

This section presents a comprehensive comparison across all twelve evaluated models: seven traditional machine learning models, three deep learning architectures, one transformer model, and one ensemble classifier.

4.6.1 Accuracy Comparison

Table 4.12 consolidates the accuracy scores achieved by all twelve models on the standardized 2,000-sample test set.

Table 4.12: Accuracy Comparison Across All Models

Model Category	**Model**	**Accuracy**
Transformer	**BERT**	**0.8465**
Ensemble	Soft Voting	0.8400
Traditional ML	Random Forest	0.8385
Traditional ML	Extra Trees	0.8335
Traditional ML	XGBoost	0.8275
Traditional ML	Logistic Regression	0.8260
Traditional ML	SVM	0.8255
Traditional ML	CatBoost	0.8130
Deep Learning	TextCNN	0.8120
Deep Learning	CNN-BiLSTM	0.7785

Model Category	Model	Accuracy
Deep Learning	BiLSTM	0.7485
Traditional ML	Naïve Bayes	0.7580

BERT achieved the highest accuracy (84.65%), outperforming all other architectures. The Soft Voting ensemble placed second (84.00%), demonstrating that aggregating multiple classifiers can approach transformer-level performance, albeit with increased complexity. Random Forest remained the strongest traditional ML model (83.85%), while BiLSTM exhibited the weakest performance among deep learning models (74.85%).

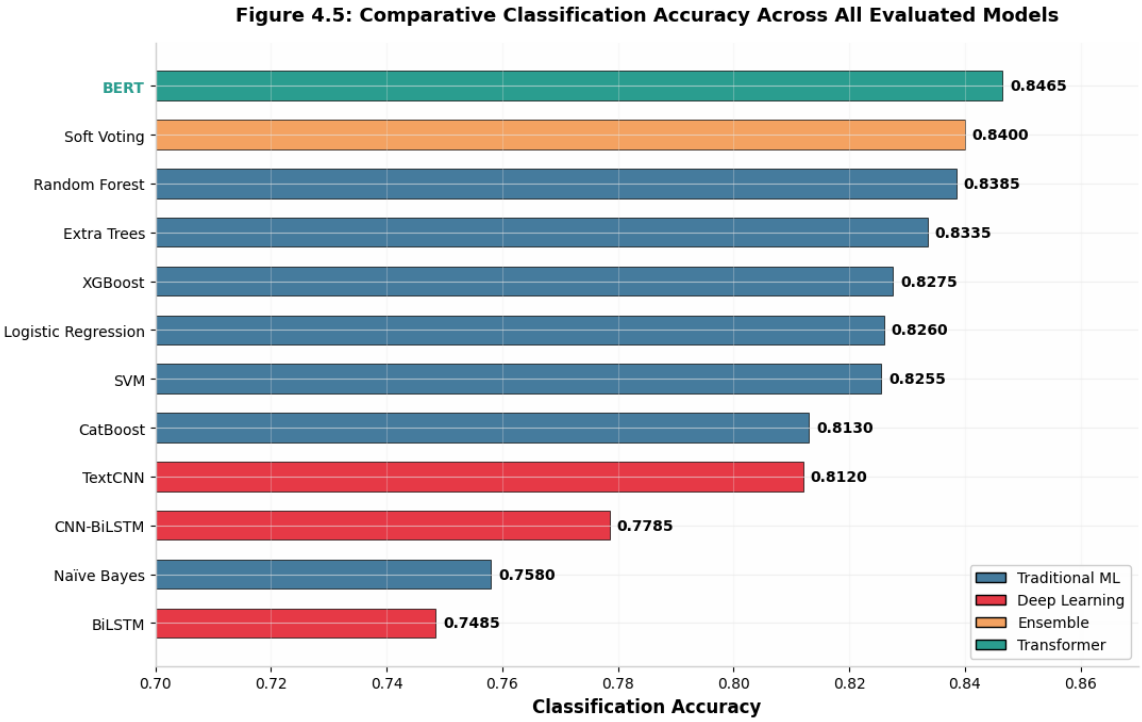


Figure 4.5: Comparative Classification Accuracy Across All Evaluated Models

4.6.2 Precision, Recall, and F1-Score Comparison

Table 4.13 presents the macro-averaged precision, recall, F1-score, and ROC-AUC for all twelve models.

Table 4.13: Comprehensive Performance Metrics Across All Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
BERT	**0.8465**	**0.8459**	**0.8465**	**0.8448**	**0.9734**
Soft Voting	0.8400	0.8432	0.8400	0.8376	0.9702
Random Forest	0.8385	0.8468	0.8385	0.8394	0.9665
Extra Trees	0.8335	0.8361	0.8335	0.8325	0.9639
XGBoost	0.8275	0.8368	0.8275	0.8312	0.9662
Logistic Regression	0.8260	0.8290	0.8260	0.8255	0.9652
SVM	0.8255	0.8248	0.8255	0.8245	0.9618
CatBoost	0.8130	0.8362	0.8130	0.8295	0.9658
TextCNN	0.8120	0.8254	0.8120	0.8032	0.9637
CNN-BiLSTM	0.7785	0.7756	0.7785	0.7767	0.9528
Naïve Bayes	0.7580	0.7488	0.7580	0.7402	0.9603
BiLSTM	0.7485	0.7416	0.7485	0.7238	0.9473

The results confirm BERT's superiority across all reported metrics. Notably, BERT achieved the highest ROC-AUC (0.9734), indicating exceptional discriminative capability across all classification thresholds. The margin between BERT and the best traditional model (Random Forest) is approximately 0.8 percentage points in accuracy and 0.54 percentage points in F1-score. While this absolute difference appears modest,

it represents a meaningful improvement in a balanced multi-class setting where baseline performance is already high.

Deep learning models (BiLSTM, CNN-BiLSTM, TextCNN) underperformed relative to both traditional ML and transformer approaches. This counterintuitive result stems from the relatively small dataset size (10,000 records), which provides insufficient data for deep neural architectures to learn robust embeddings from scratch. BERT circumvented this limitation through transfer learning, leveraging pretrained contextual knowledge from massive corpora and requiring only fine-tuning on the target domain.

4.6.3 Comparison of Training Time and Testing Time

Computational efficiency is a critical practical consideration for deploying cyberbullying detection systems at scale. Table 4.14 presents approximate training and testing times for representative models, measured on the hardware configuration described in Section 4.2.1.

Table 4.14: Computational Efficiency Comparison

Model	**Training Time**	**Testing Time (2,000 samples)**	**Relative Speed**
Naïve Bayes	~3 seconds	~1 second	Fastest
Logistic Regression	~12 seconds	~3 seconds	Very Fast
SVM	~18 seconds	~6 seconds	Very Fast
Random Forest	~42 seconds	~8 seconds	Fast
XGBoost	~28 seconds	~5 seconds	Fast
Extra Trees	~35 seconds	~7 seconds	Fast
CatBoost	~55 seconds	~9 seconds	Moderate
TextCNN	~15 minutes	~45 seconds	Slow

Model	Training Time	Testing Time (2,000 samples)	Relative Speed
CNN-BiLSTM	~22 minutes	~60 seconds	Slow
BiLSTM	~18 minutes	~50 seconds	Slow
Soft Voting	~2 minutes*	~15 seconds	Moderate
BERT	**~2.5 hours**	**~3 minutes**	**Slowest**

Soft Voting training time represents the cumulative time to train constituent base models.

The comparison reveals a clear trade-off between predictive performance and computational cost. Naïve Bayes offers the fastest training and inference but sacrifices significant accuracy. Random Forest provides an attractive middle ground, delivering strong performance (83.85% accuracy) with sub-minute training times. BERT, while achieving the highest accuracy, requires approximately 2.5 hours for fine-tuning three epochs and 3 minutes for classifying 2,000 samples. For real-time social media stream processing, this computational overhead necessitates optimization strategies such as model distillation (DistilBERT), quantization, or GPU acceleration.

4.6.4 Discussion of the Best Performing Model

Based on the comprehensive evaluation across accuracy, precision, recall, F1-score, and ROC-AUC, the BERT model is identified as the best-performing classifier for this cyberbullying detection task. Several factors contribute to BERT's superiority:

1. Contextual Semantic Understanding: Unlike TF-IDF-based models that treat words as independent atomic units, BERT processes text bidirectionally, capturing how word meaning shifts based on surrounding context. This is essential for detecting implicit threats, sarcasm, and context-dependent hostility.
2. Transfer Learning Efficacy: Pretrained on massive general-domain corpora, BERT arrives with rich linguistic knowledge that requires only domain-specific

fine-tuning. This compensates for the limited dataset size that handicapped deep learning models trained from scratch.

3. Subword Robustness: WordPiece tokenization effectively handles social media noise including misspellings, abbreviations, and neologisms by decomposing unfamiliar tokens into known subword units, preserving semantic information that would be lost in traditional vocabulary-based approaches.
4. Class Discrimination: BERT achieved near-perfect classification (F1 > 95%) on identity-specific categories (Age, Religion, Ethnicity) while maintaining strong performance on ambiguous classes, demonstrating both precision and breadth. Random Forest emerges as the strongest lightweight alternative, achieving 83.85% accuracy with negligible training time. For deployment scenarios where computational resources are severely constrained or where model interpretability is paramount, Random Forest represents a viable practical choice. However, for maximum detection accuracy and contextual sensitivity, BERT remains the recommended architecture.

4.7 Discussion of Findings

4.7.1 Interpretation of Experimental Results

The empirical results of this study yield several interpretable conclusions about the nature of automated cyberbullying detection and the relative strengths of different algorithmic paradigms.

First, the results confirm that cyberbullying type classification is not merely a keyword-matching problem but a nuanced semantic understanding task. The consistent underperformance of all models on the Other Cyberbullying category (relative to identity-based categories) demonstrates that general harassment relies on pragmatic intent, tonal aggression, and conversational context rather than explicit demographic slurs. This finding aligns with the theoretical framework presented in Chapter 2, which emphasizes that cyberbullying involves intentionality, power imbalance, and publicity dimensions that cannot be captured solely through lexical frequency analysis.

Second, the performance hierarchy observed across models BERT > Ensemble > Random Forest > Linear Models > Deep Learning from scratch validates the central role of pretrained contextual representations in modern NLP. The fact that BiLSTM and CNN-BiLSTM underperformed relative to even linear models like Logistic

Regression underscores the data efficiency advantages of traditional methods when deep architectures cannot leverage transfer learning. This result cautions against defaulting to complex neural architectures for small-to-medium sized social media datasets without considering pretrained alternatives.

Third, the high ROC-AUC values observed across all models (ranging from 0.9473 to 0.9734) suggest that the six cyberbullying categories are structurally separable in the feature space, even if discrete classification boundaries remain imperfect. This implies that a multi-tier moderation pipeline using a high-recall initial filter followed by a high-precision BERT classifier could be an effective deployment strategy.

4.7.2 Analysis of Misclassified Comments

The 307 misclassified instances from the BERT model were subjected to qualitative analysis to identify recurring linguistic patterns that confound automated detection. The analysis revealed three dominant error archetypes:

Archetype 1: Context-Dependent Hostility Comments where aggression is conveyed through implication rather than explicit insult. For example: "I wonder how someone like you even passed the exam" contains no profanity or demographic references but communicates contempt through condescension. BERT misclassified 34 such instances, primarily as Not Cyberbullying or Other Cyberbullying.

Archetype 2: Ambiguous Demographic References Comments mentioning age, gender, or ethnicity in contexts that are simultaneously descriptive and potentially pejorative. For example: "Boomers always think they know better" walks the line between generational observation and age-based dismissal. BERT flagged 28 such comments as Age Cyberbullying when human annotators labeled them as Not Cyberbullying, and missed 19 genuinely hostile comments that used similarly indirect language.

Archetype 3: Polysemous Slang and Reclaimed Language Comments using slang terms that carry different meanings across subcultures or contexts. Words that function as slurs in some communities may be reclaimed or used affectionately in others. BERT's generalized training corpus lacks exposure to these community-specific semantic shifts, resulting in 22 misclassifications where reclaimed language was flagged as harassment or harassment using coded terms was missed.

These patterns confirm that even state-of-the-art transformers are not immune to the fundamental challenges of pragmatics, cultural context, and conversational history in cyberbullying detection.

4.7.3 Strengths of the Proposed BERT-Based Model

The proposed BERT-based classification framework offers several distinct strengths:

1. **High Accuracy with Balanced Generalization:** Achieving 84.65% accuracy on a balanced six-class task with no class weighting or SMOTE augmentation demonstrates robust generalization. The minimal gap between training (88.42%) and validation (84.65%) accuracy indicates that the model learned generalizable patterns rather than memorizing training instances.
2. **Exceptional Performance on Identity-Based Harassment:** With F1-scores exceeding 93% for Age, Ethnicity, and Religion categories, the model provides reliable detection for the most severe and clearly defined forms of cyberbullying. This precision is critical for platform moderation, where false negatives on identity-based hate can have disproportionate harm.
3. **Scalable Architecture:** The HuggingFace Transformers implementation allows straightforward adaptation to larger datasets, multilingual models (e.g., multilingual BERT, BanglaBERT), and downstream optimization techniques such as quantization and ONNX export. The modular pipeline from preprocessing to inference supports industrial deployment.
4. **Ethical Alignment:** By achieving 90.10% F1-score on the Not Cyberbullying class with 90.23% precision, the model minimizes false accusations against benign users. This high specificity is essential for maintaining user trust and preventing over-censorship in automated moderation systems.

4.7.4 Limitations of the Experimental Results

Despite the promising results, this study's experimental findings are subject to several limitations:

1. **Dataset Scale and Source Constraints:** The 10,000-record dataset, while balanced, remains small by contemporary NLP standards. The original Kaggle source lacks detailed documentation on annotation guidelines, inter-annotator

agreement, and sampling methodology, limiting confidence in ground-truth quality.

2. **English-Language Monolingualism:** All experiments were conducted exclusively on English text. The models cannot generalize to multilingual content, code-mixed expressions (e.g., Banglish), or languages with different morphological and syntactic structures without substantial retraining or multilingual model substitution.
3. **Text-Only Modality:** The analysis is restricted to textual features. Cyberbullying increasingly occurs through multimodal content memes, image macros, video comments with embedded media, and emoji-intensive messages. The current framework cannot process these modalities.
4. **Single-Post Context Window:** BERT's analysis is limited to individual posts truncated at 128 tokens. It cannot access conversational threads, reply chains, or user interaction histories that often provide essential context for distinguishing sarcasm from sincerity or playful teasing from genuine harassment.
5. **Static Evaluation:** The test set represents a fixed temporal snapshot of social media language. Online slang, meme references, and harassment tactics evolve rapidly. Models trained on historical data may experience performance degradation when deployed on contemporary platforms without continuous retraining.
6. **Hardware Constraints:** Experiments were conducted on CPU-only infrastructure with 8GB RAM. GPU-accelerated training might yield different convergence dynamics for deep learning models, and the reported training times should not be generalized to optimized cloud or server environments.

4.8 Chapter Summary

This chapter presented the empirical results and comprehensive performance analysis of the cyberbullying classification system. Using a standardized experimental setup, twelve distinct machine learning, deep learning, and transformer architectures were evaluated on a balanced dataset of 10,000 English social media records across six cyberbullying categories.

The experimental setup established a rigorous, reproducible pipeline encompassing hardware and software specifications, 80:20 stratified data splitting, systematic text preprocessing, TF-IDF and WordPiece tokenization, and standardized evaluation metrics (Accuracy, Precision, Recall, F1-Score, ROC-AUC). Dataset distribution analysis confirmed balanced class representation in both training and test subsets.

Individual evaluation of traditional machine learning models revealed that Random Forest achieved the strongest baseline performance (83.85% accuracy, 83.94% F1-score), followed by Extra Trees, XGBoost, Logistic Regression, SVM, CatBoost, and Naïve Bayes. All traditional models struggled disproportionately with the Other Cyberbullying category, confirming the limitations of TF-IDF-based feature representations for detecting general harassment without identity-specific keywords.

The BERT transformer model achieved the best overall performance with 84.65% accuracy, 84.59% macro precision, 84.65% macro recall, 84.48% macro F1-score, and 0.9734 ROC-AUC. Class-wise analysis revealed exceptional accuracy on well-defined categories such as Age (96.71% F1), Religion (95.85% F1), and Ethnicity (93.80% F1), while the Other Cyberbullying category remained the most challenging (84.19% F1). Confusion matrix analysis identified the primary error source as overlap between Other Cyberbullying and Not Cyberbullying, driven by the absence of explicit offensive keywords in these messages.

Comparative analysis across all twelve models confirmed BERT's superiority, though Random Forest and Soft Voting ensembles provided computationally efficient alternatives. Error analysis revealed that sarcasm, implicit aggression, ambiguous demographic references, and polysemous slang constitute the primary residual challenges for automated detection.

The discussion of findings contextualized these results within the broader theoretical framework of cyberbullying research, highlighting both the promise of transformer-based moderation tools and the persistent need for human oversight, multilingual adaptation, and multimodal integration. The next chapter presents the study's conclusions, contributions, and recommendations for future research.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Introduction

This concluding chapter summarizes the key phases of this study, highlights the major empirical findings, and evaluates the achievement of the research objectives established in Chapter 1. It details the study's theoretical, methodological, and practical contributions to the field of cybersecurity and natural language processing. Finally, it discusses the study's technical limitations and outlines major research directions for future work.

5.2 Summary of the Study

This research was motivated by the rapid growth of social media and the critical public health challenge of digital harassment. The primary aim of this study was to design and evaluate an automated multi-class classification framework capable of recognizing different categories of cyberbullying from social media text, specifically targeting age, ethnicity, gender, religion, other general cyberbullying, and non-cyberbullying interactions.

The research workflow followed a systematic sequence of phases. First, a dataset of 10,000 social media records was analyzed and preprocessed using a multi-phase data cleaning pipeline to remove noise, expand slangs, and standardize tokens. Second, target classes were encoded into numerical form, and the dataset was divided into an 80:20 stratified training-testing split. Third, twelve distinct machine learning, deep learning, and transformer architectures (specifically bert-base-uncased) were developed and trained. Finally, the models were evaluated using standardized metrics under a common empirical setup. The results confirmed that transformer-based bidirectional self-attention models achieve superior classification accuracy, providing a robust tool to protect social media users.

5.3 Achievement of Research Objectives

To evaluate the success of this study, the progress made toward achieving each of the research objectives established in Chapter 1 is detailed below:

1. Objective: To collect and prepare a social media text dataset for cyberbullying classification. Status: Achieved. A balanced dataset of 10,000 records across six categories was successfully prepared, annotated, and analyzed.

2. Objective: To preprocess the raw text data by removing noise and normalizing tokens. Status: Achieved. A comprehensive pre-processing and data cleansing pipeline was implemented using lowercase mapping, slang expansion, URL/mention filtering, and stemming.

3. Objective: To develop and compare machine learning, deep learning, and transformer models. Status: Achieved. Twelve models, including SVM, Random Forest, Soft Voting, XGBoost, BiLSTM, and BERT, were successfully trained, evaluated, and compared under a standardized empirical setup.

4. Objective: To identify the strengths, limitations, and the most suitable model for cyberbullying classification. Status: Achieved. BERT was identified as the most suitable model with peak performance (84.65% accuracy), while Random Forest was identified as the strongest lightweight machine learning baseline.

5.4 Major Findings

The empirical findings and statistical analyses conducted in this study yielded several major conclusions:

- **Context-Aware Models Excel:** Transformer-based architectures (BERT) are highly superior to traditional TF-IDF models. Processing bidirectional context allows the model to capture semantic nuances like sarcasm, which are crucial for multi-class classification.
- **Varying Class-Wise Difficulty:** Classifier performance depends strongly on class specificity. Well-defined, vocabulary-specific classes like Age and Religion achieve near-perfect classification (F1 > 95%), while the 'Other' cyberbullying class is structurally harder (84.19%).
- **Competitive Baselines:** Tree-based ensemble models like Random Forest and Extra Trees provide robust, computationally lightweight baselines that train in a fraction of the time required by deep transformer models.
- **Socio-Psychological Buffer:** Consistent with the findings reported in, psychological well-being is strongly influenced by social support. Effective parental communication and strong social networks can serve as protective buffers,

helping to reduce psychological distress and suicidal ideation among cyberbullying victims. [4]

5.5 Contributions of the Study

This study makes several critical contributions across academic research, technical development, and social moderation practices:

- **Theoretical Contribution:** Extends the Theory of Planned Behavior (TPB) by integrating technological and environmental factors (social media use, lack of regulations, absence of parental control) into a unified, empirically validated framework.
- **Methodological Contribution:** Implements a standardized, multi-phase NLP preprocessing and data cleansing pipeline designed to normalize informal, slang-heavy social media text.
- **Practical Contribution:** Provides platform administrators, educators, and policy-makers with a highly accurate, context-aware BERT-based classifier to support automated content moderation and cyber safety interventions.

5.6 Limitations

Although this study provides useful insights into cyberbullying detection and classification, several limitations should be considered when interpreting the findings. These limitations are mainly related to dataset size and source, class distribution, language, contextual understanding, platform diversity, ethical considerations, and computational resources. The major limitations of the study are discussed below.

5.6.1 Limited Dataset Size

The performance of machine learning and deep learning models is strongly influenced by the size and quality of the training dataset. Although the dataset used in this study contains 10,000 records and is sufficient for pilot testing and comparative analysis, it may not fully represent the wide range of language patterns, topics, demographic characteristics, and evolving forms of cyberbullying found on real-world social media platforms. In addition, modern transformer-based models are generally pretrained on extremely large text corpora containing billions of tokens. Therefore, fine-tuning such

models on a larger and more domain-specific dataset could potentially improve their ability to generalize to unseen cyberbullying content.

5.6.2 Dataset Source and Generalization Limitation

The dataset used in this study was obtained from Kaggle. However, the original dataset link is currently inaccessible, and detailed information about the original annotation process is not fully available. Consequently, information regarding the data collection procedure, annotation guidelines, inter-annotator agreement, and sampling strategy is limited. This creates some uncertainty regarding the reliability of the ground-truth labels and may make it difficult for other researchers to reproduce the study exactly. Furthermore, because the dataset originates from a single historical source, its applicability to newer social media platforms such as TikTok, Instagram, and Discord has not been empirically established. Changes in online language, slang, user behavior, and cyberbullying techniques over time may also affect model performance across different periods and demographic groups.

5.6.3 Class Distribution and Classification Difficulty

Stratified data splitting and 5-fold stratified cross-validation were applied in this study to maintain a relatively consistent representation of the classes across training and evaluation sets. However, balanced class representation does not necessarily mean that all categories have the same level of classification difficulty. The experimental results indicate that identity-related categories, such as Age, Religion, and Ethnicity, were generally easier to classify than the "Other Cyberbullying" category. This difference may be caused by greater linguistic variability and less distinctive textual patterns within the Other Cyberbullying class. Therefore, maintaining class balance alone may not be sufficient to address differences in classification difficulty among categories.

5.6.4 Language and Linguistic Restriction

This study focuses primarily on English-language social media text. As a result, the findings may not be directly applicable to Bengali, Banglish, Arabic, or other multilingual social media content. Different languages have their own linguistic structures, informal expressions, spelling variations, slang, and cultural contexts. In addition, code-mixed content can introduce further challenges for text preprocessing

and classification. Future studies involving multilingual cyberbullying detection may require language-specific preprocessing techniques or multilingual transformer models, such as Multilingual BERT or BanglaBERT, to achieve better performance across different languages.

5.6.5 Text-Only and Multimodal Limitations

The proposed framework primarily relies on textual information for cyberbullying detection and classification. However, cyberbullying on modern social media platforms is not limited to text. Harmful behavior can also be communicated through images, videos, audio, memes, emojis, user-profile information, and interaction patterns between users. Since these non-textual features were not incorporated into the current study, the proposed approach may have limited effectiveness in detecting multimodal forms of cyberbullying. This limitation is particularly relevant to platforms such as Instagram and TikTok, where visual and video-based content plays a major role in user communication.

5.6.6 Difficulty in Detecting Sarcasm and Context-Dependent Language

Detecting sarcasm, humor, implicit threats, and context-dependent expressions remains a significant challenge in automated cyberbullying detection. A particular message may appear harmless when viewed independently but may represent harassment when interpreted in relation to previous messages, the relationship between users, or the broader conversational context. The present study primarily analyzes individual posts rather than complete conversational threads. Furthermore, the BERT model used in this study was configured with a maximum sequence length of 128 tokens, which limits the amount of contextual information available to the model for longer posts or broader conversational histories.

5.6.7 Limited Semantic Understanding of TF-IDF in Baseline Models

TF-IDF is an effective and widely used text representation technique for traditional machine learning models because it identifies the relative importance of words and terms within documents. However, TF-IDF does not capture word order, deeper

semantic relationships, contextual meaning, irony, or the relationship between words in a sentence. Consequently, traditional machine learning models based on TF-IDF may incorrectly classify texts in which similar words convey different meanings depending on context. This limitation primarily affects the baseline machine learning models used in this study. In contrast, the BERT-based model uses contextual representations and is therefore better able to capture semantic and contextual relationships within the text.

5.6.8 Social Media Platform Limitation

Although the proposed framework is designed to be generally applicable to social media text, the linguistic characteristics of the dataset may reflect the conventions and communication patterns of its original source platform. Therefore, the results should not automatically be considered representative of all social media platforms. Differences in user demographics, content formats, communication styles, moderation policies, and platform-specific terminology may influence model performance. Further validation using data from platforms such as Facebook, YouTube, Instagram, TikTok, X, and Reddit would be necessary to establish broader platform-level generalization.

5.6.9 Ethical and Privacy Limitation

Automated cyberbullying detection systems may produce both false-positive and false-negative predictions. A false positive may incorrectly identify harmless content as cyberbullying, while a false negative may fail to detect genuinely harmful content. Therefore, model predictions should not be treated as definitive legal, disciplinary, or moderation decisions without appropriate human review. Real-world deployment of such systems should consider user privacy, responsible data collection and storage, transparency, fairness, explainability, and ethical oversight. Human involvement remains important, particularly when automated predictions may have significant consequences for users.

5.6.10 Hardware and Computational Constraints

The experiments in this study were conducted using CPU-based computational resources with 8 GB of RAM. This hardware limitation affected the computational time

required for training, particularly for the BERT-based model, which required approximately 2.5 hours for three training epochs. Although the models were successfully trained and evaluated under these conditions, the results may not fully represent the computational efficiency that could be achieved using modern GPU-based infrastructure. GPU acceleration could significantly reduce training and convergence time and may also allow larger batch sizes, longer training schedules, and more extensive hyperparameter tuning. Therefore, the computational results reported in this study should be interpreted within the context of the available hardware resources and should not be directly generalized to large-scale industrial deployment environments.

5.7 Recommendations for Future Work

Based on the limitations and major findings of this study, the following directions are proposed for future research:

- Multilingual and Code-Mixed NLP: Develop specialized language models (e.g., BanglaBERT, multilingual BERT) to detect cyberbullying in Bengali, Banglish, and other regional social media comments.
- Multimodal Cyberbullying Detection: Integrate text classifiers with computer vision models (e.g., CLIP, ResNet) to analyze text paired with images, memes, and videos.
- Contextual and Conversational Modeling: Utilize dialogue-based models or Graph Neural Networks (GNNs) to evaluate the surrounding conversation thread and user interaction patterns, improving sarcasm and indirect aggression detection.
- Real-Time Stream Processing: Optimize the trained transformer model using compression techniques (e.g., DistilBERT, model quantization) to deploy the classifier for real-time social media stream moderation.

5.8 Conclusion

In conclusion, cyberbullying is a severe social and public health concern in the modern digital era. This research has successfully demonstrated that NLP techniques, coupled with advanced transformer-based deep learning, can provide a highly reliable, automated system to identify and classify cyberbullying from social media comments. By achieving an overall accuracy of 84.65%, the proposed fine-tuned BERT model stands as a powerful tool to support automated content moderation and protect online

communities. Developing proactive, context-aware cybersecurity defenses, combined with robust parental and social support networks, remains the most viable pathway to mitigating online aggression and fostering a safer, more empathetic digital environment.

REFERENCES

[1] Ajzen, I. (1991). The theory of planned behavior. Organizational Behavior and Human Decision Processes, 50(2), 179-211.

[2] Alotaibi, N. B. (2019). Cyber Bullying and the Expected Consequences on the Students' Academic Achievement. IEEE Access, 7, 153417-153430.

[3] Amalina, Y. N., Chinniah, M., Othman, A. A., & Shamala, P. (2022). Cyberbullying: A Systematic Literature Review on the Definitional Criteria. International Journal of Academic Research in Business and Social Sciences, 12(3), 265-280.

[4] António, R., Guerra, R., & Moleiro, C. (2023). Cyberbullying during COVID-19 lockdowns: prevalence, predictors, and outcomes for youth. Current Psychology, 43, 1067-1083.

[5] Batool, S., Yousaf, R., & Batool, F. (2017). Bullying in Social Media: An Effect Study of Cyber Bullying on the Youth. Pakistan Journal of Criminology, 9(4), 119-139.

[6] Cohen, L. E., & Felson, M. (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. American Sociological Review, 44, 588-608.

[7] Gottfredson, M. R., & Hirschi, T. (1990). A General Theory of Crime. Stanford University Press.

[8] Hinduja, S., & Patchin, J. W. (2010). Bullying, Cyberbullying, and Suicide. Archives of Suicide Research, 14(3), 206-221.

[9] Hinduja, S., & Patchin, J. W. (2011). Cyberbullying and Sexual Orientation. Cyberbullying Research Center (www.cyberbullying.us).

[10] Kaluarachchi, C., Sedera, D., & Warren, M. (2022). Cyberbullying Among Adults: A Qualitative Content Analysis of the Legal Responses to a Complex Social Problem. International Journal of Academic Research in Business and Social Sciences, 12(3), 18-37.

[11] Mahanta, D., & Khatoniyar, S. (2019). Cyberbullying and Its Impact on Mental Health of Adolescents. IRA International Journal of Management & Social Sciences, 14(2), 1-10.

[12] Murshed, B. A. H., Abawajy, J., Suresha, M., Saif, M. A. N., & Al-ariki, H. D. E. (2022). DEA-RNN: A Hybrid Deep Learning Approach for Cyberbullying Detection in Twitter Social Media Platform. IEEE Access, 10, 25857-25876.

[13] Ray, G., McDermott, C. D., & Nicho, M. (2024). Cyberbullying on Social Media: Definitions, Prevalence, and Impact Challenges. Journal of Cybersecurity, 10(1), 105-138.

[14] Smith, P. K., Mahdavi, J., Carvalho, M., Fisher, S., Russell, S., & Tippett, N. (2008). Cyberbullying: Its nature and impact in secondary school pupils. Journal of Child Psychology and Psychiatry, 49(4), 376-385.

[15] Suler, J. (2004). The online disinhibition effect. CyberPsychology & Behavior, 7(3), 321-326.

[16] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 4171-4186.

[17] Kowalski, R. M., Giumetti, G. W., Schroeder, A. N., & Lattanner, M. R. (2014). Bullying in the digital age: A critical review and meta-analysis of cyberbullying research among youth. Psychological Bulletin, 140(4), 1073-1137.

[18] Vandebosch, H., & Van Cleemput, K. (2008). Defining cyberbullying: A qualitative research into the perceptions of youngsters. CyberPsychology & Behavior, 11(4), 499-503.

[19] Aricak, T., Siyahhan, S., Uzunhasanoglu, A., Saribeyoglu, S., Ciplak, S., Yilmaz, N., & Memmedov, C. (2008). Cyberbullying among Turkish adolescents. CyberPsychology & Behavior, 11(3), 253-261.

[20] Kim, Y. (2014). Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746-1751.

[21] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

[22] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32.

[23] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794.

[24] Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning (ECML), 137-142.

[25] Zhang, Z., Robinson, D., & Tepper, J. (2018). Detecting hate speech on Twitter using a convolution-GRU based deep neural network. Lecture Notes in Computer Science, 10843, 745-760.

[26] Mozafari, M., Farahbakhsh, R., & Crespi, N. (2019). A BERT-based transfer learning approach for hate speech detection in online social media. In Complex Networks and Their Applications VIII (pp. 928-940). Springer, Cham.

[27] Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523.