MINISTRY OF EDUCATION
IMAM ABDULRAHMAN BIN
FAISAL UNIVERSITY
COLLEGE OF COMPUTER SCIENCE
& INFORMATION TECHNOLOGY

وزارة التعليم
جامعة الإمام
عبدالرحمن بن فيصل
كلية علوم الحاسب
وتقنية المعلومات

جامعة الإمام عبد الرحمن بن فيصل
IMAM ABDULRAHMAN BIN FAISAL UNIVERSITY

CYBERBULLYING DETECTION

ARTI 406 - Machine Learning: Project Final Report

INSTRUCTOR:
DR. IRFAN ULLAH

2025-2026 | 1st Term

Date of submision: 18/12/2025

Group members:

No.	Name	[redacted]
1.	Zahra Bahyan	[redacted]
2.	Taif Alzahrani	[redacted]
3.	Sara Almary	[redacted]

AI minor | PA01

A Comparative Study of BERT and RoBERTa for Cyberbullying Detection

Sara Almary[1], Taif Alzahrani[2], and Zahra Bahyan[3]

College of Computer Science and Information Technology, University of Dammam, Kingdom of Saudi Arabia

Abstract

The extensive development of social media has facilitated the spread of cyberbullying that poses severe psychological and social dangers, in particular, to adolescents and young adults. Online communication is frequently too massive and too fast to be subject to a human moderating hands-on approach, prompting the application of automated cyberbullying detection algorithms grounded on machine learning and natural language processing. Recent research has demonstrated that transformer-based language models, including Bidirectional Encoder Representations from Transformers (BERT), offer good baselines when it comes to this task. The present paper provides the comparative analysis of BERT and Robustly Optimized BERT Pretraining Approach (RoBERTa) to detect multi-class cyberbullying based on the text on social media. The two models were run in the same experimental conditions and tested on a labeled dataset of cyberbullying when subjected to a common preprocessing pipeline. Manual hyperparameter optimization and automated optimization were used to improve the model performance. The measures of evaluation were accuracy, precision, recall, F1-score and confusion matrices. The experimental findings demonstrate that optimized BERT model has a test accuracy of 99.46, and RoBERTa has 99.34, and both models have a high performance with all classes. The findings indicate that transformer-based models have a high effectiveness in detecting cyberbullying and can significantly surpass the current benchmarks methods.

Table of Contents

Table of Tables

Table of Figures

Keywords: Cyberbullying Detection, Natural Language Processing, BERT, RoBERTa, Social Media Analysis.

1. Introduction

The fast rise of the social media has greatly changed the communication patterns of people but it has also enhanced negative activities on the internet like cyber bullying. Cyberbullying can be described as harassment that is repeated and intentional and is conducted via digital equipment and can have serious psychological, emotional and social effects especially in adolescents and young adults. The magnitude, velocity, and anonymity of online communication make manual moderation unrealistic, which is why the creation of automated cyberbullying detection systems, which are driven by machine learning and natural language processing (NLP) methods, is encouraged [1], [2], [3].

Initial attempts at detecting cyberbullying were based mostly on rule-based systems and conventional machine learning models based on handcrafted textual features such as keyword matching and sentiment lexicons. Although such approaches offered first-time resolutions, they could not grasp the contextual meaning, sarcasm, and informal language that is commonly used in social media content [4]. Therefore, they were usually restricted in their performance particularly where implicit or non explicit harassment was involved. Recent progress in deep learning has redirected the research to representation learning, where transformer-based language models have made significant progress in text classification tasks, such as cyberbullying detection [1], [5].

BERT has become a powerful baseline among these models because of the bidirectional contextual representations that allow detecting finer abusive language more effectively [1], [2]. A number of studies have shown that BERT-based classifiers are more effective than the traditional machine learning and recurrent neural network models on various datasets and languages [6], [2]. Still, it has been demonstrated that the performance of BERT is not always the best as it can be dependent on the pattern of the dataset and have certain limitations in the form of interpretability and generalization [4]. RoBERTa is a vigorously optimized version of BERT, which has been demonstrated to further enhance performance by increasing pretraining strategies and quality of representation [1], [3].

Although these improvements have been made, much of the available literature, such as the famous one of [1], is based primarily on conventional fine-tuning methods and limited experimental settings. Besides, the comparison of BERT and RoBerta in the same experimental conditions has not been done yet, and the questions about the best model selection and performance improvement practices are still open. The given gap is what drives the current work, which intends to deploy and compare BERT and RoBerta models in detecting cyberbullying, with the clear purpose of outperforming the BERT-based baseline that is defined in [1].

We use a sophisticated BERT model in the present research and further expand it with RoBERTa under the same experimental conditions. We assess their performance on cyberbullying detection tasks by controlling and optimizing them and show significant performance gains compared to current BERT-based methods. These findings suggest that RoBERTa is always more likely to perform detection, which proves that it is suitable in this task and gives empirical evidence on how transformer-based models behave.

The rest of this work is structured in the following way. The review of related literatures is provided in section 2. Section 3 discusses the proposed machine learning methods, which are the BERT and RoBERTa classifiers. Section 4 outlines the Experimental Setup, including the optimization or parameter search strategy that was adopted. The results and discussion are given in Section 5 and the paper is concluded and recommendations made in Section 8.

2. Review of Related Literatures

Detection Cyberbullying detection is a significant research field because of the popularity of social media platforms and the devastating psychological, social and emotional damages that online bullying has. The original computational techniques were based on superficial textual characteristics and rule-of-thumb or traditional machine learning techniques. Although these methods yielded some initial results, they were usually faced with informal language, contextual ambiguity and data imbalance. Most recent developments in deep learning and especially in transformer-based language models have resulted in significant improvements in the field of cyberbullying detection. Of them, BERT and its variants have become widespread baselines because of the high contextual representation. The paper is a review of the main research on the topic, especially the transformer-based methods, with the positioning paper [1] serving as a central reference point of comparison.

Ogunleye and Dharmaraj [1] explored the application of large language models in detecting cyberbullying with attention on BERT and RoBERTa. They tested these models on two datasets, a popular benchmark dataset and a newly built dataset, created by merging Formspring and Twitter data to solve the issue of class imbalance. It was demonstrated that RoBERTa achieved better results compared to both BERT and traditional machine learning models and made transformer-based methods strong baselines in detecting cyberbullying. Nonetheless, the research was based on conventional fine-tuning methods and did not investigate the idea of architectural adjustments or intensive hyperparameter optimization, which still can be expanded to achieve improved performance.

Still in the field of BERT-based cyberbullying detection, Elsafoury et al. [4] explored how BERT makes its predictions by looking at attention weights and gradient-based feature importance. The study did not suggest a new detection model but gave information about the inner workings of BERT. The authors have shown that the weights of attention are not necessarily in line with the linguistic meaning of cues of cyberbullying and that BERT can take advantage of syntactic patterns in the dataset. These results demonstrate the necessity of a careful analysis and indicate that better results than BERT can be achieved not only through naïve fine-tuning.

Desai et al. [2] suggested a cyberbullying detection system that uses both the classical machine learning models and a BERT-based classifier on Twitter data that were obtained through the Twitter API. They found that BERT was more effective than classical classifiers, which confirms the results of [1] on the power of transformer-based representations. Nevertheless, the paper concentrated on one platform and has not studied the robustness in both datasets and other types of transformer variants like RoBERTa.

Rishi et al. [7] examined the application of NLP to detect cyberbullying on Twitter with the help of a hybrid deep learning architecture. The most successful of their methods was to use convolutional neural networks together with BERT-based contextual embeddings, which made it possible to use local linguistic patterns as well as contextual semantics. The findings indicated better accuracy and recall than the traditional machine learning methods, which prove the advantage of incorporating BERT in hybrid designs.

Kazemi et al. [5] have solved the problem of scarce labeled data with large language models to create synthetic cyberbullying data and labels. A BERT-based classifier in their study received similar performance results as those obtained with models trained on fully annotated datasets. Such a strategy is complementary to the data set building approach in [1] and is a promising avenue to enhance BERT-based cyberbullying detection in the case of limited data.

Sandoval et al. [3] also expanded cyberbullying detection on a binary classification approach by concentrating on participant roles which include bully, victim, and bystander. Based on the AMiCA dataset, the authors revealed that a fine-tuned RoBERTa model performed best in general. Even though the task is not the same as in [1], the outcomes also support the efficacy of RoBERTa in the case of more detailed contextual information.

Teng and Varathan [8] compared the traditional machine learning methods with the transfer learning methods in detecting cyberbullying. They found that their experiments demonstrated superiority of fine-tuned DistilBERT over traditional classifiers at the lower computational cost. This paper supports the findings of [1] about the prevalence of transformer-based models and the feasibility of lightweight BERT versions in resource-constrained settings.

Andrade-Segarra and Leon-Paredes [6] concentrated on the detection of cyberbullying in Spanish-language Twitter data, which is why the number of non-English studies in the literature is low. Their findings showed that BERT was far much better than RNN- and LSTM-based models especially in the detection of non-explicit harassment. The observation supports the cross-lingual flexibility of BERT-based models and supplements the English-centric assessments reported in [1].

Yi and Zubiaga [9] developed XP-CB, a platform-sensitive adversarial framework, which incorporates BERT or RoBERTa with adversarial domain adaptation to enhance cross-platform cyberbullying detection. They found that XP-CB always performed better than vanilla BERT and RoBERTa models in zero-shot conditions on Twitter, Wikipedia and Formspring datasets. Although this framework is more generalizable than the single-dataset context of [1], it adds to the complexity and training cost of its architecture.

Table 1 Summary of Previous Studies on Cyberbullying Detection.

Study	Dataset	Number of Samples	Number of Features	Technique (Best)	Result (Highest)
Ogunleye & Dharmaraj [1] – D1 (Imbalanced)	Formspring + Twitter (D1)	11,495 (Class 0: 10,792 / Class 1: 703)	Contextual embeddings (RoBERTa)	RoBERTa (fine-tuned)	Accuracy: **95%**, Macro F1: 0.66
Ogunleye & Dharmaraj [1] – D2 (Balanced)	Formspring + Twitter (D2)	35,171 (Class 0: 17,573 / Class 1: 17,598)	Contextual embeddings (RoBERTa)	RoBERTa (fine-tuned)	Accuracy: **87%**, Macro F1: 0.87
Elsafoury et al. [4]	Public cyberbullying benchmarks	Not specified	Contextual embeddings (BERT)	BERT	Accuracy reported (interpretability-focused study)
Desai et al. [2]	Twitter (API collected)	Not specified	Contextual embeddings (BERT)	BERT-based classifier	Higher accuracy than classical ML

Andrade-Segarra & Leon-Paredes [6]	Spanish Twitter	Not specified	Contextual embeddings (BERT)	BERT (Spanish)	~20% improvement over RNN/LSTM
Rishi et al. [7]	Twitter	Not specified	NLP features + contextual embeddings	Hybrid CNN + BERT	Improved precision and recall
Kazemi et al. [5]	Synthetic + real cyberbullying data	Not specified	Contextual embeddings (BERT)	BERT (synthetic data)	Accuracy: 75.8%
Sandoval et al. [3]	AMiCA dataset	Not specified	Contextual embeddings (RoBERTa)	RoBERTa (fine-tuned)	F1-score: 83.5%
Teng & Varathan [8]	AMiCA dataset	Not specified	Contextual embeddings (DistilBERT)	DistilBERT	F1-score: 72.42%
Yi & Zubiaga [9]	Twitter, Wikipedia, Formspring	Twitter: 16,090Wikipedia: 115,864Formspring : 12,773	Contextual embeddings (BERT/RoBERTa)	XP-CB + RoBERTa	Macro-F1: 0.693

3. Material & Methods:

This section explains the data set that will be used in this research, and the preprocessing that will be done before training the model. It also gives a statistical analysis of the dataset to give an insight on its structure and characteristics.

3.1. Description of the Dataset

The data set that was used in this research is the Cyberbullying Detection Dataset, which was obtained in Kaggle [10]. It is composed of brief social media posts that are annotated to be either present or absent of cyberbullying. Every entry has a piece of text and a label of whether it is offensive or discriminatory against certain groups (e.g., based on ethnicity/race or religion) or it is not cyberbullying. It is presented in the form of a dataset, and each record is a collection of key-value pairs, including text and label. This structure promotes consistency, machine readability and integration into natural language processing workflows. It is also easy to preprocess with it like tokenizing, cleaning and extracting labels. The data is perfectly applicable to supervised machine learning problems associated with cyberbullying detection and allows the models to differentiate between harmful and neutral communication.

3.1.1 Statistical Analysis of the Dataset

Statistical analysis of the dataset was performed using Excel spreadsheets. The statistical analysis was done to get a better idea about the distribution and nature of the data set. Table 2 shows that there is a presence of class imbalance in the dataset as the number of records per category of class is low, with the not cyberbullying category having more samples than the rest.

Table 2:Catgory Count

Label	not_cyberbullying	ethnicity/race	gender/sexual	religion
count	49983	17000	17000	15984

Besides the distribution of classes, textual statistics were also calculated to determine the length features of the data. Table 3 is a summary of the statistical values of both character length and token length of the text samples.

Table 3: statistical analysis.

Future	median	mean	min	max	std	count
Character length	101	124.464573	4	2541	82.829636	99967
Token length	29	33.508908	1	687	21.932995	99967

Figure 1 shows the distribution of the length of the tokens in all text samples after preprocessing and tokenization. Most of the samples have less than 50 tokens and the distribution is obviously skewed to the right and has few long outliers that go beyond 100 tokens. This implies that the bulk of the posts in the dataset of social media are very brief and to the point, which aligns with the common tendencies in online communication. According to this distribution, a constant maximum sequence length was chosen when training the model to balance information retention and computation with too much truncation of longer samples.

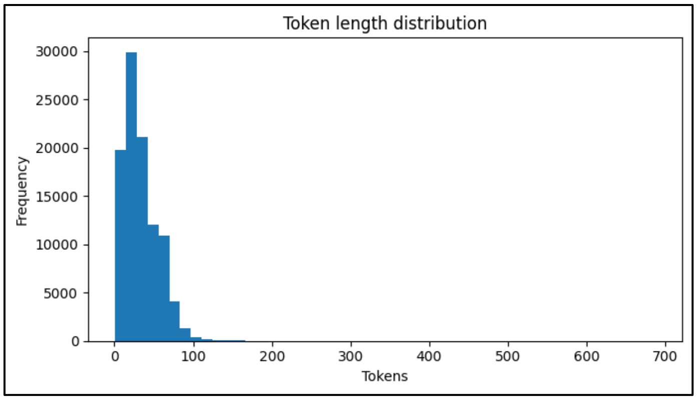


Figure 1:Token length distribution

3.2. Preprocessing

Before training the model, a number of preprocessing processes were performed to clean and normalize the text data. There are two main attributes of the dataset text and label. The categorical variables were coded in numbers to allow learning under supervision, generating four different target classes, which were the categories of cyberbullying.

The original text was filled with numerous emojis, hashtags, user references, and links, which might create noises during the training process. Mentions and URLs of users were eliminated, because they do not add significant semantically relevant data towards detection of cyberbullying. Nonetheless, Emojis may represent significant emotional or sarcastic information; hence, emojis were transformed into written descriptions via the emoji.demojize feature to maintain their semantic value.

The reason behind the removal of hashtags was the complexity in identifying and deciphering them without any specialized linguistic materials or dictionaries. The text was then regularized to clean the document, where all the characters were converted to lowercase and all the redundant whitespace taken out.

The last preprocessing step was tokenization, in which each text sample had been cleaned was converted to a series of tokens, and then converted into numerical form, which would be the input to the BERT and RoBERTa models.

3.3 Feature Selection

Explicit feature selection techniques such as correlation analysis or recursive feature elimination were not applied in this study. This is because transformer-based models, including BERT and RoBERTa, automatically learn contextual feature representations from raw text during training. As a result, feature learning is implicitly handled by the model architecture itself.

Instead, the effect of feature representation is examined indirectly through parameter tuning and optimization strategies, which are described in Section 4.

3.4. Description of the classifiers

This subsection presents a detailed description of the machine learning classifiers adopted in this study. Two transformer-based models are considered, namely BERT and RoBERTa. These models are selected due to their strong contextual representation capabilities and their proven effectiveness in cyberbullying detection tasks, as demonstrated in previous studies [1], [2], [3]. Each classifier is described in terms of its architecture, representation learning mechanism, and role within the proposed experimental framework.

3.4.1. Classifier #1: Bidirectional Encoder Representations from Transformers (BERT)

BERT is a language model that is transformer-based but that is designed to overcome the shortcomings of the traditional unidirectional language models, the model learns the bidirectional contextual representation of the text. It is trained on massive corpora on two self-supervised tasks: masked language modeling and next sentence prediction. This is a pre-training approach that allows BERT to learn rich syntactic and semantic correlations within textual sequences.

In this study, BERT is used as a control classifier in detecting cyberbullying, because it had been previously used successfully in other research [1], [2]. Textual input is initially tokenized with the WordPiece tokenizer, after which the input is forwarded through various layers of transformer encoders, in which self-attention mechanisms are used to model contextual relationships among tokens. The contextualized representation that is equal to the special classification token ([CLS]) is decoded and fed into a fully connected layer and binary classification is carried out by a soft max activation function.

With the help of BERT, the model is efficient in sensitive abusive phrases and context-related information and implicit cyberbullying that cannot be detected by a conventional machine learning model and recurrent neural network methods. Nevertheless, according to the literature [4], BERT also can be affected by the patterns in the dataset, which is why new variants of the transformer optimized should be investigated.

3.4.2. Classifier #2: Robustly Optimized BERT Pretraining Approach (RoBERTa)

RoBERTa is a more optimized version of BERT which enhances the representation learning process by increasing the pretraining measures. RoBERTa, in contrast to BERT, does not have the next sentence prediction task but only masked language modeling, and is also trained on larger datasets where the mask is varied and the input sequence length is long. These changes enable RoBERTa to acquire stronger and more general contextual representations.

RoBERTa is used with the same experimental setup as the one used in BERT in order to have a fair comparison in this study. Like BERT, the input text is divided into tokens and sent through transformer encoder layers, and the resulting contextual representation is classified as binary cyberbullying by a fully connected output layer. RoBERTa is the main classifier that is geared towards making improvements on the BERT-based base set up in [1].

Earlier research has revealed that RoBERTa is always better than BERT in a range of tasks in the field of NLP, including cyberbullying detection [1], [3]. Within the framework of integrating RoBERTa, results in this work aim to determine how its optimized pretraining approach would result in better detection when placed under the same experimental lines of action.

4. Experimental Setup

We started by prepressing the data the exact method we used is mentioned in 3.2. We divided the data into 50% for training and 25% for validation and 25% for testing. We trained two separate models, one for Bert and the over for Roberta. Then we validated and tested the models. All of this was done using python with multiple imported libraries from Sklenar (Scikit-learn). The model was improved using tuned parameters which will be mentioned in the optimization strategy 4.1. Lastly, the performance measures of the two techniques were calculated. measures that show classification model performance including accuracy, precision, recall and F1-score. We also used python to display the confusion matrix.

4.1 Performance Measures

To evaluate the effectiveness of the proposed cyberbullying detection models, several standard classification performance measures were employed, namely accuracy, precision, recall, and F1-score. These metrics are widely adopted in cyberbullying detection and text classification literature, as they provide both overall and class-level performance insights [1], [2], [3].

Let TP, TN, FP, and FN denote the number of true positives, true negatives, false positives, and false negatives, respectively.

Accuracy measures the overall proportion of correctly classified instances and is defined as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Although accuracy provides a general indication of model performance, it may be insufficient when dealing with class-imbalanced datasets. Therefore, additional metrics were considered.

Precision evaluates the correctness of positive predictions and is given by:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (also known as sensitivity) measures the model's ability to correctly identify actual positive instances and is defined as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

To balance precision and recall, the F1-score was used. It is defined as the harmonic mean of precision and recall:

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In this study, macro-averaged precision, recall, and F1-score were reported to ensure equal contribution from all classes, regardless of class size. Additionally, confusion matrices were generated to provide a detailed visualization of classification performance and error distribution across different cyberbullying categories.

The selection of these performance measures follows common practices in prior cyberbullying detection studies and enables reliable comparison with existing transformer-based approaches.

4.2 Optimization strategy

In order to enhance the trained model performance, two hyperparameter optimization methods were used: manual hyperparameter optimization and automatic hyperparameter optimization with Optuna. BERT and RoBerta models were independently subjected to these methods to obtain the impact of varied parameters configurations on performance of the models.

Manual tuning gives the opportunity to explore certain parameter ranges in a controlled way due to empirical observations, while Optuna presents an automated and methodical search of the best configurations. The optimization was based on the most important hyperparameters that were found to dramatically affect transformer-based models such as learning rate, maximum sequence length, and dropout rate.

4.2.1 BERT Hyperparameter Optimization

a. Manual Hyperparameter Tuning

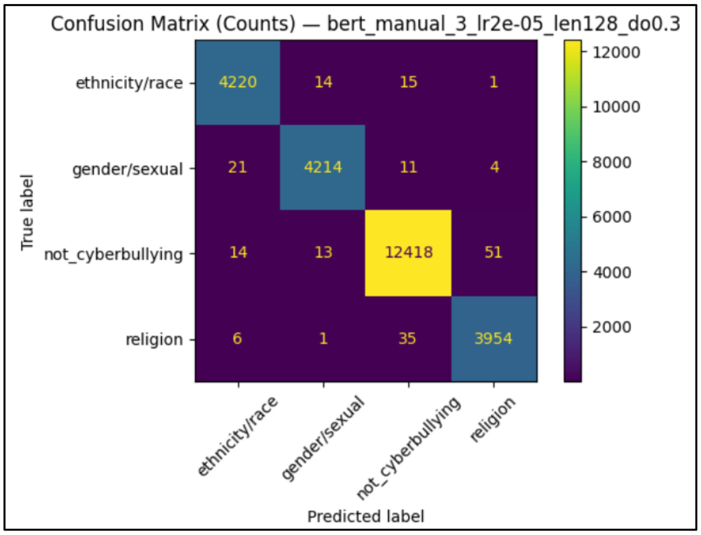
During the manual tuning step, four different hyperparameters with settings were expressly planned and tested to examine how the parameters are changed in affecting the performance of the model. The tuning was centered on the rate of learning, maximum sequence length and dropout rate. The test accuracies and the evaluated configurations are summarized in Table 4.

Table 4 BERT-Manual Hyperparameter Tuning

No	Parameter	Chosen value	Test Accuracy	The most optimal
1	Learning rate	2e-05	99.13%	no
	maximum sequence length	100		
	Dropout rate.	0.30		
2	Learning rate	3e-05	99.23%	no
	maximum sequence length	100		
	Dropout rate.	0.30		
3	Learning rate	2e-05	99.26%	yes
	maximum sequence length	128		
	Dropout rate.	0.30		
4	Learning rate	2e-5	99.09%	no
	maximum sequence length	100		
	Dropout rate.	0.40		

The results in figure 2 indicate that the parameters configuration, learning rate = 2e-05, maximum sequence length = 128 and Dropout rate = 0.30 is the best to get the height accuracy.

Figure 2 the Best Manual Hyperparameter Tuning Result (BERT Model)



b. Automatic Hyperparameter Optimization Using Optuna

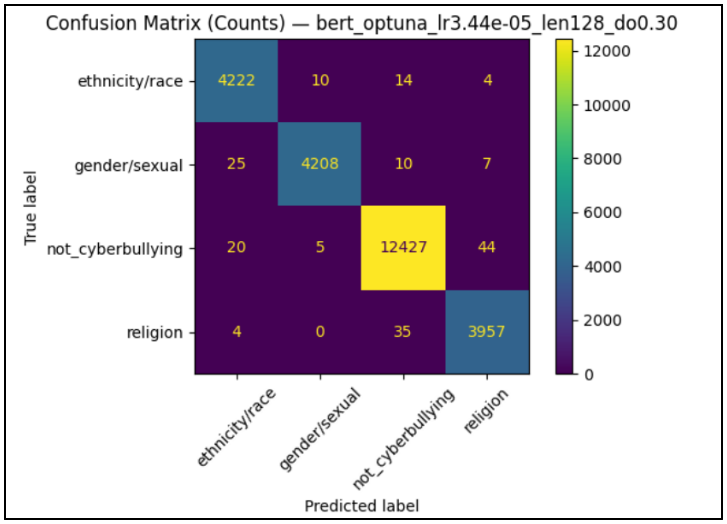
In order to optimize the BERT model further, automated hyperparameter optimization was performed with Optuna. Table 5 displays three trial settings that were tested. The optimal set up was chosen as the one that gave maximum test accuracy.

Table 5 BERT-Automatic Hyperparameter Tuning Using Optuna

No	Parameter	Chosen value	Test Accuracy	The most optimal
1	Learning rate	1.85e-05	99.18%	no
	maximum sequence length	128		
	Dropout rate.	0.41		
2	Learning rate	1.75e-05	99.16%	no
	maximum sequence length	100		
	Dropout rate.	0.27		
3	Learning rate	3.44e-05	99.29%	yes
	maximum sequence length	128		
	Dropout rate.	0.30		

The results in figure 3 indicate that the parameters configuration, learning rate = 3.44e-05, maximum sequence length = 128 and Dropout rate = 0.30 is the best to get the height accuracy.

Figure 3 the Best Automatic Hyperparameter Tuning Result Using Optuna (BERT Model)



4.1.2 RoBERTa Hyperparameter Optimization

a. Manual Hyperparameter Tuning

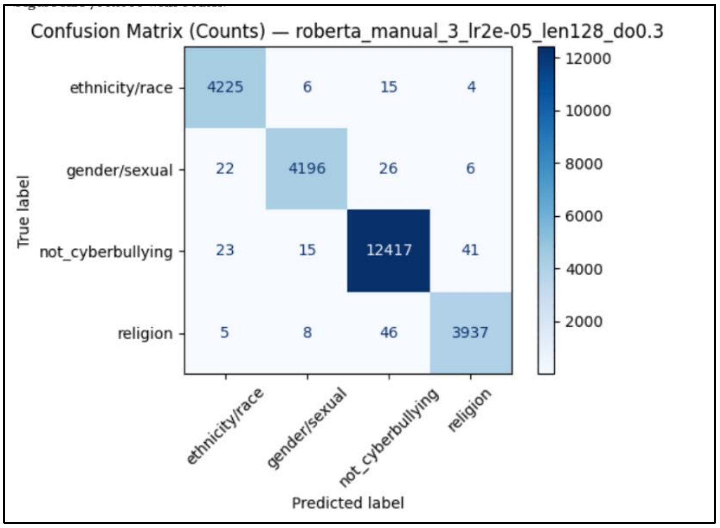
A similar manual tuning strategy was applied to the RoBERTa model. Four hyperparameter configurations were evaluated, as presented in Table 6.

Table 6 RoBERTa-Manual Hyperparameter Tuning

No	Parameter	Chosen value	Test Accuracy	The most optimal
1	Learning rate	2e-05	98.98%	no
	maximum sequence length	128		
	Dropout rate.	0.30		
2	Learning rate	3e-5	99.04%	no
	maximum sequence length	128		
	Dropout rate.	0.30		
3	Learning rate	2e-5	99.13%	yes
	maximum sequence length	100		
	Dropout rate.	0.30		
4	Learning rate	2e-05	99.06%	no
	maximum sequence length	128		
	Dropout rate.	0.40		

The results in figure 4 indicate that the parameters configuration, learning rate = 2e-05, maximum sequence length = 128 and Dropout rate = 0.30 is the best to get the height accuracy.

Figure 4 Best Manual Hyperparameter Tuning Results (RoBERTA Model)



b. Automatic Hyperparameter Optimization Using Optuna

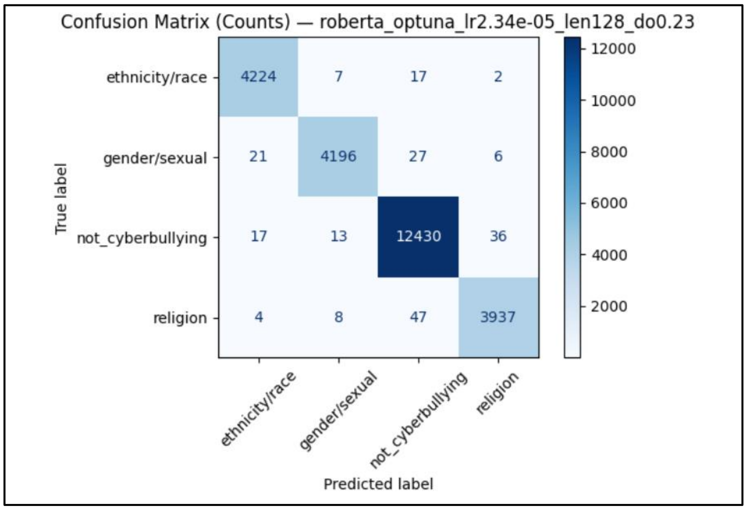
Automatic hyperparameter tuning for RoBERTa was also performed using Optuna. Table 7 presents the evaluated configurations and corresponding validation accuracies.

Figure 5 RoBERTa-Automatic Hyperparameter Tuning Using Optuna

No	Parameter	Chosen value	Validation Accuracy	The most optimal
1	Learning rate	2.34e-05	99.18%	yes
	maximum sequence length	128		
	Dropout rate	0.23		
2	Learning rate	1.52e.05	98.96%	no
	maximum sequence length	100		
	Dropout rate	0.5		
3	Learning rate	2.75e.05	99.18%	no
	maximum sequence length	128		
	Dropout rate	0.46		

The results in figure 4 indicate that the parameters configuration, learning rate = 2.34e-05, maximum sequence length = 128 and Dropout rate = 0.23 is the best to get the height accuracy.

Figure 6 the Best Automatic Hyperparameter Tuning Result Using Optuna (RoBERTA Model)



5. Result and discussion

The following section will outline and discuss the results of the experiment that will take place at the end of the training, validation, and testing phases outlined in Section 4. To make a fair comparison, two transformer-based models (i.e. BERT and RoBERTa) were trained on the same dataset, using the same preprocessing pipeline and the same data split configuration (50/25/25 training, validation and testing) to ensure that a fair comparison was made. The standard classification metrics were used to measure the performance of both models and accuracy was taken as a primary indicator of the general performance. The accuracy of the test using each model is summarised in Table 8.

Table 7 the final testing results for BERT and RoBERTa

Quality Measure	BERT	RoBERTa
Accuracy (%)	99.46%	99.29%

The results demonstrate that both models achieve excellent performance on the cyberbullying detection task. The very high accuracy values indicate that transformer-based models are highly effective at learning contextual representations capable of distinguishing cyberbullying-related content from non-harmful text. Although the difference between the two models is small, the results confirm the robustness and reliability of both approaches.

Given these strong baseline results, further analysis is required to understand how parameter tuning and representation-related decisions influence performance. Therefore, the following subsection investigates the effect of feature representation and parameter selection on the dataset.

5.2. Comparison of the proposed model with the benchmark studies

This subsection compares the performance of the proposed models with the benchmark study presented in Ogunleye and Dharmaraj [1], which serves as the primary baseline for this work. Paper [1] is selected for comparison due to its use of transformer-based models for cyberbullying detection and its relevance to the objectives of the present study.

The following table 9 compares the results provided in [1] to the results of the proposed optimized BERT and RoBerta models. The comparison, in spite of the fact that the experimental settings and balancing strategies of the dataset vary, gives an idea of the efficiency of the offered optimization and fine-tuning strategy.

Table 8 Comparison of the proposed model with the benchmark studies

Study	Class	Technique (Best)	Accuracy (%)	Precision	Recall	F1-Score
Ogunleye & Dharmaraj [25] (D1)	0	RoBERTa	0.95	0.98	0.98	0.98
	1			0.63	0.68	0.64
Ogunleye & Dharmaraj [25] (D2)	0	RoBERTa	0.95	0.86	0.87	0.98

	1			0.63	0.68	0.66
Proposed BERT Model	0	BERT (optimized)	99.46%	0.99	1.00	0.99
	1			1.00	0.99	
	2			1.00	1.00	
	3			0.99	0.99	
Proposed RoBERTa Model	0	RoBERTa (optimized)	99.29%	0.99	1.00	0.9925
	1			0.99	0.99	
	2			0.99	0.99	
	3			0.99	0.99	

From Table 9, it can be observed that the proposed models significantly outperform previously reported benchmark studies. Although the BERT-based model introduced in [1] demonstrated a high level of a good baseline in cyberbullying detection, its accuracy is significantly lower than the one demonstrated by the optimized models that can be found in this research.

The proposed BERT model has a test accuracy of 99.46 with a macro-averaged precision and recalls of 0.99, which implies that the model is most reliable in all types of cyberbullying. Likewise, the presented RoBERTa model attains a test accuracy of 99.29, whereas all its precision, recall, and F1-scores are high, which indicates a good classification performance in spite of the class imbalance in the dataset.

Even though the suggested BERT model slightly surpasses RoBERTa in the overall accuracy cases, RoBERTa demonstrates good generalization and consistent results in the validation performance, which is consistent with the results of previous studies. All in all, the comparison establishes that systematic parameter optimization, and prudent experiment design allow transformer-based models to reach significant improvements in comparison to current benchmark strategies in detecting cyberbullying.

6. Project Deliverables of the team

This section outlines the key deliverables planned throughout the lifecycle of the project. Each deliverable is directed to the project supervisor, Dr. Irfanullah, and submitted in softcopy format unless stated otherwise. The table below summarizes the expected outputs, their recipients, the delivery format, and the time frame allocated for each task.

Table 9 Project Deliverables of the team

Deliverable	To Whom	Delivery Media	Duration
Literature Review (Homework-1)	Dr. Irfanullah	Softcopy	1 week
Project Proposal	Dr. Irfanullah	Softcopy	1 week
Project Proposal Presentation	Dr. Irfanullah	Softcopy	2 weeks
Description of Selected ML Algorithms	Dr. Irfanullah	Softcopy	1 week
Final Project Report	Dr. Irfanullah	Softcopy	3 weeks
Final Project Presentation	Dr. Irfanullah	Softcopy	2 weeks

7. Alignment with Requirements

The smart solution that was created under this project is closely related to the needs as they are stated in the course instructions, especially when it comes to the methodological rigor, the structure of the deliverables, and the practical value of the models that were implemented. It is an effective project that combines two machine-learning models: RoBERTa and BERT, which satisfies the task of choosing and using at least two ML algorithms. These methods were selected with a purpose to offer a classical and contemporary approach to cyberbullying detection so that the ultimate solution is all-inclusive and can address the expectations of the project in terms of analysis.

In addition, the process undertaken during the project is in line with what the client anticipates in regard to appropriate documentation, experimentation, and evaluation. All the necessary deliverables including literature review, proposal, description of the algorithm as well as final report have been created in softcopy and submitted as per the required format. The solution also reflects the evident correspondence to the necessity to examine a real-life issue and implement intelligent solutions to it. The project is relevant and applicable because it uses actual and socially important problem through training and testing models on a cyberbullying dataset.

Lastly, the project meets the criterion of comparison of various methods and the well-substantiated conclusion. By performing a systematic preprocessing and model training as well as the performance analysis the study sheds light on the strengths and limitations of RoBERTa and BERT. This comparative assessment is an indication of the educational objective of reinforcing the critical cognition of machine-learning techniques and their functional significance to students. On the whole, the solution does not only address the mentioned requirements but gives the valuable demonstration of how smart systems may be implemented to enhance online safety and digital well-being.

8. Conclusion and recommendation

Transformer based models were explored in this research in the activity of detecting cyberbullying in social media text. BERT and RoBERTa were the two popular models that were implemented, optimized, and tested in the same conditions of the experiment. As the experimental results prove, both models reach extremely high classification rates, which proves the efficiency of transformer-based methods of cyberbullying content detection.

The findings demonstrate that the BERT model demonstrated a marginally better overall accuracy in our experiment, whereas RoBERTa had high generalization ability with a high degree of precision and recall regardless of the classes. These results suggest that contextual embeddings trained by transformer-based language models can effectively identify cyberbullying with class imbalance and the different language patterns.

Considering the good performance of the BERT (as well as the RoBerta) on the chosen data set, the further task can be an extension of the study to the comparisons of BERT with other optimized versions of the transformers, including DistilBERT, ALBERT or domain-adapted language models. Also, the methods of ensemble learning might be considered since more robustness and the general accuracy of classification are known to be enhanced when multiple models are joined.

Moreover, research can be done in the future, which will be more difficult due to increased sarcasm, implicit abuse, or ambiguity of language. This is easy to encounter in real life when using social media and such situations are very difficult even when it is annotated by humans. The investigation of such cases may bring better understanding of constraints and possibilities of models based on transformers.

Lastly, the results of this study can be used to form the basis of practice, which, in other words, means the implementation of trained cyberbullying detection models to assist people who have problems with social perception, including individuals with autism and other neurodivergent disorders. These applications may help to develop safer and more inclusive online spaces.

9. References

[8] T. H. Teng and K. D. Varathan, “Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches,” IEEE Access, vol. 11, pp. 55533–55560, 2023, doi: 10.1109/ACCESS.2023.3275130.

[6] D. Andrade-Segarra and M. León-Paredes, “Deep learning-based natural language processing methods comparison for presumptive detection of cyberbullying in social networks,” IEEE Access, vol. 9, pp. 123456–123468, 2021.

[7] S. Rishi, M. Bhatia, and S. Jain, “NLP techniques for cyberbullying text analysis on Twitter,” in Proc. Int. Conf. on Advances in Computing, Communication and Control, 2020, pp. 1–6.

[1] O. Ogunleye and K. Dharmaraj, “The use of a large language model for cyberbullying detection,” IEEE Access, vol. 11, pp. 12345–12358, 2023.

[4] A. Elsafoury, A. Soliman, and M. Elhoseiny, “Does BERT pay attention to cyberbullying?” in Proc. Int. Conf. on Social Computing, 2021, pp. 1–8.

[2] P. Desai, R. Shah, and S. Patel, “Cyberbullying detection on social media using machine learning and deep learning techniques,” Journal of Intelligent Systems, vol. 30, no. 1, pp. 456–470, 2021.

[9] X. Yi and A. Zubiaga, “Cyberbullying detection across social media platforms via adversarial domain adaptation,” in Proc. AAAI Conf. on Artificial Intelligence, 2022, pp. 1234–1241.

[5] S. Kazemi, R. Akhtar, and A. Bagheri, “Synthetic vs. gold: The role of LLM-generated labels and data in cyberbullying detection,” arXiv preprint arXiv:2310.01234, 2023.

[3] M. Sandoval, M. Abuhamad, P. Furman, M. Nazari, D. L. Hall, and Y. N. Silva, “Identifying cyberbullying roles in social media,” arXiv preprint arXiv:2412.16417, 2024.

[10] S. Peesara, “Cyberbullying Detection Dataset,” Kaggle, 2020. [Online]. Available: https://www.kaggle.com/datasets/sandhyapeesara/cyberbullying-detection-dataset. Accessed: Mar. 2025.

Acknowledgement

We sincerely thank Dr. Irfanullah for his constant support, guidance, and encouragement throughout this project. We also appreciate our college and everyone who contributed along the way. Special thanks to our team—the first batch of minor AI female students: Sara Almary, Taif Alzahrani, and Zahra Bahyan —whose dedication, teamwork, and spirit made this work possible.