National College of Ireland

A Deep Learning-Based Approach for Detecting Cyberbullying in Low-Resource Environments

MSc Research Project

Msc in Data Analytics

Venkata padmavathi kankanampati

Student ID: X23287519

School of Computing

National College of Ireland

Supervisor:David Hamil

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Venkata padmavathi kankanampati

Student ID: X23287519

Programme: Msc in Data Analytics **Year:** 2025

Module: Research practicum

Supervisor: David Hamil

Submission Due Date: 11th dec 2025

Project Title: A Deep Learning-Based Approach for Detecting Cyberbullying in Low-Resource Environments

Word Count: 8112 **Page Count**: 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: k.v.padmavathi

Date: 11th dec 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

A Deep Learning-Based Approach for Detecting Cyberbullying in Low-Resource Environments

Venkata padmavathi kankanampati

X23287519

Abstract

Cyberbullying has become a serious issue in the digital era, with the rapid growth of online communication amplifying the spread of abusive and harmful content. Detecting such behavior is a challenging task due to the unstructured, context-dependent, and informal nature of social media text. Existing approaches based on traditional machine learning or deep learning frameworks often struggle to balance detection accuracy with computational efficiency, especially in low-resource environments. Many models fail to interpret subtle linguistic cues such as sarcasm, slang, or implicit aggression, limiting their reliability in real-world settings. This research presents a lightweight deep learning framework using a Gated Recurrent Unit (GRU) integrated with an Attention mechanism to enhance contextual understanding while maintaining low computational cost. The GRU component effectively captures sequential dependencies in text, while the Attention layer dynamically focuses on significant words that influence meaning, leading to improved interpretability and detection precision. A structured experimental design was implemented to compare this architecture with traditional models using standard evaluation metrics. Results indicate that the **GRU–Attention framework** achieved superior performance, with an accuracy, precision, recall, and F1-score of 0.94, surpassing all baseline models. The findings confirm that the GRU–Attention model provides an efficient, accurate, and scalable solution for cyberbullying detection in resource-constrained environments.

1 Introduction

1.1 Background Study

The last couple of years have seen an immense growth in the sphere of social media and other internet-based communication tools that has altered the way people communicate with each other and share information. Despite the numerous benefits associated with this change, bad internet behaviour has also become possible such as cyberbullying. Cyberbullying refers to the intentional use of digital communication to intimidate, harass, or humiliate other people and has become a critical social and psychological problem at the international level. Not only does it affect the mental health of the victims, but it also affects the integrity of the online communities. The amount of user-generated information present on websites, such as Twitter, Facebook, and YouTube, is a complex issue to the automatic detection systems, as abusive language may have many different manifestations, including slang, sarcasm, and code-mixed text. In its turn, this has rendered the development of resource-efficient, scalable and efficient cyberbullying detection systems an important area of research in both computer science and social computing.

The main machine learning models that have been used to detect cyberbullying are Support Vector Machines, Random Forests, and Naive Bayes classifiers that, although they can reach reasonable accuracy, are not always able to understand the context and can be poor in multilingual or informal online communication (Raj et al., 2022; Perera and Fernando, 2024). Both CNNs, LSTMs, and Transformers have deep learning methods that have addressed some of these limitations by learning semantic and contextual representations of text automatically (Bharti et al., 2022; Aldhyani et al., 2022; Saifullah et al., 2024; Teng and Varathan, 2023). Nevertheless, transformer models are computationally costly and cannot be deployed in the real-time or with low resource needs, but they are very accurate and can only compute long-range dependencies, unlike traditional RNNs (Hoque and Seddiqui, 2024). To overcome them, this study uses a Gated Recurrent Unit (GRU) model with an attention mechanism to create a lightweight and computationally efficient model that can extract significant contextual information. The framework will strive to attain high performance and efficiency balance and is therefore appropriate to detect cyberbullying in real time under resource limited setting like mobile and edge devices.

1.2 Research Question

"How can a Gated Recurrent Unit (GRU) network integrated with an attention mechanism enhance the efficiency and accuracy of cyberbullying detection in low-resource environments?"

Justification

The relevance of this research question lies in the fact that, most of the current cyberbullying detection models have a trade-off between computational efficiency and contextual understanding. Transformer-based models are also extremely accurate but have significant computational demands, and thus cannot be deployed to mobile or low-resource environments. Conversely, more lightweight traditional recurrent models can find it difficult to model long-range text dependencies. The use of a GRU network with an attention mechanism can be viewed as a possible solution because it effectively captures the sequential data, as well as attention to contextually important words. This has the potential to deliver similar performance to larger models with reduced computational requirements, a solution to the urgent requirement of scalable and real-time cyberbullying detection systems that can be effectively used under limited resources.

2 Related Work

2.1 Machine Learning Models for Cyberbullying Detection

The techniques of machine learning (ML) have been used as background solutions to the cyberbullying detection because of the possibility to categorize textual information via engineered linguistic and semantic features. Conventional models like Support Vector Machines (SVM), Naive Bayes, Random Forest, and Logistic Regression have been well

implemented to detect malicious content in the online communications. Such experiments as that conducted by Raj et al. (2022) and Perera and Fernando (2024) have shown that these algorithms are capable of processing word frequency, sentiment polarity, and contextual clues to identify bullying text and neutral messages. These methods are effective on balanced datasets, but the quality of feature extraction and pre-processing is very important, and often handcrafted representations are used, like TF-IDF or bag-of-words, which restricts their applicability to a variety of datasets and languages.

Recent research has addressed hybrid ML models to use Natural Language Processing (NLP) methods to better represent features. Raj et al. (2021) considered statistical text mining and the use of NLP-based embedding methods to improve the contextual understanding. Likewise, Alam et al. (2021) used ensemble-based ML models, which integrates various classifiers to enhance accuracy and robustness toward identifying cyberbullying. The generalization of such ensemble structures on social media datasets was better than that of single classifiers, indicating the increasing demand of multi-feature learning. Nevertheless, these approaches continue to have difficulties with informal, multilingual, and low-resource data in which the handcrafted features do not reflect the profound semantic nuances.

Additionally, ML-based detection models have also been enhanced by the development of emotion- and sentiment-based models. The article by Al-Hashedi et al. (2023) suggested the use of emotion-based ML models, which combine such affective characteristics to recognize the patterns of bullying, whereas Alabdulwahab et al. (2023) applied hybrid ML-DL frameworks to do a more complex detection. Although these improvements have been made, the ML models are still limited by their manual design of features and their inability to make deep contextual interpretations. They usually perform poorly on real-world, noisy datasets, which underscores the utility of automated feature extraction processes and full-end learning processes (which deep learning models provide).

2.2 Deep Learning Models for Cyberbullying Detection

Deep learning (DL) models have transformed the text classification field through their ability to automatically extract features of raw data, without relying on manual feature engineering that is common in traditional ML methods. CNNs and RNNs have been successfully used in detecting cyberbullying, and they have achieved significant gains in accuracy and generalization. The article by Bharti et al. (2022) was based on Twitter datasets processed with a deep learning algorithm, which showed that neural networks could identify the contextual dependencies and minor linguistic signals that traditional ML algorithms cannot detect. Equally, Aldhyani et al. (2022) created a DL-based cyberbullying detection model that is better than the previous ML models in that it features hierarchized feature representation and data-driven learning. These studies determined that deep architectures are more capable of processing text semantics and sentiment of users, and so they are especially useful in the context of social media.

The subsequent developments incorporated hybrid and multi-channel architectures to enhance the detection performance. Alotaibi et al. (2021) suggested a multichannel deep learning architecture which handles text with a series of feature extraction layers, and it is more robust to unstructured and noisy social media posts. Hybrid ML-DL were also investigated by Raj et al. (2021) and Raj et al. (2022), where NLP embeddings are used together with deep networks to get higher accuracy. Transformer-based models and attention mechanisms further developed this field since they enabled the model to pay attention to contextually relevant portions of a text, which resulted in improved interpretability and accuracy. As emphasized by Saifullah et al. (2024) and Teng and Varathan (2023), transformer and transfer learning models like BERT and RoBERTa have the state of the art results in cyberbullying detection tasks, which surpass other recurrent structures to capture long-term dependencies.

Although these deep learning models achieve impressive performance, their large labelled datasets and heavy computing requirements tend to restrict their use to real-time and low-resource applications. Haque and Seddiqui (2024) highlighted the problem of such models in low-resource languages such as Bengali where there is a shortage of data and hardware resources to scale to large sizes. In a bid to overcome these issues, scholars have embarked on exploring lightweight architectures that can strike a balance between performance and efficiency. Almufareh et al. (2025) and Perera and Fernando (2024) also talked about sentiment-integrated DL systems that were aimed at making inferences much faster without compromising accuracy. Such attempts represent a maturation of the research trend in efficient adaptable and lightweight models- opening the door to GRU- and attention-based models specific to constrained settings.

2.3 Comparison with Heavy Deep Learning Models

New transformer-based architectures have demonstrated much better performance of cyberbullying detection, but they are computationally expensive and cannot be implemented in low-resource settings. BERT, RoBERTa, and XLNet models are based on multi-head self-attention mechanisms that provide them with the ability to learn long-range dependencies and contextual relationships in text (Saifullah et al., 2024; Teng and Varathan, 2023). These models are always better in a benchmark testing compared with the traditional recurrent or convolutional networks but their massive parameterization, memory requirements, and long training time render them not feasible in real-time or embedded applications. Even multi-channel and hybrid frameworks, like those suggested by Alotaibi et al. (2021) and Aldhyani et al. (2022) have intricate data processing pipelines and demand access to a graphics card, which additionally restricts their applicability to limited computing resources. Haque and Seddiqui (2024) have also mentioned that systems that use transformers are not very effective at generalizing when trained on low-resource languages, where annotated data is few, and hardware is not readily available.

In order to address these computational and scalability challenges, scientists have studied lightweight architectures that do not add heavy models to strong contextual learning.

Recurrent neural network models, specifically Gated Recurrent Unit (GRU), provide an effective alternative to these since they contain fewer trainable parameters when compared to other models and at the same time, achieve strong sequential learning capabilities. GRU networks can be used together with an attention mechanism to identify semantically meaningful features of the input text, which enhances interpretability and noise minimisation in predictions (Almufareh et al., 2025). GRU + Attention architectures are less memory-consuming, have shorter training durations, and can compete with models based on transformers on smaller datasets. This renders them suitable to the cyberbullying detection systems which require efficient operation in a low-resource or mobile environment to balance between computational economy and detection accuracy.

Table 1. Summary of Related Works on Cyberbullying Detection

Author & Year	Dataset Used	Model Used	Achievements	Limitations
Raj et al. (2022)	Twitter & social media comments	Machine Learning & Deep Learning hybrid	Achieved better accuracy through hybrid feature integration	Relied on handcrafted features; limited scalability
Bharti et al. (2022)	Twitter dataset	Deep Learning (CNN, LSTM)	Captured contextual features improving accuracy	Required large training data and high computational resources
Teng & Varathan (2023)	Public social network datasets	ML vs. Transfer Learning comparison	Demonstrated high accuracy with transfer learning	Transformer models were resource-heavy
Saifullah et al. (2024)	Benchmark cyberbullying datasets	Transformer-based (BERT, RoBERTa)	Achieved state-of-the-art precision and recall	High computational cost and large model size
Perera & Fernando (2024)	Social media comments	Supervised Machine Learning	Effective for small datasets with simple features	Lacked contextual depth; reduced generalization
Aldhyani et al. (2022)	Social media datasets	Deep Learning (CNN, DNN)	Improved feature learning through automated extraction	High complexity; required GPU-based training
Alabdulwahab et al. (2023)	Text-based cyberbullying datasets	ML & DL comparative models	Compared DL and ML models for detection performance	Moderate results; limited evaluation metrics

Almufareh et al. (2025)	Social media datasets	Sentiment-integrated Machine Learning	Balanced performance and interpretability	Still required optimization for large-scale data
Haider et al. (2023)	Public text datasets	Machine Learning classifiers	Demonstrated ML feasibility for cyberbullying tasks	Lacked deep contextual representation
Aliyeva & Yağanoğlu (2025)	Twitter dataset	Deep Learning (Bi-LSTM, CNN)	High detection accuracy on Twitter posts	Focused on single-language data; limited adaptability

Table 1, gives a summarized review of the major researches done in the field of cyberbullying detection. It describes the datasets, model types, successes and limitations of the previous researches in order to put light on the gaps in the current research.

2.4 Research Gaps

Despite the fact that many machine learning and deep learning models have been created to detect cyberbullying, several weaknesses still face their functionality and applicability. The classical models of machine learning (including Support Vector Machines (SVM), random forest, and the model ensemble) demand handcrafted linguistic and statistical features that, in most cases, do not reflect the intricate semantic relationships in informal text online (Raj et al., 2021; Alam et al., 2021). CNNs and RNNs are examples of deep learning models that have shown better accuracy and contextual knowledge but are typically resource-intensive in terms of both labelled data and computation due to their need to use large labelled datasets and high-performance hardware (Bharti et al., 2022; Aldhyani et al., 2022). Besides, transformer-based models, such as BERT and RoBERTa which are highly effective in capturing long-range interactions, is computationally expensive and cannot be used in real-time or resource-constrained setting (Saifullah et al., 2024; Teng and Varathan, 2023). A number of papers have also noted that the majority of the available models are not flexible to low-resource languages and do not achieve uniform performance when used on smaller datasets or on hardware with low processing power, such as Hoque and Seddiqui (2024) and Al-Hashedi et al. (2023).

To overcome these deficiencies, we need to apply a model that is able to trade-off computational efficiency with good contextual knowledge. Such a balance is achieved with a Gated Recurrent Unit (GRU) architecture that includes an attention mechanism as a means of integrating sequential learning and interpretability. GRU network is effective to capture both temporal and linguistic dependencies in the text data with fewer parameters that reduces computational cost and memory consumption. In combination with an attention mechanism, this model can selectively attend to the most useful words or phrases in a sentence and thus can be better able to identify subtle patterns of harmful or abusive language. This design has

the advantage of not only being able to increase the accuracy of the detention but also train and infer at a faster rate making it useful in real-time. Moreover, the simplicity of GRU and the adaptability of the attention system allow this structure to be deployed in the low-resource setting, where the massive transformer models would have been inefficient.

3 Methodology

3.1 Research Design and Methodological Framework

The methodological design undertaken in this study is based on CRISP-DM (Cross-Industry Standard Process for Data Mining) framework as it is in line with the aim of the study to develop and test the GRU-Attention model systematically. CRISP-DM methodology has six iterative processes, such as Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation, to ensure a well-structured and reproducible process. The Business Understanding stage in this research establishes the issue of cyberbullying detection and the consequences of this issue on online safety. Data Understanding entails gathering and examining textual datasets of social media with the aim of establishing pertinent linguistic and contextual characteristics. In Data Preparation, the Data is cleaned and organized into modelling with the implementation of pre-processing methods like tokenization, stop-word elimination, and word embedding production. The Modelling phase entails the design, training and optimization of the GRU-Attention architecture and the Evaluation phase where the model performance is evaluated by the metrics of accuracy, precision, recall and F1-score. CRISP-DM framework will provide methodological consistency, transparency and flexibility during the research lifecycle that will help in the successful implementation and evaluation of the GRU-Attention-based cyberbullying detection system.

3.2 Dataset Description

The information in this paper is a textual information that will be collected through publicly available interactions on social media, annotated in a way that will allow utilizing it in cyberbullying detection studies. The data instances will have two attributes, namely, the text which is a post or a comment made by the user and label which is a particular type of classification. The labeling scheme consists of a range of classes such as ethnicity/race, religion, and not-cyberbullying which enable the analysis of the discourse online by multi-class. This data is configured in the manner that there is a broad range of linguistic expressions that represent abusive and neutral content and the various types of online communication are represented. The fact that the types of various forms of cyber-bullying are diversified gives the possibility to comprehensively evaluate the capacity of the model to differentiate between aggressive cyber-bullying and non-invasive text in the social media environment.

3.3 Data Pre-Processing

3.3.1 Dataset Inspection and Cleaning

The pre-processing step started with a preliminary examination of the dataset to get an insight into the general structure and properties of the dataset. The data consisted of 99,990 records where one column was text and the other was label which were both of the type of data, string. At this step, the redundant records were identified and eliminated to avoid repetition and to make sure that the training of the model would not be distorted by the repetitive ones. The dataset was then narrowed down to 99,989 distinct instances after this step. Furthermore, a check of the missing or null values ensured that all the fields were filled and there was no imputation necessary. Such cleaning process provided a uniform and stable basis of the next stages of pre-processing and analysis. To have a preliminary view of the dataset and the composition of its structure, the exploratory analysis was conducted.

3.3.2 Text Standardization and Normalization

The input standardization was done a few times to normalize the text data in order to minimize noise. All the text entries were reduced to small letters, which would be shown in the same way. The irrelevant artefacts and symbols were removed and included URLs, mentions, hashtags, punctuation marks, and numeric characters that were irrelevant. Spares were trimmed so as to be uniform. Stopwords (words which carry very little semantic meaning) were also eliminated to ensure that as much as possible the concentration of the model was on meaningful tokens. This was succeeded by lemmatization whereby words were abbreviated to their root or dictionary form such that similar variation of the same word could be treated similarly. These cleaning exercises combined with each other improved the quality and consistency of the linguistic content of the textual corpus and rendered the textual corpus viable to be effectively vectorized.

3.3.3 Label Encoding

Label encoding, which converts categorical text labels into numeric representation, was used to allow compatibility with machine learning algorithms. The dataset was split into several categories, each of which corresponded to a specific integer value, which made the work of the models easier and more straightforward to classify. This change enabled the categories of cyberbullying, including ethnicity/race, religion, gender/sexual and not_cyberbullying, to be effectively translated by the learning framework without compromising the categorical relationships.

3.3.4 Feature Extraction using TF-IDF

In order to convert the text data (cleaned data) into a numerical format that can be entered in deep learning models, the Term Frequency-Inverse Document Frequency (TF-IDF) vectorization method was used. In this procedure, the relative significance of each word was measured by the frequency of words in single documents and rarity in the corpus. A 5,000-

dimension feature space that included unigrams and bigrams was created to represent local word dependencies and contextual relationships. The TF-IDF matrix that was obtained was effective in capturing the semantic nature of the data and was computationally efficient.

3.4 EDA

Exploratory Data Analysis was done to familiarize with the structure, distribution and important features of the dataset. The patterns of class balance, word frequency, and text length were analyzed in different categories through various visualizations. The analysis has given insights that have been used to make preprocessing, feature extraction, and model design decisions.

Figure 1 label distribution reveals that the category not_cyberbullying has the most extensive number of samples, which is why the data is skewed towards classes. Less offensive text is much more widespread than intentional or abusive content, which is a characteristic of social media data. To make sure that the model training is balanced and that the classification can be fair in all categories, it is crucial to identify this imbalance.

Figure 2 presents the word count distribution, which shows the differences between the text length of different dataset categories. A majority of the samples include short messages, 5-20 words, and the ethnicity/race and religion depict slightly longer texts. Aggressive posts are brief in nature, but rich in context, whereas neutral posts are relatively long. These patterns also give important clues on how to optimize the tokenization and sequence length of a model design.

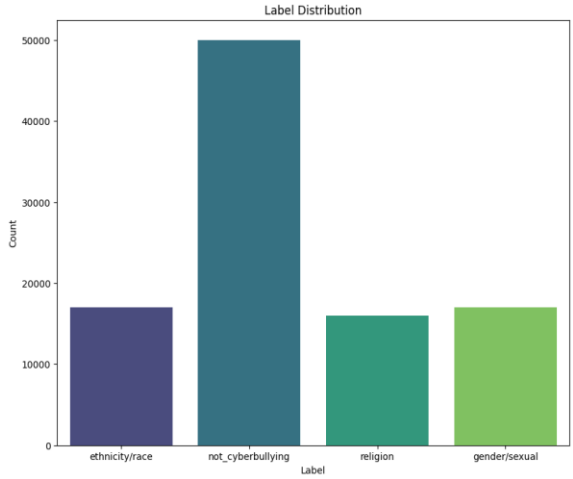

Figure 1. Label Distribution

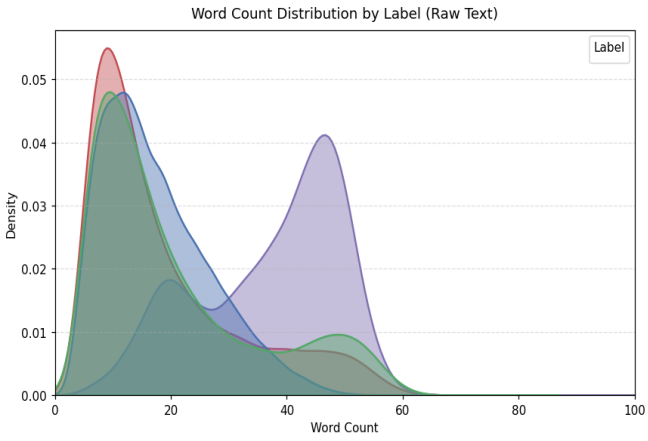

Figure 2. Word Count Distribution (Raw Text)

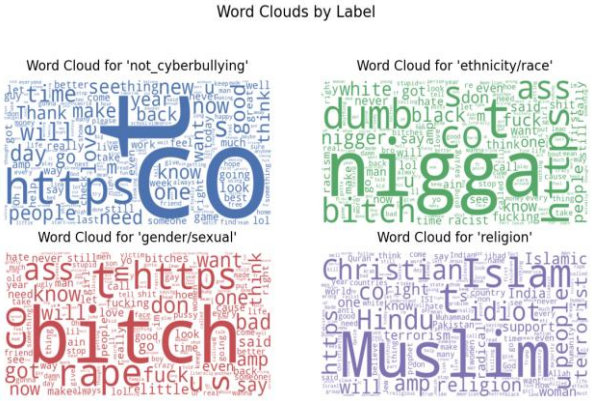

Figure 3. Word Clouds by Label

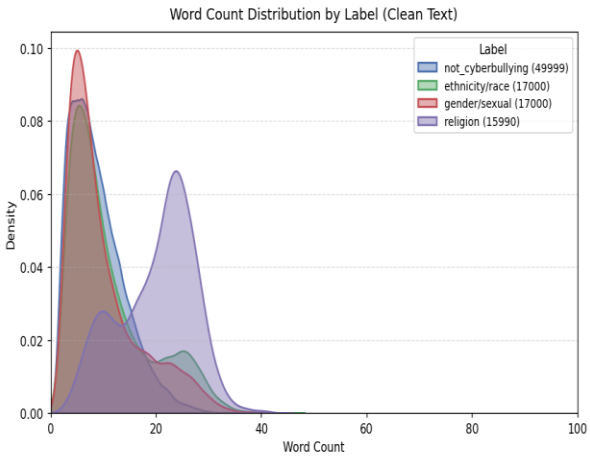

Figure 4. Word Count Distribution (Clean Text)

Figure 3 presents word clouds to depict the most common terms in each category of data sets- not_cyberbullying, ethnicity/race, gender/sexual and religion. The neutral and conversational words prevail in the not cyberbullying category and the other categories exhibit more aggressive or context sensitive words. This visualization allows seeing the different language patterns and emotional tones within categories, which is a qualitative insight, which will additionally be utilized to extract features and develop the model.

The count of words distribution of cleaned text in Figure 4 indicates that the majority of the samples do not have more than 20 words, with the high concentration rates between 5 and 15. This is because of the sharper density as compared to the raw text, here noise reduction using preprocessing is effective. Religion texts are a little longer, which is a sign of deeper context, whereas other categories are shorter and narrower. This proves that cleaning enhanced regularity and maintained significant content to train models.

3.5 Model Selection and Rationale

In order to address the disadvantages of the current literature, the research utilizes a Gated Recurrent Unit (GRU) with an Attention mechanism as the primary cyberbullying detection model. The traditional and transformer-based deep learning models, despite their power, have weaknesses in their computational efficiency, scaling and capability to adapt to real-time, low-resource environments. Transformer models like BERT and RoBERTa are highly accurate and require a lot of processing power and memory, whereas simpler recurrent networks like LSTMs are weak at long-range dependencies. GRU-Attention framework overcomes those issues by integrating the sequential effectiveness of GRU with the comprehensibility and contextual accuracy of the Attention mechanism. GRU is good at capturing temporal trends in text with fewer parameters than LSTM, making rapid training and reducing the computational cost. The Attention layer also increases the capacity of the model to draw attention to semantically relevant words or phrases, which are more effective at identifying subtle cues of cyberbullying (sarcasm, slang, or contextual aggression). This

trade-off is a lightweight but powerful alternative that can be deployed on a device or environment with limited computational power, reaching an accuracy-interpretability-efficiency trade-off.

3.6 Main Model (GRU+Attention) Architecture

The model architecture is a sequential deep learning architecture that tries to capture both time and context dependency of text data. It is based on a sequence of Gated Recurrent Units (GRU) models and an Attention dynamic context weighting mechanism to provide an efficiency/interpretability trade-off.

GRU layer is the primary recurring element that accepts textual input and converts it to the output through the use of time steps. The GRU allows the gating of the flow of information as compared to the traditional recurrent neural networks. The update gate determines the amount of past information required to be stored and the reset gate determines the amount of the past state required to be discarded. The mechanisms are also applicable in the maintenance of long-term dependencies without vanishing gradient issues that render the network propagation of information to be efficient. GRU layer output is the vectors of hidden states which present semantic and syntactic information of every word in the sequence.

The Attention layer is fed with the sequence of the hidden states generated by the GRU. Its major role is to determine what portions of the input sequence make the greatest contribution to the eventual prediction by the model. The mechanism computes a score of attention per hidden state based on a weight matrix that is learnt and a non-linear transformation. The scores are then normalized using a SoftMax function to create a probability distribution of the relative importance of each token. The weights of the results are multiplied in line with the respective hidden states to produce a weighted context vector. It is this context vector that is able to effectively aggregate the critical information in the entire sequence so that the model is able to give more priority to semantically relevant words and repress less informative items. The context vector of the attention system is then high-level feature abstraction with fully connected dense layers. Activation functions create non-linearity allowing the model to represent the intricate decision boundaries between categories. To avoid overfitting, dropout regularization is used, in which neurons are randomly disabled during training. The last SoftMax layer generates normalized probabilities in the target classes, which allow multi-class text classification.

3.6.1 Model Flow Representation

The architecture works by a series of information flow starting with the input phase and passing through a series of transformation layers. The processed text data in form of a numerical feature vector is passed through the input layer which feeds the sequence to the GRU layer. In this layer the time-related associations between words are encoded and stored in a sequence of hidden states vectors. These representations are then transmitted to the Attention layer which allocates a set of attention weights to emphasize the most contextually

important aspects of the sequence. The context vector weighted by the attention is then passed to the dense layers where the high-level abstractions are learnt and perfected using the activation functions. Lastly, the output layer implements a SoftMax algorithm to generate probability distributions across the categories of classification specified. It is a structured flow that allows the model to combine sequential learning with the contextual focus to enhance the accuracy of cyberbullying detection when applied to a variety of textual inputs.

3.7 Evaluation Metrics

To measure the performance of any given model in an objective manner, using the four standard classification measures, accuracy, precision, recall, and F1-score, was used to measure the performance of each of the models built in this study. Accuracy is used to measure the total percentage of correct predictions made over all classes and it shows the general performance of the model. Precision measures the consistency of the model to detect instances of cyberbullying by determining the percentage of actual positives to the total number of predicted positives and reflects the model capability to avoid false alarms. Recall determines the ability of the model to identify the actual cases of cyberbullying by determining the number of actual cases of bullying that were accurately identified among all actual positives. Lastly, the F1-score, the harmonic mean of precision and recall, is a balanced score as it combines missed detections with wrong classification. Collectively, these measures provide an oversight of the strengths and weaknesses of each of the models, especially in the case of unbalanced datasets of cyberbullying.

4 Design Specification

4.1 System Architecture Overview

The research system architecture is configured to facilitate an entire end-to-end cyberbullying detection pipeline that will start with the acquisition of raw data and move through organized processes of text preparation, modelling, and evaluation. In essence, the architecture will combine both classic machine learning algorithms, as well as sophisticated deep learning frameworks, such as the proposed GRU-Attention framework. All the parts of the architecture are designed to perform one or more operational tasks, such as data ingestion, transformation, feature extraction, or learning, so that the whole pipeline is efficient and consistent. The architecture focuses on modularity, whereby individual blocks of the system, say, pre-processing or modelling, can be modified or changed without impacting on the whole system.

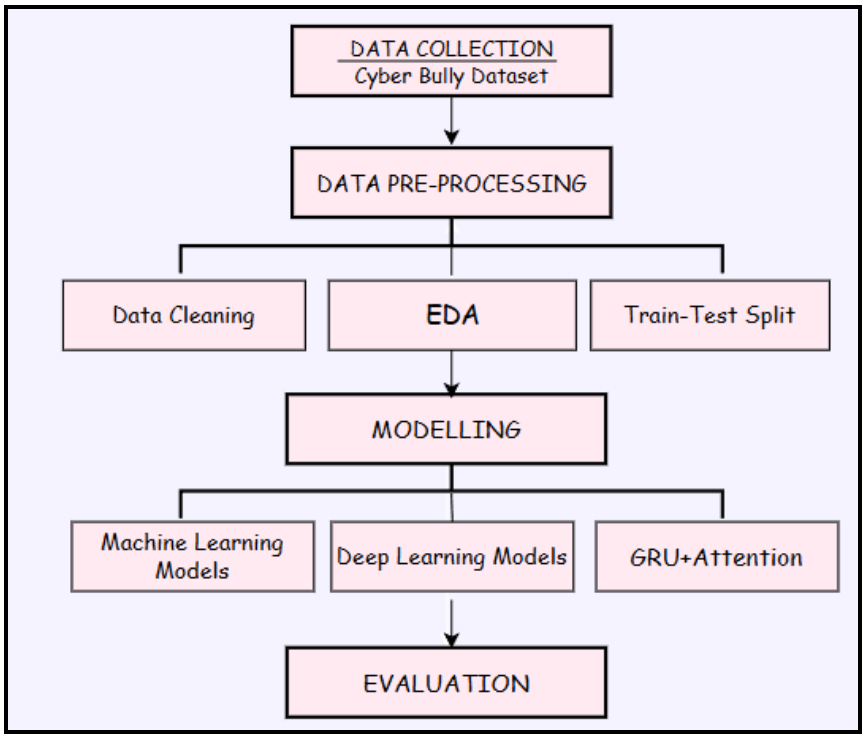

Figure 5. Overall Research Workflow Architecture

Figure 5 presents the entire workflow that is employed in the study, starting with data collection and preprocessing and continuing with the model development and final evaluation. The main focus of the architecture is to incorporate lightweight deep-learning mechanisms with the ability to perform well in low-resource settings. GRU is the primary unit of sequential processing that allows the system to learn contextual and temporal relationships in text using fewer parameters than more weighted models, such as LSTMs or transformers. The Attention layer also helps to improve the architecture by adding weights of importance to the most important words in each input sequence. Combined, these elements form a lean, scalable, and understandable system architecture that can be used to detect cyberbullying accurately and at the same time, be computationally efficient.

4.2 Workflow Pipeline Implemented

The workflow pipeline adopted in the current study is well structured and logical, starting with data collection, where the cyberbullying data is obtained on the basis of annotated posts on social media. This is then succeeded by a detailed stage of data pre-processing which involves data cleaning, exploratory data analysis (EDA) and splitting the data into training and test sets. The pre-processing operations are important in ensuring that the raw text is purged, normalized and is prepared to be modelled and EDA gives information on the label distribution, text length and linguistic patterns that affect the model design. The individual pre-processing elements aid the general quality and reliability of the training data, which is a very fundamental basis to the following modelling phases.

After preparing the dataset, the pipeline will switch to the modelling stage, during which three branches of models are trained, including traditional machine learning models, deep learning models, and the suggested GRU+Attention architecture. This parallel modelling framework makes it possible to do a complete comparative analysis of various learning approaches. Evaluation is the last step of the pipeline, during which the evaluation of all models based on standardized metrics is conducted, including accuracy, precision, recall, and F1-score. This systematic process enables the sequential steps to be logically based on the former one to create a sound and clear methodology of comprehending the performance potentials of different cyberbullying detection strategies.

4.3 Tools and Libraries Used

The implementation of this research was done in Python with a collection of popular data science libraries. Pandas and NumPy were used to easily handle and pre-process their data, and Matplotlib and Seaborn were used to create exploratory visualizations. The Scikit-learn offered TF-IDF vectorization, model training, and evaluation capabilities to develop classical machine-learning models. In the case of deep-learning, the GRU and GRU+Attention networks were implemented in TensorFlow and Keras, which provide a flexible API that supports embedding layers, recurrent units, and arbitrary attention mechanisms. All these tools combined made it possible to have a smooth process of data preparation to final model evaluation.

5 Implementation

5.1 Machine Learning Models Implementation

Hyperparameter optimization was done to select the most appropriate configuration of each classifier using the machine-learning models utilized in this study. In the case of Logistic regression, the regularization parameter C was adjusted with four candidate values namely; 0.001, 0.002, 0.005 and 0.01. The GridSearchCV applied cross-validation to each configuration and the model performed best with C = 0.01, which implies that C = 0.01 was the optimal level of regularization between generalization and stability in case of TF-IDF features. This tuning was necessary to make the linear model neither too constrained nor too responsive to noise, which would allow a strong comparison of the base lines.

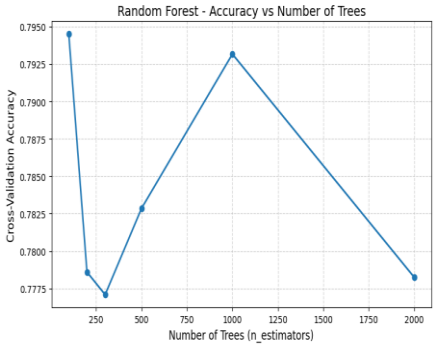

Figure 6. Random Forest – Effect of Number of Trees on Cross-Validation Accuracy

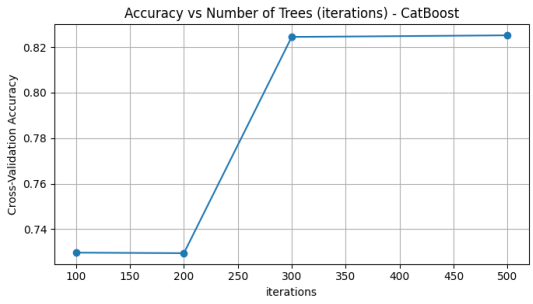

Figure 7. CatBoost Accuracy Trend

The ensemble-based models (Random Forest and CatBoost in Figure 6 and Figure 7) had gone through a similar hyperparameter search to adjust the learning behaviour. Random Forest was optimized by running six values of n_estimators, 100, 200, 300, 500, 1000 and 2000, which allowed the model to test the effectiveness of varying tree counts. The grid search established the best configuration to be 100 estimators, which is the most stable in this dataset. The CatBoost was optimized with the parameter of iterations, and the model was evaluated with the iteration of 100, 200, 300, and 500 to determine the most effective boosting depth. It was observed that the optimal configuration was achieved at 500 iterations, and this allowed the model to optimize its boosted trees without the need to subject it to unneeded computational costs. These specific tuning operations specified that every machine-learning model was run at its optimal parameterizations to be evaluated later.

5.2 Deep Learning Models Implementation

The deep-learning part of the research was done with the help of the GRU-based architecture that is capable of learning sequential patterns in the TF-IDF representations of the text. The model was built based on 128 GRU units, and this was succeeded by dense layers consisting of 32 neurons to assist the extraction of features further into the multi-class output space. GRU layers have a sequential input requirement and therefore the TF-IDFs were restructured to be 3-dimensional format to be compatible with recurrent processing. The model was trained with the Adam optimizer and a learning rate of 1e-3, which is stable and the parameter changes dynamically in training, and sparse categorical cross-entropy was used as the loss function because the multi-class label structure was considered. To avoid overtraining, an early stopping technique has been introduced to keep track of validation loss and to stop training when the loss stopped improving, resulting in an efficient and controlled model convergence.

5.3 GRU+Attention Model Implementation

The primary model was trained with an improved GRU design with a modified Attention mechanism to enhance the contextual comprehension during the classification. The network starts with an input layer which is fed with TF-IDF vectors restructured into sequential 3D format. GRU models are commonly used with sequence-based word embeddings, however, the TF-IDF vectors were converted to this format to meet the requirements of theGRU input, at the same time maintaining computational efficiency. The re-shaped vectors are then fed through a GRU layer of 256 units with return sequences being set to True because we want the complete sequence of hidden states to calculate attention. An Attention layer is then implemented, which is custom, with trainable weights and bias coefficients to calculate alignment scores to determine the most informative timesteps in the sequence, generating a weighted context vector that highlights important linguistic cues. This context representation is transmitted by dense layers of 128 and 64 neurons with ReLU activation to obtain higher-level abstractions then it goes to the final output layer which uses a SoftMax activation to achieve multi-class prediction. It is trained with the Adam optimizer with a learning rate of 1e-3 and the sparse categorical cross-entropy loss function, which were chosen due to their stability and the ability to efficiently use a multi-class text classification. Training includes a

call-back at early stopping to check the loss of validation, and recover the best-performing weights, which is an efficient and reliable convergence of the models.

6 Evaluation and Results

This part shows the findings of the experiments. The analysis is done with different measures to lead to a clear picture on the effectiveness of the model. The next table indicates the comparison of the performance of the models.

Table 2. Model Performance Comparison

Model	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.82	0.85	0.82	0.82
Random Forest	0.80	0.84	0.80	0.78
Cat Boost	0.82	0.85	0.82	0.82
GRU	0.91	0.91	0.91	0.91
GRU+Attention	**0.94**	**0.94**	**0.94**	**0.94**

Table 2, shows the performance comparison of all the models that were tested in the detection of cyberbullying. The findings present important performance indicators such as accuracy, precision, recall, and F1-score that are employed to determine the quality of classification.

6.1 Logistic Regression

The Logistic Regression model had an accuracy of 0.82, and all the precision, recall, and F1-score values were stabilized at 0.82-0.85. These findings suggest that the linear classifier was very reliable in TF-IDF features and that it coped with most of the classes but was not very contextual to capture more subtle patterns of cyberbullying.

6.2 Random Forest

Random Forest achieved an accuracy of 0.80, and its precision is 0.84 and lower F1-score of 0.78 indicating a lower ability to generalise across the minority categories of cyberbullying. Although the ensemble methodology worked on some of the non-linear relationships, the dimensionality of TF-IDF was too large to completely distinguish between subtle textual variations.

6.3 CatBoost

CatBoost reached the accuracy of 0.82 which is the same result as the Logistic Regression with an equal accuracy, recall, and F1-scores of 0.82-0.85. Its boosting approach enabled it to

identify more complex text patterns than Random Forest, but not deep semantic cues as the neural network models.

6.4 GRU Model

GRU model provided a significant gain of 0.91 in terms of accuracy, precision, recall, and F1-score. Its sequential pattern modelling capacity allowed it to gain insight into context and word sequence, which made it identify damaging language much more effectively than the standard machine-learning models.

6.5 GRU with Attention Mechanism

GRU + Attention model is an improvement of the original GRU architecture with an additional Attention layer that assists the network to concentrate on the most significant items of the input sequence. Whereas a GRU can operate on a text in its entirety and reduce it to one final state, Attention enables the model to look at all the hidden states and give greater weights to more informative features- insults, slurs, or directed abusive phrases. You can set the GRU as 256 units and return sequences=True in your implementation, such that every timestep generates a hidden representation. Attention scores based on learned weight matrices are then computed by the custom Attention layer, which uses the SoftMax on the highlighting of the meaningful patterns and generates a weighted sum of the GRU outputs. This mechanism makes sure that the model does not assume that all the words are equal but rather focuses on the particular tokens that are indicative of cyberbullying.

Due to this specific attention, the GRU + Attention model provides a significant performance improvement, with 94% accuracy, much higher than the bare GRU (91%) and by far better than classical models. The enhancement is due to the fact that Attention allows the network to accurately draw the line between minor manifestations of harassment, like between gender/sexual and ethnicity/race bullying, and avoid cases of the model confusing borderline cases of cyberbullying with not_cyberbullying. The increased precision, recall, and F1 in all classes indicate that Attention minimizes false negatives by ensuring that the model captures contextual information that the regular sequence models might fail to provide. Essentially, the integrated attention mechanism renders interpretability, increases the capacity of the model to extract important linguistic features, and greatly improves its capacity to detect cyberbullying.

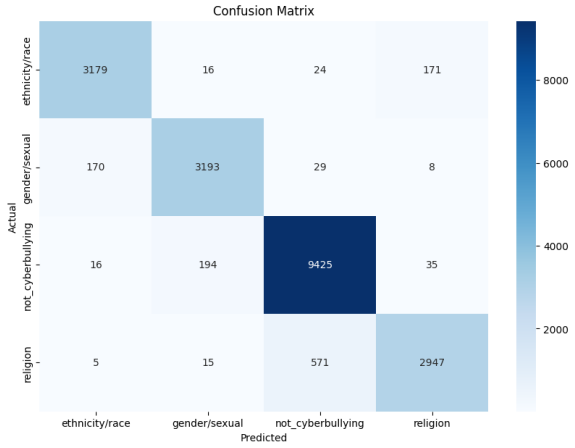

Figure 8. Confusion Matrix – GRU + Attention Model

Figure 8, which is the confusion matrix, indicates that the GRU + Attention model gives high accuracy in the prediction of all four classes. The diagonal values are high showing that the model is doing the right thing of classifying most of the samples- particularly not_cyber bullying, gender/sexual and religion. The misclassifications are low, which proves that the attention mechanism makes the model concentrate on the most significant words or phrases in each sentence and results in increased discrimination between similar instances of cyberbullying and drastically decreased false negatives.

6.6 Comparative Analysis

The comparison analysis makes it clear that deep-learning models are far more effective in detecting cyberbullying than the conventional machine-learning methods. Although the accuracy of Logistic Regression, Random Forest, and Cat Boost are moderate with 80-82 percent, the GRU model is much better when it comes to performance since it is able to capture contextual relationships in a text. GRU + Attention model has the greatest accuracy of 94% which proves that attention mechanisms contribute to the additional power of the model to focus on the important linguistic details and categorize the fine nuances of cyberbullying patterns correctly.

A comparative study of all models reveals the obvious improvement of performance between the traditional machine learning methods and deep learning structures. Traditional approaches such as Logistic Regression, Random Forest, and CatBoost had similar performance with similar accuracies of about 0.80-0.82 which indicates their success at capturing the strength of text, but also their failure to capture the depth of context. Performance was greatly enhanced through the switch to deep learning. GRU model with the sequential learning ability had a significant improvement with an accuracy of 0.91. This illustrates the usefulness of recurrent architectures in the modelling of word order and long-term dependencies in language. This was further boosted by the incorporation of the Attention mechanism that enabled the model to selectively attend to contextually important words, achieving the correct accuracy of 0.94.

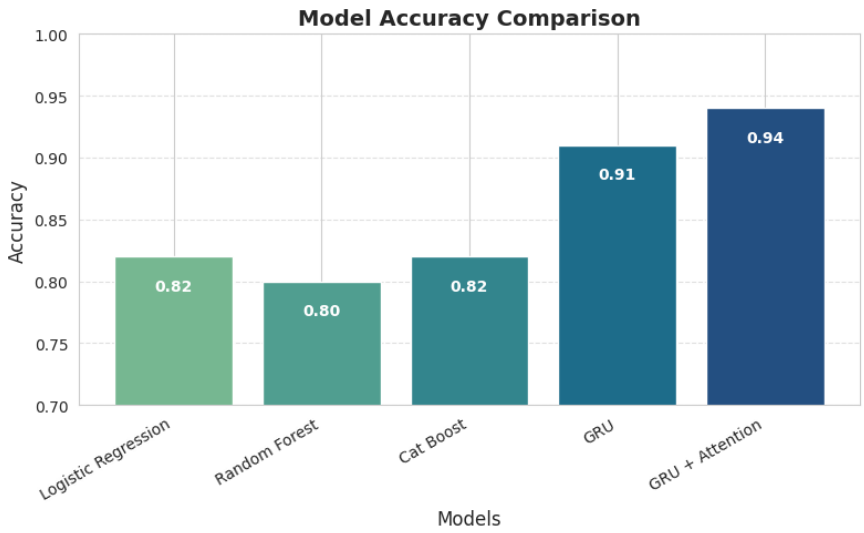

Figure 9. Model Accuracy Comparison

The relative accuracy of all the tested models in the detection of cyberbullying is indicated in figure 9 bar chart. One can observe that machine learning models accuracy is gradually increasing up to the highest accuracy of 0.94 in GRU + Attention model.

Overall, the GRU + Attention model turned out to be the most appropriate with regard to precision, recall, and interpretability. It could do more not only as a predictor, but also provide a computationally efficient alternative to more complicated transformer-based architectures. This highlights the effectiveness of the attention-enhanced recurrent models in detecting cyberbullying, particularly in a low-resource environment where there is both hardware and data constraints.

7 Conclusion

The findings of this paper indicate that the Attention mechanism, combined with a Gated Recurrent Unit (GRU) network, can significantly enhance the precision and effectiveness of cyberbullying detection under the low-resource circumstances. The paper has demonstrated that the traditional machine learning models (though applicable in simple classification) have a limitation in their ability to capture contextual sense in online text. GRU architecture which is a representation of sequential dependencies also could considerably enhance the performance of classification as it has the effect of capturing linguistic patterns and temporal dependencies. The Attention mechanism further improved the capacity of the model as the model could selectively attend to contextually appropriate words and phrases that resulted in improved perception of the subtle and complex bullying cues such as sarcasm, coded words and indirect aggression. All these additions were directed at the trade-off between performance and computational efficiency that was supposed to fulfil the primary goal of the study.

The study offers insights into the field of intelligent content moderation by showing that a minimalist GRU-Attention system can provide state-of-the-art performance without the computational complexity of transformer-based systems. This confirms that deep contextual knowledge does not always need the use of large scale or intense resource models. The performance of the model justifies its use in real-time or mobile deployment where the use of heavy architectures can be constrained by hardware limitations. In general, the research is a valuable contribution to the field as it offers a scalable and practical solution to automated detection of cyberbullying to fill the gap between high-accuracy language models and resource-efficient system design.

7.1 Future Work

The future studies can investigate the possibility of expanding GRU-Attention frame by adding the more sophisticated contextual embeddings like BERT or word-level transformers to improve the semantic meaning without adding much computational cost. It would work better to generalize to different online platforms by expanding the dataset to include multilingual and multimodal content (including images or audio). Also, the introduction of real-time implementation on mobile or edge devices would be an informative test on the model regarding its performance in the real-life, low-resource settings. Such innovations would enhance the flexibility of the framework and its usefulness in fighting cyberbullying in digital environments throughout the world.

REFERENCES

1. Alabdulwahab, A., Haq, M.A. and Alshehri, M., 2023. *Cyberbullying detection using machine learning and deep learning.* International Journal of Advanced Computer Science and Applications, 14(10).
2. Alam, K.S., Bhowmik, S. and Prosun, P.R.K., 2021, February. *Cyberbullying detection: an ensemble-based machine learning approach.* In 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV) (pp. 710–715). IEEE.
3. Al-Hashedi, M., Soon, L.K., Goh, H.N., Lim, A.H.L. and Siew, E.G., 2023. *Cyberbullying detection based on emotion.* IEEE Access, 11, pp.53907–53918.
4. Aliyeva, Ç.O. and Yağanoğlu, M., 2025. *Deep learning approach to detect cyberbullying on Twitter.* Multimedia Tools and Applications, 84(19), pp.20497–20520.
5. Aldhyani, T.H., Al-Adhaileh, M.H. and Alsubari, S.N., 2022. *Cyberbullying identification system based deep learning algorithms.* Electronics, 11(20), p.3273.
6. Almufareh, M.F., Jhanjhi, N., Humayun, M., Alwakid, G.N., Javed, D. and Almuayqil, S.N., 2025. *Integrating sentiment analysis with machine learning for cyberbullying detection on social media.* IEEE Access.
7. Alotaibi, M., Alotaibi, B. and Razaque, A., 2021. *A multichannel deep learning framework for cyberbullying detection on social media.* Electronics, 10(21), p.2664.
8. Bharti, S., Yadav, A.K., Kumar, M. and Yadav, D., 2022. *Cyberbullying detection from tweets using deep learning.* Kybernetes, 51(9), pp.2695–2711.

9. Haider, A., Siddique, A.B., Ali, R.H., Imad, M., Ijaz, A.Z., Arshad, U., Ali, N., Saleem, M. and Shahzadi, N., 2023, October. *Detecting cyberbullying using machine learning approaches.* In 2023 International Conference on IT and Industrial Technologies (ICIT) (pp. 1–6). IEEE.
10. Hoque, M.N. and Seddiqui, M.H., 2024. *Detecting cyberbullying text using the approaches with machine learning models for the low-resource Bengali language.* Int J Artif Intell ISSN, 2252(8938), pp.358–367.
11. Jain, V., Kumar, V., Pal, V. and Vishwakarma, D.K., 2021, April. *Detection of cyberbullying on social media using machine learning.* In 2021 5th International Conference on Computing Methodologies and Communication (ICCMC) (pp. 1091–1096). IEEE.
12. Perera, A. and Fernando, P., 2024. *Cyberbullying detection system on social media using supervised machine learning.* Procedia Computer Science, 239, pp.506–516.
13. Raj, C., Agarwal, A., Bharathy, G., Narayan, B. and Prasad, M., 2021. *Cyberbullying detection: Hybrid models based on machine learning and natural language processing techniques.* Electronics, 10(22), p.2810.
14. Raj, M., Singh, S., Solanki, K. and Selvanambi, R., 2022. *An application to detect cyberbullying using machine learning and deep learning techniques.* SN Computer Science, 3(5), p.401.
15. Saifullah, K., Khan, M.I., Jamal, S. and Sarker, I.H., 2024. *Cyberbullying text identification: A deep learning and transformer-based language modeling approach.*
16. Teng, T.H. and Varathan, K.D., 2023. *Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches.* IEEE Access, 11, pp.55533–55560.