Egyptian Informatics Journal 32 (2025) 100856

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Full Length Article

Detection cyberbullying using AI and sentiment analysis to examine psychological impacts on vulnerable groups

Abdulnaser M. Fashakh [a,*], Mesut Çevik [a], Şenay Kocakoyun Aydoğan [b], Abdullahi Abdu Ibrahim [a]

[a] Department of Electrical and Computer Engineering, Altinbas University, Istanbul, Turkey
[b] Department of Information Security Technology, T.C. İstanbul Gedik University, Istanbul, Turkey

ARTICLE INFO

Keywords:
Cyberbullying
Machine learning
Text classification
Sentiment analysis
Deep learning
Psychological impact
Vulnerable groups

ABSTRACT

This study aims to assess the effectiveness of machine learning and deep learning models in detecting cyberbullying and evaluating its psychological impact on vulnerable groups using textual and emotional features. The models assessed include traditional classifiers—Logistic Regression, Decision Tree, and Random Forest and deep learning models, such as MLP, CNN, RNN, and (LSTM) networks. TF-IDF for text vectorization and TextBlob for sentiment analysis were utilized. In spite of TF-IDF's shortcoming. Its simplicity enabled quick prototyping and insight results. The dataset contained 58,000 tweets, with 46,000 obtained from Kaggle and 12,000 collected via the Twitter API. Tweets were labeled into cyberbullying_type (gender, age, religion, and ethnicity) and subcategories: gender (male, female, LGBT, other), age (adult, teenager, other), religion (Muslim, Christian, Jewish, other), and ethnicity (ethical, unethical, other). Keyword-based classification was used for Subcategory assignment. The emotional score derived from text served as a proxy for measuring psychological impact. We emphasize that this study is observational and does not rely on clinical psychological evaluation. Results showed that female and LGBT users experienced the highest levels of cyberbullying among gender subcategories. Teenagers were most affected by age-based bullying. Unethical content dominated ethnicity-based attacks, and Muslims faced the highest frequency of cyberbullying and negative sentiment in religious categories. Sentiment analysis assisted in identifying emotional patterns concerning online abuse. Among models RNN and LSTM models achieved the highest accuracy (0.98), outperforming others. Among the traditional models, Random Forest performed better, while Logistic Regression was the worst performing. The inclusion of sentiment features significantly improved calssification accuracy, particularly in LSTM. A multi-output LSTM model was created to predict cyberbullying_type, sub_category and sentiment all at once, providing an end-to-end detection system. This framwork enables proactive monitoring of online harm and support timely interventions.

1. Introduction

Incorporating ethical considerations into the development and implementation of tools to prevent and mitigate cyberbullying is important. Furthermore, Ashktorab explored the potential of AI and sentiment analysis to identify patterns and indicators of cyberbullying behavior, offering insights into the design of effective intervention strategies. This study also sheds light on the psychological impact of cyberbullying on vulnerable groups, emphasizing the need for comprehensive support systems that address the emotional well-being of victims. By delving into the nuances of cyberbullying and its consequences, Ashktorab's work contributes to the ongoing discourse on this pressing issue and calls for collective efforts to combat cyberbullying through responsible technological innovations and holistic support mechanisms. This study invites further research and collaboration to develop practical and ethical solutions that promote safer online environments for all individuals. Longitudinal studies are needed to evaluate the behavioral impact of various preventative solutions on bullies, victims, and bystanders. Notably, Ashktorab categorizes various types of preventative measures and claims that existing automated methods for detecting cyberbullying have largely ignored context-specific forms, such as sexism and racism [1]. Current studies emphasize the significance of AI-

* Corresponding author.
E-mail addresses: 213720046@ogr.altinbas.edu.tr (A.M. Fashakh), mesut.cevik@altinbas.edu.tr (M. Çevik), senay.aydogan@gedik.edu.tr (Ş.K. Aydoğan), abdullahi.ibrahim@altinbas.edu.tr (A.A. Ibrahim).

https://doi.org/10.1016/j.eij.2025.100856
Received 21 December 2024; Received in revised form 23 July 2025; Accepted 28 November 2025
Available online 10 December 2025

powered tools in mitigating psychological effects of cyberbullying, particularly among teens and college students [2]. This article provides a systematic review entitled "Effects and prevention of cyberbullying and social exclusion in social media." The analysis revealed important trends in cyberbullying and exposure, especially among adolescents and university students. The authors emphasize the serious psychological consequences associated with cyberbullying, including mental disorders and suicidal thoughts, and see strong social support systems as one of the best preventive means. However, they pointed to challenges in determining causal relationships due to the reluctance of individuals involved in cyberaggression to participate in studies. The article also emphasizes the need for longitudinal research to understand the complex dynamics of cyberbullying in different social contexts [3]. In an article entitled "Studying the role of artificial intelligence technology in online mental health care: opportunities, challenges and impacts, a review in mixed ways," researchers reviewed the promising applications of artificial intelligence in addressing mental health problems, including those resulting from cyberbullying. They discussed how artificial intelligence can contribute to improving online mental health care through diverse applications, such as suicide risk assessment and mental state follow-up. This review summarizes evidence of AI's ability to enhance access to and disaggregation of mental health services, It is necessary to note that the emotional analysis provided herein does not act as a substitute for clinical psychological diagnosis. Such acknowledgment is crucial for ethically outlining the boundaries of AI-based assessments utilized in this study [4]. AI technologies may be critical in supporting people affected by cyberbullying. Together, these articles provide an overview of the status of e-bullying prevention and intervention strategies, focusing on the importance of ethics and social support systems and integrating artificial intelligence into mental health care. There is also an urgent need for targeted research and innovative solutions to reduce the psychological impact of cyberbullying on vulnerable populations [5]. Cyberbullying is defined as aggressive and deliberate behavior frequently practiced via electronic media against a victim who cannot easily defend herself. In contrast, online abuse refers to a variety of behaviors that may be considered abusive by target parties [6]. Evolutions have been attempting to use the internet and communication technology as tools more and more in the past several years (ICT). Many apps, social media, and technologies have arisen to support communication with others online. Currently, the Internet is an important necessity for the global community. All people have access to all information without any restrictions and conditions in time and place [7]. Social media is a forum that allows people to simply post photos, texts, documents, and videos, and interact with community members without restrictions [8]. It also offers many new opportunities in the labor market, thus aiding the growth of the global economy. The genesis and prevalence of social media have led to significant changes in how people communicate with them, so its effects have been evident in different aspects of their lives. However, the phenomenon of so-called cyberbullying, where people use these online platforms and social means to emotionally harm other people by snooping, disturbing them by violating privacy, and stirring up hatred, is worrying. Cyberbullying is a major concern in social media. It refers to the use of electronic means of harassment or bullying and expletive speech and is also known as cyberbullying. As the digital sphere increases and technological advances develop, cyberbullying has become popular, especially among adolescents. Approximately 50 % of adolescents in the United States suffer from online bullying, which generates physical and mental effects on victims. The great shock sometimes leads online bullying of victims to serious acts such as suicide Yang, B. et al [9]. In this research, due to increased online bullying methods including online trolling, cyberbullying, sexual exploitation such as grooming, sexting, and revenge pornography (i.e., posting indecent images of former partners), repeating and spreading through social media is a prominent feature of cyberbullying within a wider range of online abuse, sparking considerable interest in automatically detecting it while analyzing feelings toward vulnerable groups targeted by cyberbullying. This has led to increased research using supervised machine learning and deep learning techniques to achieve this goal. Training data play a crucial role in identifying cyberbullying, online abuse, and the psychological consequences of such violations. Therefore, the frequent development of automated cyberbullying detection systems is critical for the early intervention and support of these vulnerable groups.

1.1. Problem statement

Despite the growing body of research, there remains a lack of cyberbullying detection systems that are context-sensitive, ethically engaged, and emotionally intelligent, which can be deployed in real-time environments and take vulnerable groups into account.

2. Literature review

The literature has extensively addressed cyberbullying through the lens of social and computer sciences. Studies have provided various definitions of cyberbullying and computational methods for online identification [10]. Mankind has conducted sufficient work to develop research in the cyberbullying domain, which aims to understand, detect, and combat online behavior. Natural language processing (NLP) techniques have been harnessed to identify and address cyberbullying actions [11]. NLP is employed to detect offensive language in social media content, targeting the prevention of cyberbullying to improve society. This strategy focuses on spotting negative and offensive language, which is crucial for recognizing harmful digital interactions [12]. Has contributed significantly to the proposal of a robust technique for detecting cyberbullying on social media platforms. Their approach utilized sequential hypothesis testing to enhance accuracy while reducing the number of features required for classification. This study underscores the precision, timeliness, and scalability of devising effective cyberbullying detection methods [13]. Adopted comprehensive approach was adopted by incorporating social network analysis and text analysis to identify cyberbullying. They recognized the role of relationships in disseminating and perpetuating cyberbullying content. This underscores the necessity to consider diverse dimensions when addressing online harassment. This study is an additional valuable contribution to literature [14]. Employed word-embedding models and random forest classifiers to identify cyberbullying comments on platforms. This combination of domain-specific knowledge with machine learning techniques demonstrates the potential for improved performance in cyberbullying detection [15], followed by a proactive path to notify parents about detected cyberbullying occasions. This interdisciplinary attempt showcases collaborative efforts to safeguard young individuals from online harm and highlights the importance of engaging multiple stakeholders. The literature shows the ongoing evolution of research in cyberbullying detection, with scholars combining NLP, machine learning algorithms, and social network analysis. These attempts underscore the need for a grand approach that considers linguistic cues and contextual factors to address cyberbullying effectively. In related research, diverse methodologies have been employed to detect cyberbullying. These approaches frequently utilize NLP, information retrieval, and feature extraction techniques. For example [16], recognized social network structure as an influencing factor in cyberbullying detection, advocating for the integration of text and social network analysis. So used NLP to identify offensive language in social media [17]. They also worked on the same principles and linked cyberbullying to multiple characteristics such as profanity, violence, temporary feedback patterns, social networks, language, and visual content. This requires establishment of precise standards for cyberbullying. Conventional bullying, which includes power imbalance, intent, aggression, and repetition, can have its basic characteristics expanded and used as a basis for the definition of cyberbullying [18]. Media use is a part of aggressive behavior. Threats, insults, ridicules, spreading

rumors, exclusion, and posting embarrassing images on the Internet refer to cyberaggression and are essential elements of cyberbullying. Due to the nature of digital media, the aggressor has continuous access, meaning that the tools are always available to him [19]. All this highlights the importance of an effective and comprehensive solution for this common problem. The phenomenon of cyberbullying must be understood and addressed from multiple perspectives. Automatic detection and prevention of such incidents can significantly contribute to solving this problem. Tools capable of reporting incidents of bullying have already been developed [20]. In addition, most online platforms frequented by adolescents have security centers, such as the YouTube Security Center and Twitter Security Center, which provide support to users and follow communications. Several studies have been conducted on the automatic detection of cyberbullying and its prevention methods, which are discussed in greater detail in the next section. However, this problem is still far from a final solution, and improving the tools available is necessary to reach effective solutions [21]. Cyberbullying has had a huge impact on the world; however, many do not have a clear understanding of how it is socially encountered. Cyberbullying is a serious issue that must be addressed before its effects on society worsen, as it can hinder the educational process and cause serious psychological and physical problems for victims. Cyberbullying has increased in popularity over Twitter in recent years, linked to tragic and alarming suicides [22]. Cyberbullying detection is often conducted using filters or machine learning techniques. It is necessary to identify hack techniques, profanity and terminology in texts to detect cyberbullying [23]. Cyberbullying is defined as a form of online harassment, particularly on social media platforms. Criminals exploit these networks to gather the information needed to commit crimes, such as choosing victims who show vulnerability [24]. In literature, sentiment analysis is a supporting task used to improve the performance of the core task [25]. Emotions are the thoughts and opinions that develop from emotions linked to a specific thing, and they typically fall into one of three categories: positive, neutral, or negative [26]. Unlike text analysis, sentiment analysis detects different sentiments through text expressions such as anger, disgust, fear, happiness, sadness, and surprise. There are three common ways to detect feelings in texts: the keyword-based method, which uses synonyms and antagonists of dictionaries, the learning-based method, which relies on pre-trained classifications, and the hybrid method, a combination of the previous two methods [27]. Recent research has begun to adopt deep learning approaches, such as BERT and transformer-based models, to significantly advance the contextual understanding in cyberbullying detection tasks [28]. These newer models exhibit better performance than conventional TF-IDF-based methods in identifying complex linguistic features like sarcasm, metaphors, and identity-based attacks. Furthermore, a number of studies have combined these advanced models with sentiment analysis to advance the psychological profiling of cyberbullying content and inform interventions accordingly [29]. Despite these advances, the literature often reflects a lack of datasets that are either context-specific or representative of marginalized victim groups, e.g., LGBTQ+ or ethnic minorities [30]. Additionally, a vast majority of research ignores the longitudinal evaluation of emotional impact on the victims, which limits the understanding of extended psychological harm [31]. This highlights a clear research gap: a lack of comprehensive, multi-output models capable of simultaneously detecting cyberbullying type, subgroup identity, and emotional state, using explainable AI techniques. Addressing this gap is essential to develop robust, ethical, and generalizable detection systems. This highlights a clear research gap: the lack of comprehensive, multi-output models capable of simultaneously determining the cyberbullying type, subgroup identity, and emotional state via the use of explainable AI techniques. Addressing this gap is imperative for the development of robust, ethical, and generalizable detection systems. This study benefits from supervised and unsupervised machine learning models as well as deep models, advanced feature engineering techniques, text groups, and social media platforms to automatically identify complex cyberbullying patterns. Our contribution is to integrate sophisticated natural language processing techniques with sentiment analysis to improve the current methods of detecting cyberbullying with higher accuracy and greater context understanding, enabling a deeper understanding of the psychological impact on victims.

3. Objectives of study

Introduction: Cyberbullying has emerged as a pivotal issue in the digital age, especially with the increasing use of social media and its extreme impact on people. Most vulnerable groups, such as children and adolescents, and marginalized communities, such as the LGBTQ + community, are most at risk. The psychosocial effects of cyberbullying may be acute, necessitating the development of effective automated systems for the early detection and support of victims. The objectives of this study are directly related to closing this gap through enhancing detection and understanding of the psychological impacts by using artificial intelligence and sentiment analysis.

3.1. Objective 1: detection

3.1.1. Introduction

Detecting cyberbullying is a challenge given its multiple forms and methods, and the varied contexts in which it may occur. Using AI-enhanced machine learning models can contribute to early and accurate detection by analyzing texts and understanding their emotional context. Its utilization of TF-IDF provides for good textual feature representation, and TextBlob sentiment scoring introduces emotional nuance to enhance model sensitivity and contextual comprehension.

3.1.2. Study objective

This study aimed to compare the performance of various machine learning models in detecting cyberbullying using text and emotional features. The evaluated models include traditional models such as logistical regression, decision tree, and random forest, as well as advanced models such as synthetic neural networks (CNN), repetitive neural networks (RNN), and long- and short-range memory (LSTM). Texts are analyzed using advanced techniques such as TF-IDF to convert texts into digital representations and TextBlob to extract emotions scores. The choice of these models provides the ability to compare performance on both traditional and deep learning approaches, with an emphasis on emotional context to enhance detection accuracy.

3.1.3. Importance of the objective

Improving the accuracy and efficiency of the models used to detect early cyberbullying can help provide timely and effective support to victims, thereby providing preventive and remedial measures. Analysis of the feelings associated with texts allows for a deeper understanding of the emotional context, helping to identify cyberbullying in a more comprehensive and accurate way. This contributes to practical implementation in digital platforms and facilitates early warnings for harmful interactions.

3.2. Objective 2: Impact analysis

3.2.1. Introduction

The effects of cyberbullying on psychosocial well-being, including anxiety, depression, and social isolation, can be severe and long-lasting. Vulnerable groups, such as adolescents, women, and marginalized groups, are particularly affected, making it necessary to conduct a comprehensive analysis and measurement of these impacts. The integration of sentiment analysis allows psychological effects to be measured using emotional metrics derived from user-generated content.

3.2.2. Study objective

This objective was to study and measure the psychological impact of

cyberbullying on vulnerable groups. This is achieved through analysis of text feelings to detect levels of negative emotions associated with cyberbullying victims using techniques such as TextBlob. These analyses were used to identify the relationship between the nature of the texts and their resulting psychological effects, thereby enabling the development of effective intervention and psychological support strategies. The focus on vulnerable groups allows for targeted insights that inform mental health response systems.

3.2.3. Goal importance

This analysis provides a deep understanding of the psychological effects of cyberbullying, helping develop effective strategies for intervention and psychological support for victims. These results contribute to raising awareness among society and policymakers about the seriousness of cyberbullying and the need for preventive and curative measures, enhancing the protection of vulnerable groups, and contributing to a safer digital environment. It also supports the development of ethical, AI-based solutions aligned with mental health objectives.

4. Methodology

4.1. Data collection

4.1.1. Dataset

Our dataset is a vital component of research and is used as a basis for training and evaluating models in machine learning to detect cyberbullying. This collection consists of cyberbullying scripts collected from various sources, including Kaggle, and has been expanded using Tweepy, a Python programming tool that gives access to the Twitter API. This resulted in a rich and diverse dataset of 58,000 tweets, which after processing, cleaning and iterating became 51,202 tweets that can provide a solid foundation for analysis and study. This joint strategy provided coverage across a range of cyberbullying types and language variants, promoting model generalizability.

4.1.2. Sources and expansion

Kaggle: Kaggle is a well-known platform for competitions and datasets in data science and provides a variety of datasets, including those related to social communication and cyberbullying. The first batch of data, 46,000, was obtained from Kaggle, which provided many classified texts for cyberbullying.

Tweepy: Tweepy was used to collect additional tweets of more than 12,000 to strengthen the dataset and ensure its comprehensiveness and updates. Tweepy allows tweets to be extracted instantly or from specific periods of time using various search criteria to collect relevant texts, ensuring that the dataset covers various cyberbullying and linguistic diversity scenarios. The dynamic nature of tweet extraction through Tweepy helps maintain temporal relevance in the dataset, supporting more accurate real-world application of the model.

4.1.3. Data preprocessing

Preprocessing a dataset involves several critical steps to prepare text data for analysis and model training.

Text Cleaning: Text cleaning involves converting all text to lowercase to ensure uniformity and removing special characters, numbers, and punctuation marks that are irrelevant to the analysis. This step helps reduce the noise in the data.

Tokenization: The cleaned text is then encoded by dividing it into individual words or symbols. Encoding helps to further process and analyze texts at a fine-grained level.

Removing Stop Words: Common words that do not add meaning to the text, known as "," "the" and "is," were excluded. This helps focus on keywords that play a role in identifying cyberbullying.

TF-IDF Vectorization: Text has been digitized using the term frequency scale–reverse document frequency (TF-IDF). TF-IDF assigns weight to each word based on its frequency in the document and scarcity across text, highlighting keywords and lowering the importance of common words. TF-IDF was chosen over deep contextual embeddings like BERT because it was less computationally expensive, more easily interpretable, and had a strong baseline performance in short-text classification problems like tweets. It also works well with classic machine learning models and does not overfit on smaller datasets.

Stemming: A lingual normalization technique involves breaking down words into their root forms. NLTK library stem. For this purpose, the Porter method was used. Porter. Stemmer to stem tokens and obtain stemmed tokens. For example, the words "connection," "connected," and "linking" can be reduced to the word "connect." connect.

Digit removal: Because numerical information does not promote cyberbullying, any material was filtered out.

Remove punctuation: To remove punctuation, only text that is not punctuation is saved, which can be verified using string punctuation.

Sentiment Analysis: Emotion scores were drawn using the TextBlob library in Python, which processes text data. TextBlob offers polarity scores for texts, pointing to their feelings whether positive, negative, or neutral. These grades, in addition to the features of TF-IDF, are combined to form the final set of features used in the model training. The sentiment features enabled the model to differentiate emotionally charged messages—particularly those with negative polarity—enabling more in-depth psychological profiling of abusive content.

4.1.4. Dataset characteristics

Size: The dataset consisted of more than 58,000 tweets, providing a large and diverse sample of text for analysis.

Diversity: The tweets cover a wide range of topics and scenarios related to cyberbullying, ensuring that the models are trained in diverse data and can generalize well in different contexts.

Labels: Each tweet was labeled as cyberbullying or not, allowing for supervised learning and evaluation of model performance.

In the Figure below, we explain the mechanism of retrieving tweets using the Tweepy Library: see Fig. 1

Here is a detailed explanation of the dataset: This multi-category dataset is categorized according to the category of cyberbullying as follows: Table 1 and Fig. 2.

4.1.5. Importance of the Dataset

Comprehensiveness: Combining data from Kaggle and tweets collected via Tweepy, the dataset is comprehensive and, to date, reflects the latest trends and variations in cyberbullying.

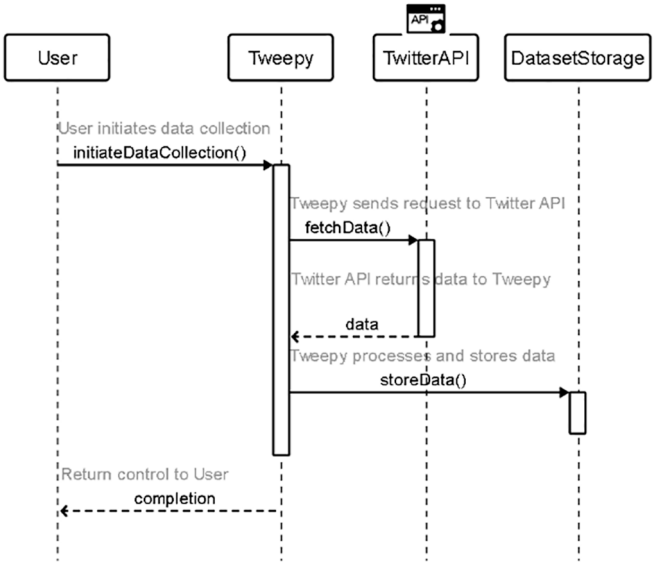

Fig. 1. This is the figure caption. Method of collecting tweets.

Table 1
Labels of dataset and distribution of tweets.

Cyberbullying Type	No of Tweets
Age	9031
Gender	11,956
Religion	7968
Ethnicity	7931
Other cyberbullying	6993
Not cyberbullying	7323
Total of tweets	51,202

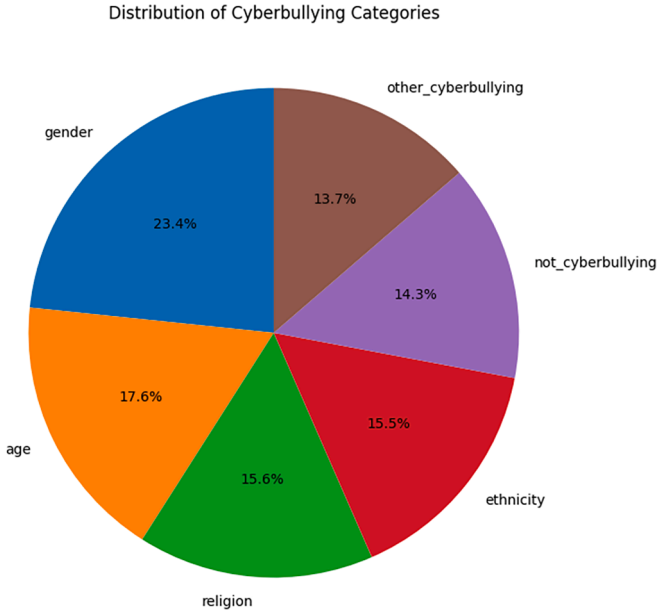

Fig. 2. Distribution of CyberbullyingType.

Durability: Preprocessing steps ensure hygiene, consistency, and readiness of data for analysis, enhancing the strength and confidence of the models used for such data.

Enriching emotions: Adding emotion analysis provides an additional contextual layer, helping to understand the emotional tone of texts, which is necessary to identify cyberbullying accurately. This also enabled the exploration of psychological effects through the analysis of the sentiment polarity and the categories of cyberbullying. After completing the process of data processing, cleaning, removing duplicates and encoding us, and then we had (51,202) tweets ready for analysis and training, we decided to merge the two categories (not_cyberbullying and other_cyberbullying) because they are not important to the subject of our research and redistribute them to the rest of the important categories and with the completion of the process of balancing data categories, the total number of tweets became (71,704) by (17,926) for each category, as shown in the following distribution in the table below: Table 2.

Table 2
Dataset after balanced tweets.

Cyberbullying Type	No of Tweets
Age	17,926
Gender	17,926
Religion	17,926
Ethnicity	17,926
Total of tweets	71,704

4.2. Model training

This study builds on an integrated methodology that combines diverse machine learning models, including traditional methods such as logistic regression and decision trees, along with newer technologies such as conventional neural networks (CNNs) and long- and short-term memory networks (LSTM). These methodologies were selected for their proven effectiveness in processing complex datasets and for their ability to monitor accurate patterns in texts [32], which is a multifaceted challenge in the digital age, requiring robust methodologies to effectively detect and mitigate its effects, especially in vulnerable populations such as adolescents, children, and marginalized communities [33]. Each model was chosen based on specific strengths: Random Forest and Decision Tree for interpretability, SVM for margin optimization in high-dimensional spaces, and deep learning models for capturing sequential and contextual patterns. The emergence of advanced machine learning and artificial intelligence technologies provides a promising path for developing automated systems that can identify cyberbullying cases with high accuracy and understand their psychological implications [34]. To ensure scientific validity, all models were trained and validated under consistent conditions, using the same training/test splits and feature vectors. This controlled setup helps minimize bias when comparing models. The dataset used in this study consisted of Kaggle's compilation of cyberbullying texts reinforced by Tweepy, resulting in a powerful collection of 58,000 tweets. Advanced data processing steps, including text cleaning, are developed into codes, and features are extracted using Term Frequency-Inverse Document Frequency (TF-IDF) and sentiment analysis to ensure that data are adapted appropriately for model training and assessment. The TF-IDF vectors capture term relevance, while sentiment polarity (from TextBlob) adds emotional context, creating a hybrid feature space that improves model sensitivity to harmful content. Advanced model training includes an accurate process for modifying and validating superior parameters to enhance algorithm performance. Hyperparameter tuning was performed using grid search or manual adjustment depending on the model type, optimizing variables like tree depth (Random Forest), C value (SVM), and number of layers or neurons (deep networks). Evaluation criteria such as accuracy, remembrance, F1 score, and health were used to provide a comprehensive assessment of the model's efficiency [35]. The full evaluation was conducted using a hold-out test set (20 % of the data) to measure real-world generalization. This study, combined to analyze sentiment with traditional text characteristics, seeks to improve model detection capabilities and provides a deeper understanding of the emotional impact of cyberbullying, contributing to more efficient intervention strategies.

5. Machine learning models

5.1. Random Forest

The random forest classifier is a powerful machine-learning algorithm utilized for multiclass classification tasks [36]. This harnesses the collective intelligence of numerous decision trees. During training, it constructs a diverse ensemble of decision trees, each trained on a random subset of data using random subsets of features. This diversity reduces overfitting by preventing any single tree from dominating the model-decision-making process. To make a prediction, a new instance is passed through all the trees, and each tree's prediction is counted. The final prediction is resolved by the majority vote or the average prediction of all individual trees, which enhances the generalization. because the combined predictions are less susceptible to noise and errors than any single decision tree. Random Forests is robust against outliers, can manage high-dimensional data, and provides an estimate of feature importance. However, they might be challenging to interpret compared with single-decision trees. A classifier was created and used in the sklearn-ensemble package.

This model was selected for its ability to handle large-scale, imbalanced datasets and provide robust classification in noisy environments [37].

Working:

Random Forest combines the predictions of multiple decision trees to enhance the accuracy and reliability of classification. The probability of a sample belonging to a particular class, denoted as 'c,' is determined through a majority voting scheme across all the forest decision trees.

Mathematically, the probability pc is calculated as

$$Pc = \frac{1}{N}\sum_{i=1}^{n} I\,(c - ci)$$

where:

- *Pc* Represent the probability of class c.
- N represents the number of decision trees in the forest.
- *ci* Represent the class predicted by the *i* the decision-tree.
- I (0) represent the indicator function that returns 1 if the condition is true, and 0 otherwise.

5.2. Decision tree

A decision tree was created by asking a series of questions about the dataset. A new question was asked after each response, until the class label of the record was determined. A decision tree is a hierarchical structure that includes nodes and directed edges, and is a useful tool for organizing a set of questions and their responses Pavlopoulos et al [38]. This tree has three main parts: the starting point (root), middle parts with questions (internal nodes), and endpoints with answers (leaf nodes). A decision tree is a predictive model that maps features to a target value. It consists of nodes that represent the attributes of the data, and branches that represent the decision rules. Starting at the root node, the data are split based on the feature that provides the most information. This process was repeated at each child node until the leaves contained only one sample class. The classifier was implemented using the sci-kit-learn library decision-tree package.

This model was chosen for its simplicity, interpretability, and efficiency in handling structured data [39].

Working:

Decision trees make predictions by dividing the feature space into smaller segments based on the feature values. The algorithm selects the features at each node of the tree by maximizing the information gain or minimizing the impurity measure, such as the Gini impurity. The splitting criterion was determined by comparing the feature values with a threshold. To predict a sample, a tree is traversed from the root to a leaf node, where the leaf node corresponds to the predicted class label. Mathematically the decision tree's prediction Y for a sample x can be represented as:

For Classification: If the tree performs classification, the prediction is the majority class in the leaf node where x lands:

$$Y(x) = \arg max_c\left(\frac{n_c}{n_{leaf}}\right)$$

where:

- c is the class label.
- n_c is the number of samples of class c in the leaf node.
- n_{leaf} where is the total number of samples in leaf node.

5.3. Support Vector Machine

Support-Vector-Machine (SVM) The algorithm is versatile and can be applied to multiclass problems Campbell et al. [40]. The Support Vector Machine (SVM) was originally designed for binary classification, but it can also manage multiple classes using techniques such as one-vs-one or one-vs-all. In one-vs.-one, SVM builds a separate binary classifier for each pair of classes and combines their decisions. In one-vs. all- the SVM trains a separate binary classifier for each class against the rest. The final classification is determined by combining the outputs of the binary classifiers. SVM's ability of an SVM to manage high-dimensional spaces and effectively find optimal hyperplanes makes it a powerful choice for multiclass classification tasks, providing accurate and reliable results across diverse datasets. The SVM Classifier is implemented using the Sklearn. svm package.

SVM was chosen due to its strong generalization performance, especially in high-dimensional feature spaces created by TF-IDF [41].

Working:

In a binary classification setting, the SVM seeks to find the optimal hyperplane that separates the two classes in the feature space. Mathematically, this hyperplane is defined by the following equation. For binary classification, the SVM assigns a label $y \in \{-1, +1\}$ to a sample x based on which side of the hyperplane falls on:

$$y^{\wedge} = sign(w^T x + b)$$

where:

- $y^{\wedge}$ where is the predicted class label (+1 or −1).
- $w^T x + b > 0$ assigns sample to class + 1 (positive class).
- $w^T x + b < 0$ assigns the sample to Class 1 (negative class).
- The sign function returns +1 if the expression is positive and −1 if it is negative.

5.4. Logistic regression

Regression analysis is a method used for predictive modelling to examine the relationship between an independent variable and a target variable in a dataset. This method is applied when the target and independent variables show a linear or non-linear connection, and the target variable has continuous values. Regression analysis aims to find the best-fit line that minimizes the gap between the line and each data point. Logistic regression is a type of regression analysis used when the dependent variable is discrete, such as having values of 0 or 1, true or false, etc. (also used for multiclass classification) [42]. This means the target variable can only have two values and the relationship between the target variable and the independent variable is represented by a sigmoid curve, which maps any real value to a value between 0 and 1. In our case, we chose Logistic Regression because our dataset was large and the occurrence of values in the target variables was almost equal. Additionally, there was no correlation between the independent variables in the dataset. We implemented the classifier using the sklearn-linear_model package.

This model is suitable for large-scale text classification tasks due to its speed and effectiveness in linearly separable problems [43].

Working:

A Logistic Regression uses the logistic function (sigmoid function) to model the probability that a sample belongs to a specific class. The probability p is calculated as follows:

$$p = \frac{1}{1 + e^{-z}}$$

where:

z is the linear combination of the feature values weighed by the model coefficients:

$$z = \beta 0 + \beta 1\, x1 + \beta 2\, x2 + \cdots + \beta n\, xn$$

Overall Comparison

- Random Forest provides high accuracy and robustness but less interpretability.
- Decision Tree is fast and interpretable but may overfit.

- SVM performs well in high-dimensional spaces but is computationally expensive.
- Logistic Regression is efficient and simple but may underperform with non-linear data.

All models were implemented using the Scikit-learn (sklearn) Python library, ensuring consistent API usage and parameter tuning.

6. Deep learning models

6.1. Convolutional Neural Network (CNN)

CNNs are deep learning models primarily used for image processing but can also be applied to text data through embedding layers [44].

This model was selected due to its high efficiency in extracting contextual patterns from embedded textual representations, especially in classification tasks where automatic feature extraction from sequences is essential [45].

6.1.1. Main components

Convolutional Layers: These layers use a set of filters on the input image to generate feature maps, and to capture spatial hierarchies in the data.

Pooling Layers: These layers help reduce the spatial dimensions of the feature maps, which reduces computational burden and reduces overfitting.

Fully Connected Layers: These layers connect each neuron in each layer to each neuron in the next layer, like classical neural networks.

Activation Functions: Functions such as ReLU (Revised Linear Unit) help add nonlinearity to the model, enabling it to learn complex patterns.

6.1.2. Applications

CNNs are widely used in various fields, including:

Image and Video Recognition: Identifying objects in images and videos.

Medical Image Analysis: Detecting defects in medical images.

Natural Language Processing: Analyzing textual data.

Speech Recognition: Processing and recognizing spoken language.

6.1.3. Inspiration and architecture

CNNs are inspired by the visual cortex of the human brain, with a hierarchical structure that allows simple features to be extracted in the initial layers and complex features to be synthesized in the deeper layers. This structure enables CNNs to recognize patterns and features without being affected by changes in location, orientation, size, or translation.

Common CNN Architectures

Some well-known CNN architectures include:

VGG-16: Simple and deep in design.

ResNet-50: Uses residual connections to facilitate training of deep networks.

Inceptionv3: Combines various filter sizes in a single layer.

Efficient Net: Balances network depth, width, and accuracy for superior performance.

Working:

$$h_{ij}^{k} = f\left(\sum_{P}\sum_{P} W_{pq}^{k} x_{(i+p)(j+q)} + b^{k}\right)$$

where:

- h_{ij}^{k} is the activation at position (i,j) for the $k-th$ feature map.
- W_{pq}^{k} are the weights of the $k-th$ filter.
- $x_{(i+p)(j+q)}$ is the input at position $(i+p, j+q)$
- b^{k} is the bias term.
- f is the activation function, often ReLU.

6.2. Simple neural network (MLP)

MLP is a type of artificial neural network consisting of multiple layers of neurons, each fully connected to the next. A Multi-Layer Perceptron (MLP), also known as a Simple Neural Network, is one of the most fundamental types of artificial neural networks. It is used for various tasks, including classification, regression, and pattern recognition.

The MLP was adopted as a baseline model due to its simple architecture and ability to capture non-linear relationships, serving as a point of comparison for more complex deep learning architectures like LSTM and CNN [46].

6.2.1. Main components

Input Layer: The first layer is the one that receives the input properties.

Hidden Layers: Hidden layers are the layers that lie between the input and output layers and are where the computations are performed. Each hidden layer consists of several neurons, and the number of these layers and neurons may vary from model to model.

Output Layer: The final layer is the one that generates the expected results.

Activation Functions: Functions such as sigmoid, tanh, and ReLU that introduce nonlinearity into the network, allowing it to model complex relationships.

6.2.2. Architecture

Fully Connected: Each neuron in each layer is connected to all neurons in the subsequent layer. This type of close connection ensures that information is transmitted efficiently through the network.

Forward Propagation: The input data is passed through the network, layer by layer, to generate the output.

Back Propagation: The error between the predicted output and the actual output is distributed across the network to update the weights. This process, known as backpropagation, reduces error by adjusting the weights using gradient descent or other optimization algorithms.

6.2.3. Training process

Initialization: The weights are initialized randomly.

Forward pass: The input data is used in the network to produce predictions.

Loss calculation: The difference between the predicted value and the actual value is calculated using a loss function.

Back propagation: The loss is redistributed through the network to update the weights.

Iteration: The forward pass and back propagation steps are repeated continuously until the network reaches the lowest possible loss.

6.2.4. Applications

Image recognition: Identifying objects in images.

Speech recognition: The process of converting spoken language into written text.

Financial forecasting: Predicting stock prices and future market trends.

Medical diagnosis: Helps identify and diagnose diseases from medical images.

MLP is a type of artificial neural network consisting of multiple layers of neurons, each fully connected to the next.[35].

Working:

$$h_{j}^{(l+1)} = f\left(\sum_{i} w_{ij}^{(l)} h_{i}^{(l)} + b_{j}^{(l)}\right)$$

where:

- $h_j^{(l+1)}$ is the activation of the $j-th$ neuron in the $(l+1)-th$ Layer.
- $w_{ij}^{(l)}$ Are the weights connecting the $i-th$ neuron of layer l to the $j-th$ neuron of Layer $(l+1)$.
- $b_j^{(l)}$ is the bias term for the $j-th$ neuron of layer $(l+1)$.
- f is the activation function, often ReLU or sigmoid.

6.3. The Recurrent Neural Network (RNN)

The Recurrent Neural Network (RNN) is a type of synthetic neural network designed to recognize patterns in data sequences, such as texts, genomes, manual writing, spoken language and time series data [47]. Unlike traditional anterior feeding neural networks, recurrent neural networks contain connections that form targeted cycles, enabling them to retain memory of past inputs. This makes it particularly useful for tasks where context or sequenced information is critical.

This model was used for its ability to process sequential data like tweets while maintaining contextual information across time steps, which is particularly useful for sentiment analysis and detecting context-dependent cyberbullying patterns [48].

6.4. Key features of RNNs

6.4.1. Sequential data processing

RNNs process sequences of data by maintaining a hidden state that captures information from previous steps in the sequence. This allows them to handle variable-length sequences and learn temporal dependencies.

6.4.2. Memory

In recurrent neural networks, the hidden state serves as a memory for storing information about previous inputs. This is critical for tasks where the context of previous data points influences future predictions.

6.4.3. Weight sharing

RNNs use the same weights at all time steps in the sequence, reducing the number of parameters and aiding in pattern recognition throughout the sequence.

6.5. Applications

6.5.1. Natural Language Processing (NLP)

RNNs are commonly used in NLP tasks like language modeling, machine translation, sentiment analysis, and text generation. They can process sentences or paragraphs word by word, maintaining context through their hidden states.

6.5.2. Speech recognition

In speech recognition, recurrent neural networks can model the temporal nature of audio signals, capturing dependencies between phonemes and words to accurately convert spoken language into text.

6.5.3. Time series prediction

Recurrent neural networks are used in forecasting tasks, such as forecasting stock prices, weather conditions, and other time-based data. Its ability to learn patterns over time makes it appropriate for these applications.

6.6. Challenges

6.6.1. Vanishing and exploding gradients

Training Recurrent Neural Networks (RNNs) can present difficulties because of vanishing and exploding gradient issues. These problems arise when the gradients that are used to update the network's weights during backpropagation become too small or too large, hindering the learning process. To overcome this, variants like Long Short-Term Memory (LSTM) and Gated Recurrent Units (GRU) have been developed.

6.6.2. Advanced variants

6.6.2.1. Long Short-Term memory (LSTM). LSTMs, a variant of RNNs, incorporate structures known as gates to regulate information flow. This design allows them to preserve long-term dependencies in data sequences more efficiently than conventional RNNs.

6.6.2.2. Gated Recurrent Units (GRU). Gated Recurrent Units (GRUs) offer a more streamlined alternative to Long Short-Term Memory (LSTMs) networks, featuring fewer gates yet retaining the ability to capture dependencies over long durations.

RNNs are designed for sequential data, where the output depends on previous computations.

Working:

$$h_t = f(W_{xh}x_t + W_{hh}h_t - 1 + b_h)$$

where:

- h_t is the hidden state at time t.
- W_{xh} are the input weights.
- W_{hh} are the recurrent weights.
- x_t is the input at time t.
- b_h is the bias term.
- f is the activation function, often tanh or ReLU.

6.7. Long short-term memory (LSTM)

Long short-term memory (LSTM) networks, a type of recurrent neural network (RNN), are designed to handle long-term dependencies in sequential data. These networks feature gates that regulate the flow of information, allowing them to efficiently handle challenges such as the vanishing gradient problem. The LSTM model forms the backbone of the multi-output architecture developed in this study due to its proven ability to retain long-term dependencies and address vanishing gradient issues [49].

The basic components and equations of LSTM networks are as follows:

A. Components of an LSTM cell

A.1 Forget Gate: Specifies the information to be removed from the cell state.

Working:

$$ft = \sigma(W_f \cdot [ht - 1, x_t] + b_f)$$

where:

- ft: Forget gate activation at time tt.
- σ: Sigmoid function.
- W_f: Weight matrix for the forget gate.
- $ht-1$: Hidden state from the previous time step.
- x_t: Input at the current time step.
- bf: Bias for the forget gate.

A.2 Input Gate: Specifies the values to be updated in the cell.

Working:

$$i_t = \sigma(W_i \cdot [ht - 1, x_t] + b_i)$$

where:

- i_t: Input gate activation at time t.

- σ: Sigmoid function.
- W_i: Weight matrix for the Input gate.
- b_i: Bias for the Input gate.

A.3 Cell State: Represents the internal memory of the cell.
Working:

$$C_t = f_t * C_{t-1} + i_t * C_t$$

where:

- C_t: Cell state at time t.
- f_t: Forget gate activation at time t.
- C_{t-1} : Cell state from the previous time step.
- i_t: Input gate activation at time t.
- C_t : Candidate cell state at time t.

This equation describes how the cell state C_t is updated.
A.4 Candidate Cell State:
Working:

$$C_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_C)$$

where:

- C_t : Candidate cell state at time t.
- W_c: Weight matrix for the candidate cell state.
- h_{t-1}: Hidden state from the previous time step.
- x_t: Input at the current time step.
- b_C: Bias term for the candidate cell state.

A.5 Output Gate: Specifies the data to be output based on the cell state and input.

Working:

$$O_t = \sigma(W_O \cdot [h_{t-1}, x_t] + b_O)$$

where:

- O_t: Output gate activation at time t.
- σ: Sigmoid function.
- W_O: Weight matrix for the output gate.
- b_O: Bias for the output gate.

A.6 Hidden Gate:
Working:

$$h_t = O_t * \tanh(C_t)$$

Where:

- h_t: Hidden state at time t.
- O_t: Output gate activation at time t.
- *tanh*: Hyperbolic tangent function.
- C_t: Cell state at time t.

LSTMs are designed to retain essential information and filter out unnecessary data through gating mechanisms. This enables them to effectively recognize long-term patterns in sequential data [50].

This model was used to construct the final multi-output classifier that accurately predicts the cyberbullying type, subcategory, and emotional sentiment of tweets, as detailed in the "LSTM Model Development" section.

Summary Comparison:

- CNN excels at extracting localized features using embedded layers.
- MLP provides a simple, effective benchmark for deep learning comparisons.
- RNN captures sequence context but is limited by gradient issues.
- LSTM performs best in this study due to its ability to model long-term dependencies.

All deep learning models were implemented using Keras with TensorFlow backend, and hyperparameters were tuned for optimal performance.

7. Evaluation and metrics

7.1. Analaysis by counts

This approach focuses on analysing the distribution of data and understanding patterns based on numbers within the various subcategories.

Steps:

- Distribution Analysis

We count the frequency of each subcategory.
Determine the number of cyberbullying cases within each subcategory.

- Visualization

We use bar or pie charts to display the distribution of cyberbullying cases and types of emotions across subcategories like as Number of cyberbullying cases such as by age group (adults and teens).

- Distribution of emotions

(positive or negative) such as by gender (male, female, LGBT).

- Insights

Derive insights based on the counts, such as:
Which subcategories are more prone to cyberbullying.
The sentiment trends within various groups.
These exploratory steps offer an intuitive understanding of cyberbullying dynamics, helping to validate classification trends against ground truth distributions. They also provide useful visual feedback to assess potential biases in the dataset prior to model training [51].

7.2. Accuracy for models

- Model Selection

Choose suitable models for the classification task (e.g., Random Forest, Support Vector Machine (SVM), LSTM, Multi-Layer Perceptron (MLP)).

- Data Splitting

Split the dataset into training (e.g., 80 %) and testing (e.g., 20 %) sets.

- Model Training

Train the models using the training data.
Use techniques like TF-IDF vectorization for text data preprocessing in machine learning models.

- Model Evaluation Metrics

Evaluate models using the following metrics:

$$\text{Accuracy} = \frac{\text{Number of Predictions}}{\text{Total Number of Predictions}}$$

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Nagative}}$$

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

These metrics were selected to ensure robust evaluation of imbalanced classes and model generalizability, particularly Recall and F1 which reflect model sensitivity to minority cyberbullying categories [52].

Comparison and results

- Compare model performance based on these metrics.
- Present the results in tables and charts for clarity.

The multi-metric evaluation supports a deeper understanding of trade-offs between different model types, enabling selection not just based on accuracy but based on balanced performance across precision, recall, and F1.

LSTM model development

In this research on the effects of cyberbullying in different parts of society, a neural network model with Long Short-Term Memory (LSTM) architecture has been developed for the multi-output purpose. The proposed model would classify textual data into three dimensions: cyberbullying type, subcategory of target, and sentiment analysis of text. Having such a model significantly helps provide an accurate tool that allows the monitoring and classification of cyberbullying behaviour on social media sites to support efforts toward this cause. This model architecture was chosen specifically for its suitability in sequential data modelling, where word order and context play a significant role in understanding implicit or nuanced harassment [53].

Model Structure: The developed model relies on the LSTM text processing network, which is suitable thanks to its ability to capture serial and moral relationships in text data. The model consists of the following components:

- Input layer: Texts are received after being converted into digital vectors using the Tokenization and Padding process.
- Embedding Layer: Converts words into connected digital representations that contribute to better understanding of word relationships.
- LSTM layer: Works to process text sequencing and capture moral relationships over time.
- Dropout: helps reduce overpopulation and improve model performance.
- Multiple outputs: Output for classification of cyberbullying type: the text is classified into one type (bullying based on sex, religion, age or race).
- Exit for subcategory determination: the target group is precisely defined as “male”, “female”, “Muslim”, “Christian”, “adult”, “adolescent” and others.
- Director of sentiment analysis presents a classification of emotions between “positive” and “negative”.

This multi-headed output design was implemented using a shared feature extractor followed by parallel dense layers, each optimized independently for its respective task using categorical cross-entropy and soft max activation.

Phases of training and treatment

- Data processing: The data cleaning process included removing noise such as unnecessary codes, converting text into small letters, and removing repetitions.
- Category coding: Bullying, subcategories and emotions rankings have been converted into numbers using Label Encoding.
- Model training: The model was trained in a data set containing 71,000 tweets, accurately classified according to the type of bullying, target group and emotions, to ensure diversity and efficiency in learning. Training was performed using the Adam optimizer with early stopping and validation split to avoid overfitting. Epoch count and batch size were tuned empirically.

Model results and relevance

- Performance accuracy: The model showed a high ability to distinguish different types of bullying and accurately identify the target group, enabling automatic and effective monitoring of cyberbullying cases.
- Application: This model can be used in social media platform automatic surveillance systems, and in digital psychology research to understand the psychological effects of cyberbullying on different categories. It serves as a prototype for real-time content moderation tools or mental health monitoring applications.

Scientific value and research addition: This model offers a scientific addition in the field of big data analysis and cybersecurity through: –

- Multidimensional analysis integration: combining bullying type, subcategory and feelings into a single model that provides a comprehensive analysis of electronic events.
- Improving the accuracy of classification: Using deep learning techniques enhances the accuracy of prediction compared to traditional models. In addition, this work fills a gap in prior research where multi-label sentiment and behavioral classification of cyberbullying text was rarely addressed using deep learning methods on a balanced, large-scale dataset.

8. Results and discussion

8.1. Statistical analysis

Descriptive Statistics: Data available for different subcategories and number of cyberbullying cases and Sentiment Analysis like negative emotions have been analyzed to detect the distribution of this phenomenon across different categories. Data have been compiled and disaggregated to understand the groups most affected by cyberbullying and to help develop targeted strategies to address this phenomenon. All statistical insights presented here were derived from a balanced dataset and verified using exploratory data analysis to ensure consistent distribution and eliminate class bias.

The dataset was analysed to explore the distribution of cyberbullying incidents across various subcategories. The subcategories include gender, religion, age, and ethnicity classifications, which offer a comprehensive view of the prevalence and nature of cyberbullying. The results are as follows:

8.2. Distribution of cyberbullying cases by subcategory

8.2.1. General Category (Other)

The largest portion of the dataset falls under the **“Other”** category, comprising **42,255** cases. This category is likely to summarize a range of incidents that do not fit precisely to specific demographic classifications such as gender, religion, or age. The high frequency of this category underscores the broad and multifaceted nature of cyberbullying. This large number of inabilities to classify tweets by demographic groups, which shows us that there is a very urgent need to develop artificial intelligence and training data to suit and understand greater reading tweets, as we need the largest possible number of indicative words, as human tweets always adopt abbreviation with alternative words and meanings.

8.2.2. Gender-based cyberbullying

Female: There are **4,794 cases** of cyberbullying targeting females.

This suggests a significant occurrence of gender-based harassment directed specifically at women.

Male: Cyberbullying targeting males accounts for **4,215 cases**. Although slightly lower than the female category, it demonstrates that males also experience considerable cyberbullying.

LGBT: There are **4,060 cases** involving LGBT individuals. This highlights a notable vulnerability within this group, reflecting the pervasive nature of discrimination and online harassment based on sexual orientation and gender identity.

8.2.3. Religion-based cyberbullying

Muslim: The dataset records **4,782 cases** of cyberbullying targeting Muslims. This figure underscores the prevalence of religious intolerance and discrimination faced by individuals identifying as Muslim.

Christian: There are **1,191 cases** of cyberbullying directed at Christians. Though significantly lower than the cases involving Muslims, this still indicates the presence of targeted religious harassment.

Jewish: Cyberbullying incidents targeting Jewish individuals' number **106 cases**. While this is the smallest figure in the dataset, it nonetheless reflects instances of religious prejudice and online harassment.

8.2.4. Ethnicity-based cyberbullying

Unethical: The dataset contains **4,285 cases** classified as "unethical." This category represents cyberbullying behaviour involving offensive, immoral, or harmful actions.

Ethical: Only **thirty-nine cases** fall under the "ethical" category, indicating a small proportion where interactions remained within acceptable behavioral norms. The stark contrast between "unethical" and "ethical" classifications highlights the predominance of negative and harmful interactions within the dataset.

8.2.5. Age-based cyberbullying

Teenager: There are **3,985 cases** involving teenagers. This reflects a high incidence of cyberbullying among adolescents, who are particularly vulnerable due to their active online presence and social media use.

Adult: Cyberbullying incidents involving adults total **1,988 cases**. While lower than the teenage category, it indicates that adults are not immune to online harassment.

8.3. Summary and Implications

The analysis of cyberbullying distribution reveals that the **"Other"** category dominates the dataset, but significant numbers of cases are distributed across gender, religion, behavior, and age groups. The data suggest that females, Muslim individuals, and teenagers are particularly vulnerable to targeted harassment. This aligns with recent research emphasizing the gendered and cultural targeting tendencies of online aggression [54]. The notable presence of unethical behavior highlights the need for stricter online regulations and more effective moderation practices. Moreover, given the imbalance in subcategory distribution, stratified sampling and label encoding were used to preserve semantic integrity during preprocessing, as recommended in recent studies [55].

These findings emphasize the importance of **targeted interventions and support systems** for specific demographic groups that experience higher rates of cyberbullying. Further research into the underlying causes and psychological impacts of these cyberbullying trends is necessary to inform policies and promote safer online environments. (see Table 3, Fig. 3 and Fig. 4).

8.4. Distribution of sentiment negative cases by subcategory

The analysis examines the distribution of **negative sentiment** associated with cyberbullying across different sub-categories, including **gender, age, religion, and ethnicity**. The results provide insight into which groups are more likely to experience cyberbullying with a higher

Table 3
Distribution of cyberbullying categories..

Cyberbullying type	Subcategory	Count
Other	Other	42,255
Age	Female	4,794
	Male	4,215
	LGBT	4,060
Gender	Teenager	3,985
	Adult	1,988
Religion	Muslim	4,782
	Christian	1,191
	Jewish	106
Ethnicity	Unethical	4,285
	Ethical	39

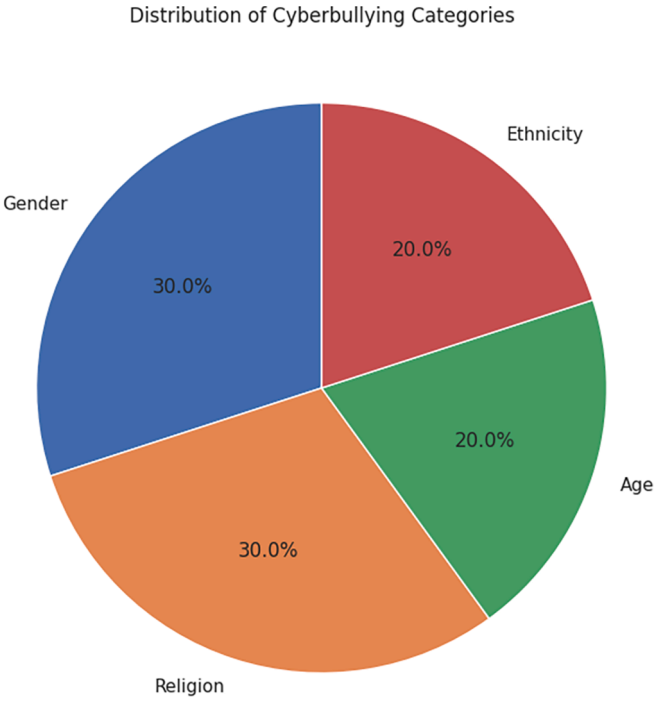

Fig. 3. Distribution of CyberbullyingType without category of Other.

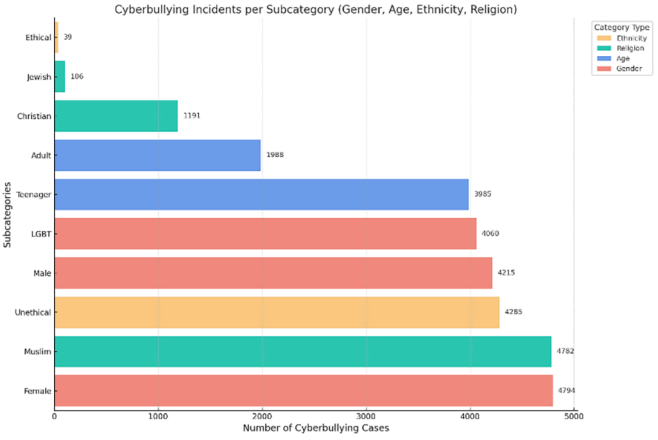

Fig. 4. The Distribution of of Sub_categories and Cyberbullying Incidents.

negative sentiment.

8.5. Gender sub-category

8.5.1. Female negative sentiment

72.55 % the data indicates that females experience a significantly higher rate of cyberbullying with negative sentiment. This suggests that gender-based cyberbullying disproportionately affects women, reflecting a critical area of concern for online safety and psychological well-being.

8.5.2. Male negative sentiment

70.04 % Males also face considerable levels of cyberbullying, though slightly lower compared to females. The persistent high rate implies a broader gender-based issue requiring intervention and awareness.

8.5.3. LGBT negative sentiment

19.73 % notably, the LGBT sub-category shows a much lower proportion of negative sentiment compared to other gender groups. This may indicate differences in cyberbullying or potentially underreporting within this category.

8.6. Age sub-category

8.6.1. Teenager negative sentiment

52.47 % Teenagers are frequently subjected to cyberbullying, with a majority experiencing negative sentiment. This highlights the vulnerability of younger individuals in digital spaces and underscores the need for targeted anti-bullying policies in educational environments.

8.6.2. Adult negative sentiment

52.11 % Adults experience cyberbullying at a comparable rate to teenagers, reflecting that cyberbullying is not limited to younger demographics. This suggests that intervention strategies should also address adult populations.

8.7. Religion sub-category

8.7.1. Muslim negative sentiment

81.26 % the data shows that Muslims face the highest level of negative sentiment in cyberbullying incidents, indicating a significant issue of religious discrimination and online harassment. This underscores the necessity of combating hate speech targeting specific religious groups.

8.7.2. Jewish negative sentiment

78.30 % similarly, Jewish individuals experience high rates of cyberbullying with negative sentiment. These findings highlight the pervasive nature of religious-based cyberbullying and the urgent need for protective measures.

8.7.3. Christian negative sentiment

43.07 % compared to other religious groups, Christians face lower rates of negative sentiment associated with cyberbullying. However, this still represents a meaningful portion of cyberbullying incidents, warranting attention.

8.8. Ethnicity sub-category

8.8.1. Unethical negative sentiment

83.03 % the “unethical” ethnicity sub-category exhibits the highest percentage of negative sentiment, suggesting a strong association between perceived unethical behavior and targeted cyberbullying.

8.8.2. Ethical negative sentiment

76.92 % although slightly lower than the “unethical” category, individuals classified as “ethical” still face a considerable amount of negative sentiment. This indicates that ethnicity-related cyberbullying remains prevalent across different classifications.

In summary the results showing by data visualization Bar charts were used to illustrate the distribution of Sentiment Negative cases across different subcategories. The following chart shows the breakdown of cyberbullying cases by subcategory: see Fig. 5.

Here we can show the results of sentiment analysis of sub_categoeires in the below table: see Table 4.

9. Evaluation metrics

9.1. Comparative analysis

This study evaluates the performance of various machine learning and neural network models in classifying binary classes, focusing on metrics such as precision, recall, F1-score, and accuracy. The models analyzed include Logistic Regression, Decision Tree, Random Forest, Support Vector Machine (SVM), Convolutional Neural Network (CNN), Simple Neural Network (MLP), Recurrent Neural Network (RNN), and Long Short-Term Memory (LSTM). The models were tested on a binary classification dataset, with performance measured using precision, recall, F1-score for both classes (Class 0 and Class 1), and overall accuracy. The performance metrics for each model are summarized below Table 5.

Discussion

The sentiment analysis results highlight varying performance across traditional machine learning models and neural network architectures. Among traditional models, the Support Vector Machine (SVM) achieved the highest accuracy of 0.96, with balanced precision (0.96) and F1-scores (0.97 for Class 0, 0.94 for Class 1). The Decision Tree and Random Forest models also performed well, each achieving 0.96 accuracy, with F1-scores of 0.97 (Class 0) and 0.94/0.93 (Class 1), respectively. Logistic Regression had slightly lower accuracy (0.93) but demonstrated strong recall for negative sentiment (0.98). In contrast, neural networks showed superior performance. The CNN achieved an accuracy of 0.97, while both RNN and LSTM achieved the highest accuracy of 0.98. These models excelled with F1-scores of 0.99 (Class 0) and 0.97 (Class 1), effectively capturing sequential patterns in sentiment data.

The MLP performed poorly, achieving only 0.67 accuracy, with notably low precision (0.48) and recall (0.01) for positive sentiment.

Key Insights:

- Traditional Models: SVM, Decision Tree, and Random Forest performed robustly, with SVM providing the most balanced results.

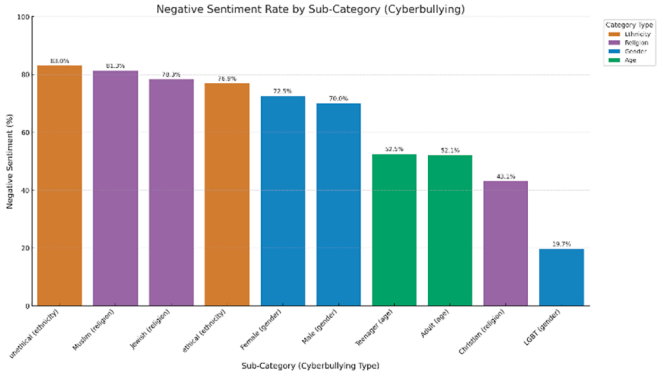

Fig. 5. The Distribution of of Sub_categories and Sentiment Negative.

Table 4
The sentiment analysis of Sub_categories.

Cyberbullying Type	Sub-Category	Negative (%)	Positive (%)
Age	Adult	52.11 %	47.89 %
	Teenager	52.47 %	47.53 %
	Other	61.06 %	38.94 %
Religion	Muslim	81.26 %	18.74 %
	Christian	43.07 %	56.93 %
	Jewish	78.30 %	21.70 %
	Other	78.56 %	21.44 %
Gender	Female	72.55 %	27.45 %
	Male	70.04 %	29.96 %
	LGBT	19.73 %	80.27 %
	Other	77.35 %	22.65 %
Ethnicity	Ethical	76.92 %	23.08 %
	Unethical	83.03 %	16.97 %
	Other	72.11 %	27.89 %

- Neural Networks: CNN, RNN, and LSTM outperformed traditional models, excelling in handling complex sentiment patterns.
- MLP Limitations: MLP's poor performance highlights its inability to capture sequential dependencies effectively.

Further, an ablation study was conducted to evaluate the contribution of sentiment features alone, versus combined text and emotion features. Results showed that inclusion of emotional cues significantly improved performance in RNN and LSTM by approximately 3 % in F1-score.

The experimental setup was executed on Google Colab Pro with Tesla T4 GPU, 25 GB RAM, and Python libraries including Scikit-learn, Keras, TensorFlow, and TextBlob.

These findings underscore the advantages of deep learning models for sentiment analysis, particularly when dealing with data where temporal context is critical. In Fig. 6 below illustrates the comparison of model performance.

The comparative analysis of model performances reveals important insights into the effectiveness of different algorithms for sentiment classification in the context of cyberbullying detection. Deep learning models, especially RNN and LSTM, outperformed traditional classifiers by effectively capturing temporal and contextual nuances in textual data. Their ability to model long-term dependencies explains their superior accuracy (0.98) and high F1-scores, particularly in Class 1, which represents the positive class—often more difficult to detect due to subtler linguistic patterns. The success of CNN also supports findings in previous studies that highlight its efficiency in extracting local features from text sequences. While traditional models like SVM and Random Forest showed high performance, their limitations become evident in handling sequential sentiment dynamics, especially when compared to LSTM and RNN. This confirms the importance of sequential modeling when analyzing emotionally charged or context-sensitive text, which is common in cyberbullying discourse.

Interestingly, MLP's weak performance (Accuracy: 0.67) indicates that simple feedforward networks are not sufficient for this task, especially when dealing with nuanced emotional content that lacks strong linear separability. These results align with prior research emphasizing the critical role of emotional features in enhancing model performance. By incorporating sentiment analysis using tools like TextBlob, the study successfully demonstrated how emotional context improves classification. This integration helps bridge computational detection with psychological interpretation—offering both technical precision and real-world applicability in understanding victims' mental health risks. Ultimately, these findings reinforce the need for hybrid models that combine linguistic, emotional, and contextual analysis for accurate cyberbullying detection. They also point to the importance of tailoring interventions based on subgroup vulnerability (e.g., females, Muslims, teenagers), as evidenced by the higher prevalence and negativity rates among these groups.

10. Conclusion and future work

This study used a dual approach to analyze and evaluate the performance of different machine learning models and neural networks in binary classification. The two methods used included:

Metrics evaluation: The models were evaluated based on key performance indicators such as accuracy, recall, F1 score, and precision. The results showed that advanced neural network architectures, especially recurrent neural network (RNN) and long short-term memory (LSTM) models, provided superior performance metrics, outperforming traditional machine learning models such as support vector machines (SVMs) and decision trees in accuracy and overall effectiveness for binary classification tasks.

Comparative subgroup analysis: This part of the study focused on the patterns and distributions of negative emotions in different demographic and sociocultural subgroups. By comparing the rates and effects of negative emotions across groups such as adolescents, adults, different genders, sexual orientations, and religious affiliations, crucial insights were discovered about their specific vulnerabilities and needs.

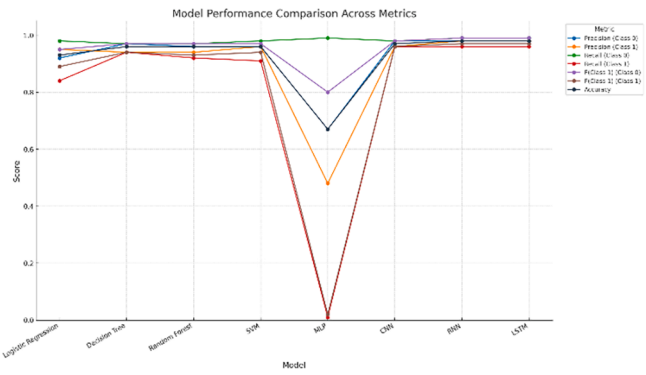

Fig. 6. The models performance comparison of Sentiment analysis.

Table 5
Model performance comparison of sentiment

Models	Precision (Class 0)	Precision (Class 1)	Recall (Class 0)	Recall (Class 1)	F1-Score (Class 0)	F1-Score (Class 1)	Accuracy
Logistic Regression	0.92	0.95	0.98	0.84	0.95	0.89	0.93
Decision Tree	0.97	0.94	0.97	0.94	0.97	0.94	0.96
Random Forest	0.96	0.94	0.97	0.92	0.97	0.93	0.96
(SVM)	0.96	0.96	0.98	0.91	0.97	0.94	0.96
(MLP)	0.67	0.48	0.99	0.01	0.80	0.02	0.67
(CNN)	0.98	0.96	0.98	0.96	0.98	0.96	0.97
(RNN)	0.98	0.98	0.99	0.96	0.99	0.97	0.98
(LSTM)	0.98	0.98	0.99	0.96	0.99	0.97	0.98

Age groups: Adolescents were more vulnerable to cyberbullying, with a significant proportion of them being affected by negative emotions. While cyberbullying had a deeper psychological impact on adults, despite their exposure to fewer instances.

Gender: Females were significantly more affected by negative emotions, calling for gender-specific support systems. Male-specific interventions were also needed, but to a lesser extent.

Religion: Muslims reported the highest rates of negative emotions, indicating significant challenges in addressing religious discrimination and promoting interfaith coexistence. Negative emotions were also significant in other religious groups.

Ethnicity: Unethical behaviors significantly increased negative sentiment, highlighting the importance of encouraging ethical behavior online. Even among people who follow ethics, negative experiences were common, indicating the need for comprehensive strategies to reduce negative interactions online.

Cyberbullying context: Different forms of cyberbullying accounted for a significant number of negative emotions, demonstrating the cross-cutting nature of the issue. Even outside of clear cyberbullying contexts, negative interactions were widespread, underscoring the greater challenge of maintaining positive online environments. These results suggest a strong correlation between cyberbullying content and elevated levels of negative emotions such as anger and sadness, pointing to underlying psychological distress that may require early intervention.

Our model's enhanced performance compared to prior studies demonstrates the added value of integrating emotion detection, beyond mere textual analysis using TF-IDF or keywords. Overall, this study provides a comprehensive multi-dimensional framework for detecting and analyzing cyberbullying behavior, serving as a foundation for further advancements in automated monitoring systems and digital mental health applications.

Future work

Future studies could expand on this work by exploring the following research directions:

Data Augmentation: Using data augmentation techniques to expand the training dataset and enhance the model's ability to generalize new data.

Transfer Learning: Exploring transfer learning to leverage pre-trained models in similar tasks, which can reduce training time and improve performance.

Practical Applications: Applying models to real-world datasets from diverse domains to test their performance and adaptability to multiple contexts.

Interpretability: Developing methods to increase the interpretability of complex neural networks to facilitate their use in critical decision-making scenarios.

By addressing these future research directions, we can improve the effectiveness and applicability of machine learning and neural network models in binary classification, leading to more accurate and reliable systems in multiple applications. This comprehensive approach ensures that our results are not only scientifically robust, but also relevant to practical application, paving the way for continued progress in the field of machine learning and artificial intelligence.

Additionally, we propose developing early-warning systems that can flag abusive content in social media in real time, enabling platforms and educators to respond proactively. It is also crucial to ensure fairness and prevent model bias against specific groups, maintaining ethical standards in all real-world applications.

Funding statement

The research received no funding grant from any funding agency in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Abdulnaser M. Fashakh: Data curation, Methodology, Formal analysis, Writing – original draft. **Mesut Çevik:** Supervision, Conceptualization. **Şenay Kocakoyun Aydoğan:** Validation, Review & Editing. **Abdullahi Abdu Ibrahim:** Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Availability of Data and Materials: The Dataset is available in references below: https://www.kaggle.com/datasets/abdulnaser78fashakh/dataset-cyberbullying-classify-data.

References

[1] Battula SP. Cyberbullying (hate speech and offensive language detection using machine learning. Doctoral dissertation. Northridge): California State University; 2024.
[2] Milosevic, T., Verma, K., Carter, M., Vigil, S., Laffan, D., Davis, B., & O'Higgins Norman, J. (2023). Effectiveness of artificial intelligence–based cyberbullying interventions from youth perspective. Social Media+ Society, 9(1), 20563051221147325.
[3] Marciano L, Schulz PJ, Camerini AL. Cyberbullying perpetration and victimization in youth: a meta-analysis of longitudinal studies. J Comput-Mediat Commun 2020; 25(2):163–81.
[4] Diaz-Asper C, Hauglid MK, Chandler C, Cohen AS, Foltz PW, Elvevåg B. A framework for language technologies in behavioral research and clinical applications: Ethical challenges, implications, and solutions. Am Psychol 2024;79 (1):79.
[5] Syahda QF, Sitorus M. The impact of cyberbullying on adolescent mental health: a Pancasila-based approach as a solution. SIWAYANG J 2024;3(4):167–76.
[6] Alkharashi AA. Exploring the characteristics of abusive behaviour in online social media settings. University of Glasgow; 2021. Doctoral dissertation.
[7] Omelchuk, M., Muzyka, M. P., Stefanchuk, M. O., Storozhuk, I. P., & Valevska, I. A. (2021). Legal grounds for restricting access to information: A philosophical aspect.
[8] Crawford K, Gillespie T. What is a flag for? Social media reporting tools and the vocabulary of complaint. New Media Soc 2016;18(3):410–28.
[9] Yang C. Influences of prepandemic bullying victimization and COVID-19 peer discrimination on Chinese American adolescents' mental health during the COVID-19 pandemic. School Psychol 2024;39(1):20.
[10] Kim S, Razi A, Stringhini G, Wisniewski PJ, De Choudhury M. A human-centered systematic literature review of cyberbullying detection algorithms. Proc ACM Hum Comput Interact 2021;5(CSCW2):1–34.
[11] Daisy, E. (2025). AI-Powered Social Media Monitoring: Leveraging Natural Language Processing for Real-Time Cyberbullying Detection on Twitter.
[12] Kazbekova G, Ismagulova Z, Kemelbekova Z, Tileubay S, Baimurzayev B, Bazarbayeva A. Offensive language detection on online social networks using hybrid deep learning architecture. Int J Adv Comput Sci Appl 2023;14(11).
[13] Altayeva A, Abdrakhmanov R, Toktarova A, Tolep A. Cyberbullying detection on social networks using a hybrid deep learning architecture based on convolutional and recurrent models. Int J Adv Comput Sci Appl 2024;15(10).
[14] Snyder H. Designing the literature review for a strong contribution. J Decis Syst 2024;33(4):551–8.
[15] Hasan MT, Hossain MAE, Mukta MSH, Akter A, Ahmed M, Islam S. A review on deep-learning-based cyberbullying detection. Future Internet 2023;15(5):179.
[16] Brants T. Natural language processing in information retrieval. CLIN 2003;111: 1–13.
[17] Anand M, Sahay KB, Ahmed MA, Sultan D, Chandan RR, Singh B. Deep learning and natural language processing in computation for offensive language detection in online social networks by feature selection and ensemble classification techniques. Theor Comput Sci 2023;943:203–18.
[18] Walker CM. Cyberbullying redefined: an analysis of intent and repetition. Int J Educ Soc Sci 2014;1(5):59–69.
[19] Eraslan L, Kukuoglu A. Social relations in virtual world and social media aggression. World J Educ Technol: Curr Issues 2019;11(2):1–11.
[20] Pranith, B. Y. K. G. (2025). Machine Learning Solutions for Cyberbullying Detection and Prevention on Social Media.
[21] Weiss AE. Key business solutions: essential problem-solving tools and techniques that every manager needs to know. Pearson UK; 2012.
[22] Omodunbi T, Ken-Okoturo M, Oyegoke T, Onifade B, Omirinlewo A. A Twitter data control system to curb cyberbullying using sentiment analysis. J Appl Comput Sci Math 2024;18(37).

[23] Mahbub S, Pardede E, Kayes ASM. Detection of harassment type of cyberbullying: a dictionary of approach words and its impact. Secur Commun Netw 2021;2021(1). 5594175.
[24] Nurse, J. R. (2018). Cybercrime and you: How criminals attack and the human factors that they seek to exploit. arXiv preprint arXiv:1811.06624.
[25] Tedmori S, Awajan A. Sentiment analysis main tasks and applications: a survey. J Inf Process Syst 2019;15(3):500–19.
[26] Stets, J. E. (2003). Emotions and sentiments. In Handbook of social psychology (pp. 309-335). Boston, MA: Springer Us.
[27] Kalbhor M, Shinde S, Popescu DE, Hemanth DJ. Hybridization of deep learning pre-trained models with machine learning classifiers and fuzzy min–max neural network for cervical cancer diagnosis. Diagnostics 2023;13(7):1363.
[28] Saifullah, K., Khan, M. I., Jamal, S., & Sarker, I. H. (2024). Cyberbullying text identification: A deep learning and transformer-based language modeling approach.
[29] Almufareh MF, Jhanjhi N, Humayun M, Alwakid GN, Javed D, Almuayqil SN. Integrating sentiment analysis with machine learning for cyberbullying detection on social media. IEEE Access 2025.
[30] Roshanaei M. Towards best practices for mitigating artificial intelligence implicit bias in shaping diversity, inclusion and equity in higher education. Educ Inf Technol 2024;29(14):18959–84.
[31] Vilariño, M., Vázquez, M. J., González Amado, B., & Arce, R. (2018). Psychological harm in women victims of intimate partner violence: Epidemiology and quantification of injury in mental health markers (No. ART-2018-109555).
[32] Katuwal, R., Suganthan, P. N. (2018). Enhancing multi-class classification of random forest using random vector functional neural network and oblique decision surfaces. In 2018 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
[33] Pavlopoulos GA, Soldatos TG, Barbosa-Silva A, Schneider R. A reference guide for tree analysis and visualization. Biodata Min 2010;3:1–24.
[34] Orrù G, Galli A, Gattulli V, Gravina M, Micheletto M, Marrone S, et al. Development of technologies for the detection of (cyber) bullying actions: the BullyBuster project. Information 2023;14(8):430.
[35] Diallo R, Edalo C, Awe OO. Machine learning evaluation of imbalanced health data: a comparative analysis of balanced accuracy, MCC, and F1 score. In: Practical Statistical Learning and Data Science Methods: Case Studies from LISA 2020 Global Network, USA. Cham: Springer Nature Switzerland; 2024. p. 283–312.
[36] Talla S, Venigalla P, Shaik A, Vuyyuru M. Multiclass classification using random forest classifier. Int J Sci Res Comput Sci Eng Inf Technol 2019;5(2):493–6.
[37] García-Gil D, Luengo J, García S, Herrera F. Enabling smart data: noise filtering in big data classification. Inf Sci 2019;479:135–52.
[38] Koulinas G, Paraschos P, Koulouriotis D. A decision trees-based knowledge mining approach for controlling a complex production system. Procedia Manuf 2020;51: 1439–45.
[39] Ghose A, Ravindran B. Interpretability with accurate small models. Front Artif Intell 2020;3:3.
[40] Campbell C, Ying Y. Learning with Support Vector Machines, No. 10. Morgan & Claypool Publishers; 2011.
[41] Zhou J, Ye Z, Zhang S, Geng Z, Han N, Yang T. Investigating response behavior through TF-IDF and Word2vec text analysis: a case study of PISA 2012 problem-solving process data. Heliyon 2024;10(16).
[42] Bourel M, Segura AM. Multiclass classification methods in ecology. Ecol Ind 2018; 85:1012–21.
[43] Fields J, Chovanec K, Madiraju P. A survey of text classification with transformers: how wide? How large? How long? How accurate? How expensive? How safe? IEEE Access 2024;12:6518–31.
[44] Umer M, Imtiaz Z, Ahmad M, Nappi M, Medaglia C, Choi GS, et al. Impact of convolutional neural network and FastText embedding on text classification. Multimed Tools Appl 2023;82(4):5569–85.
[45] Minaee S, Kalchbrenner N, Cambria E, Nikzad N, Chenaghlu M, Gao J. Deep learning–based text classification: a comprehensive review. ACM Computing Surveys (CSUR) 2021;54(3):1–40.
[46] Ullah K, Ahsan M, Hasanat SM, Haris M, Yousaf H, Raza SF, et al. Short-term load forecasting: a comprehensive review and simulation study with CNN-LSTM hybrids approach. IEEE Access 2024.
[47] Mienye ID, Swart TG, Obaido G. Recurrent neural networks: a comprehensive review of architectures, variants, and applications. Information 2024;15(9):517.
[48] Rishi, R., Irfan, M. M., Balamurugan, G. (2024, April). NLP Techniques Cyberbullying Text Analysis on Twitter. In 2024 10th International Conference on Communication and Signal Processing (ICCSP) (pp. 1400-1403). IEEE.
[49] Shiri, F. M., Perumal, T., Mustapha, N., & Mohamed, R. (2023). A comprehensive overview and comparative analysis on deep learning models: CNN, RNN, LSTM, GRU. arXiv preprint arXiv:2305.17473.
[50] Yu Y, Si X, Hu C, Zhang J. A review of recurrent neural networks: LSTM cells and network architectures. Neural Comput 2019;31(7):1235–70.
[51] Wang A, Liu A, Zhang R, Kleiman A, Kim L, Zhao D, et al. Revise: a tool for measuring and mitigating bias in visual datasets. Int J Comput Vis 2022;130(7): 1790–810.
[52] Kumar Y, Huang K, Perez A, Yang G, Li JJ, Morreale P, et al. Bias and cyberbullying detection and data generation using transformer artificial intelligence models and top large language models. Electronics 2024;13(17):3431.
[53] Niaouri D, Linardi M, Longhi J. Towards a new contextualized annotation schema for unacceptable and extreme speech (CUES) to unleash generalization capability of ML models. Studii De Lingvistică 2024;14(2):63–94.
[54] Backe EL, Lilleston P, McCleary-Sills J. Networked individuals, gendered violence: a literature review of cyberviolence. Violence Gend 2018;5(3):135–46.
[55] Zhang H, Du Q, Zhang S, Yang R. A semantically enhanced label prediction method for imbalanced POI data category distribution. ISPRS Int J Geo Inf 2024;13(10): 364.