International Journal of Advances in Scientific Research and Engineering (ijasre)

E-ISSN : 2454-8006

DOI: 10.31695/IJASRE.2025.3.3

Volume 11, Issue 3
March - 2025

Subject Review: Cyberbullying and Detection Methods

Amal Abbas Kadhim[1], Zainab Khyioon Abdalrdha[2], and Wedad Abdul Khuder Naser [3]

[1]Assistant Professor, [2]Assistant Professor, Dr., [3]Assistant Professor

[1,3]Mustansiriyah University/ College of Education, Department of Computer Science

[2]Mustansiriyah University/ College of Basic Education, Department of Computer Science

Baghdad- Iraq

ABSTRACT

Cyberbullying is a prevalent issue on social media, causing significant mental and social harm to victims. This review highlights cutting-edge cyberbullying detection technologies, concentrating on machine learning (ML), deep learning (DL), and natural language processing (NLP). The learners are categorized into supervised, unsupervised, and hybrid models, highlighting their pros and cons. This article covers common research datasets, including social media comments, and addresses difficulties including data imbalance, linguistic diversity, and context interpretation. Future efforts include developing context-sensitive models, improving on-the-fly detection, and addressing ethical concerns in automated system deployment.

Keywords: Cyberbullying, Deep learning, Bullying detection, Natural Language Processing (NLP).

1. INTRODUCTION

Modern technology relies on the internet for communication, which may encourage undesirable conduct. Such disruptive and harmful behavior includes cyberbullying. Research suggests cyberbullying is to online or digital bullying. It can occur on cellphones, desktops, and other digital platforms as text-based interactions, messages, comments, forum postings, or image sharing. Cyberbullying is the repeated broadcast of cruel or offensive content on social media to injure or distress others [1]. It is a new type of bullying that differs from traditional harassment. Cyberbullying is not limited by time or location, and its anonymity can reach a wider audience and result in more serious abuse. Cyberbullying, especially among younger generations, has increased due to internet connectivity and social media platforms like Facebook, Instagram, and Twitter. This aggressive behavior deeply impacts victims' mental health and lives and can inspire other teens or group members [2]. Text messages, instant chats, social media, and online gaming can all be used for cyberbullying. Statisticbrain.com says Facebook is the most common cyberbullying site [3].

The most common cyberbullying media:

- Author-to-recipient email.
- Instant messaging online.
- Live online chat rooms with like-minded strangers.
- Short phone texts.
- Social networks unite people with similar backgrounds, interests, and connections.
- Personal, corporate, and government sites.

Cyberbullying is varied [4][5]:

- Flooding: Bullies repeat dumb posts to silence victims.
- Masquerade: Bullies imitate victims.
- Bullies incite hatred in victims as individuals or groups.
- Trolling posts opposing comments to provoke conflict.

- Posting hateful comments is illegal.
- Bullies humiliate victims.
- Bullies lie to ruin relationships and reputations.
- Bullies post embarrassing content online. Cyberbullying is degrading. The expedition may involve the bully and victim.
- Online group exclusion proposed. Teen cyberbullying is rising.
- Giving false information to harm reputation or friendships is impersonation.
- Trickery: Post embarrassing or sensitive information online.
- Cyberstalking involves repeated harassment, threats, or anxiety.

Although there is extensive research on cyberbullying, the majority of survey articles focused on certain topics and left out others. A comprehensive investigation into the use of websites for this type of criminal activity is still limited, so the current study will review recent publications spanning the last four years between 2020 and 2024. Ten separate sections make up the paper's structure. The first section introduces the topic. A survey of the literature makes up the second section. The effect of social networks on cyberbullying is discussed in the third section. The fourth section provides a comparison of instances of cyberbullying on social networks and methods for identifying them. Comparisons of bullying intelligence technologies for cyberbullying in social networks are provided in the fifth section. The sixth section covers the study's findings and its prospects. Preventive measures and cyberbullying protection systems make up the sixth section. The tenth section summarizes the paper's findings and offers recommendations for further research in the field of social networks. The ninth section provides data set availability.

2. A SURVEY OF THE LITERATURE

This section provides a succinct review of prior research on identifying cyberbullying in social networks. The main goal is to summarize the previous research, highlighting the most important works that fall under the purview of the literature review. S. M. Sh. Sha, Mahmuda K., and others [6] offer a system for evaluating Twitter data to detect terrorist attacks using machine learning. The model employs ternary search to give weights to previously defined keywords and Aho-Corasick automata for pattern matching. The model categorizes tweets into three groups: Terror Attack, Severe Terror Attack, and Normal Data. This model can detect if a terrorist incident has happened by using KNN and SVM classifiers. Using machine learning. B. Y. AlHarbi, M. S. AlHarbi, and others [7] suggest using Logistic Regression (LR) and Ridge Regression (RR), which are excellent at analyzing sentiment in English, to find cyberbullying on Arabic social media. By describing the potential of cyberbullying detection algorithms with an Arabic focus, this study addresses a knowledge vacuum in the existing literature. In this study, Sh. T.T. and B. R. Jeetha [8] employ Twitter models based on deep learning and sentiment analysis to identify instances of cyberbullying. An HRecRCNN is used for feature classification after the Modified Fruit Fly Algorithm (MFFA) is used for feature selection. In terms of accuracy, precision, recall, F-measure, and time complexity, the study found that the suggested framework performed better than the usual methods. A text-based machine learning method for predicting cybercrimes on Twitter is described by Sheila M. M., Christian M. D. S., and others in their publication [9]. In light of the growing epidemic of cyberbullying and related dangers, this study emphasizes social media prediction and prevention. The framework trains many machine learning models, such as Naïve Bayes, Decision Tree, Random Forest, SVM, and K-Nearest Neighbor. It also cleans tweets and uses TF-IDF to pull out features. Random Forest Classifier and SVM predicted cybercrime with the highest accuracy, reaching 93%. Cybercrime sentiment and trending phrases on Twitter are tracked by social media monitoring and law enforcement. Karan Shah, etc [10] introduced Hinglish cyberbullying detection that uses machine learning and natural language processing. It uses feature extraction methods like TF-IDF and strong machine learning classifiers to fix problems with biased datasets and keyword matching. The researchers can predict cyberbullying in real-time by analyzing tweets, WhatsApp conversations, and YouTube comments. B. Abd M., Jemal A., and others [11] introduce a hybrid deep-learning approach for detecting cyberbullying on Twitter. The model uses an improved Dolphin Echolocation Algorithm (DEA) and Elman Recurrent Neural Networks (RNN) to find the best RNN parameters, improve performance, and cut down on training time. With an accuracy of 90.45%, precision of 89.52%, recall of 88.98%, and F1-score of 89.25% on a Twitter dataset consisting of

DOI: 10.31695/IJASRE.2025.3.3

10,000 tweets, the model surpassed previous deep learning and machine learning models. N. Ashraf, D. Mahmood, and others [12] propose using NLP and deep learning to detect illegal conduct on social media. It targets aggressive, abusive, and antisocial conduct to identify dangers. The system categorizes activities using real-time social media data to help cybersecurity agencies prevent crime. The authors Ali A. J, and Ahmed H. A of this study [13] , present casual and unstructured Arabic social media texts, which they classify using deep learning multi-task models. Topic classification, sentiment analysis, sarcasm identification, and emotion categorization are all parts of the research process. We recommend a deep learning model that combines CNN with BiLSTM. The model is equipped with 9 million Arabic tweets that span nine categories—art, literature, politics, sport, economy, religion, science, health, and tech—and is powered by ArSarcasm-v2 and ArSAS, which enable the detection of sarcasm, sentiment, and emotion. In the simulation, the topic categorization accuracy was 97.58%, which means it was correct 97% of the time. The F1 score for sarcasm detection was 95%, which means it was correct 86% of the time. The F1 score for sentiment analysis was 86%, which means it was correct 81.6% of the time, and the F1 score for emotion recognition was 82%. Zainab Kh. A., Abbas M. B., and others [14] introduce a CNN-LSTM model and XGBoost for Arabic crime tweet detection and classification. The system uses Aho-Corasick's smart vocabulary to detect unlawful activities using keywords and contextual data. While XGBoost got a perfect score of 100 on 18,493 tweets, the hybrid deep learning model got 99.75%. This structure assists law enforcement in detecting Arabic criminal tweets. The authors S. Tofayet and N. Jahan [15] detail an ML system that uses optical character recognition and natural language processing to identify instances of cyberbullying in screenshots and photographs shared on social media. The proposed method finds instances of bullying or non-bullying language in images. Eight ML classifiers employ TF-IDF and BuW for feature extraction. The accuracy rate is 96% when using logistic regression with linear SVC. One framework was made by Zainab Kh. A., Abbas M. B., and others[16] to use Convolutional Neural Networks (CNNs) to find Arabic crime tweets. The framework used five hyperparameter optimization methods: SO-Roulette Wheel, SO-Tournament, SO-Linear Rank, SO-Exponential Rank, and Grey Wolf Optimizer. Improving CNN's ability to recognize tweets about crimes requires tweaks to embedding dimensions, dense units, learning rate, and batch size. The most accurate optimization approach, SO Tournament, has a 99.60% success rate. Silvia S, T Islam et al [17], The study proposes an NLP-based model using machine learning and deep learning algorithms to detect and classify Bengali comments on social media, specifically cyberbullying comments. The model uses a dataset of 56308 Bengali comments, including categories like not bully, trolls, sexual, and threats. The Recurrent Neural Network achieves the best accuracy of 86%, promoting morality. Israt T and V. Nunavath [18], proposed a hybrid deep learning strategy employing text and image data for social media cyberbullying detection and classification. Late fusion of transformer-based models (RoBERTa, BERT, DistilBERT) for text analysis and vision-based models (ResNet, CNN, ViT) for picture analysis is proposed. With 99.20% and 0.992 on public and 96.10% and 0.959 on private datasets, the RoBERTa+ViT model exceeds previous hybrid models in accuracy and F1-score. Khalid S., M.I. Khan1, and others [19] introduce Bully Filter Net, a deep learning framework for Bengali text cyberbullying detection and classification. The method addresses OOV and contextual feature extraction in low-resource languages. The study detects Bengali cyberbullying more accurately using transformer-based models like Bangla BERT. The model's 88.04% accuracy beats transformer-based techniques. Researchers N. GS, A. Shenoyy, and colleagues set out [20]to identify instances of cyberbullying in both English and Hinglish on social media. The study employs machine learning for classification and uses natural language processing for text preprocessing. Utilizing TF-IDF and Count Vectorizer for feature extraction, we examined comments made in real-time on Twitter, WhatsApp, and YouTube. The model's high accuracy makes it a useful tool for preventing cyberbullying. Zainab Kh. A., Abbas M. B., and others [21] use Snake Optimizer to fine-tune Convolutional Neural Networks (CNNs) to detect crime-related tweets. Optimizing CNN performance scores increases crime-related tweet identification. This study shows that the suggested framework works on a bespoke tweet dataset and outperforms existing optimization methods. To identify cyberbullying on social media, W. Tapaopong, A. Ch., and colleagues [22] investigate Transformer models. To categorize the severity of cyberbullying, this work employs transfer learning and fine-tuning on five Transformer models: BERT, RoBERTa, ALBERT, DistilBERT, and ConvBERT. When it comes to 5-class categorization tasks, DistilBERT excels with an accuracy of 94.36% and the best precision and recall. To identify cyberbullying on social media, Aigerim et al. [23] suggest a mixed-model deep learning system that

DOI: 10.31695/IJASRE.2025.3.3

combines convolutional neural networks (CNNs) with long short-term memory (LSTMs). This approach utilizes the hierarchical feature extraction of CNNs and the long-term text dependencies of LSTMs. While showcasing AI-driven solutions for digital safety, the paper reveals data asymmetry, overfitting, and ethical concerns. Zainab Kh. A., Abbas M. B., and colleagues [24] analyzed and predicted Arabic crime tweets. Using natural language processing and a vocabulary generated by a genetic algorithm, it screens tweets for criminal activity. The three machine learning algorithms that make up the prediction model are Decision Tree (DT), Random Forest (RF), and Logistic Regression (LR). DT and LR both achieve metrics of 97% accuracy, but LR comes in at 94.43%. The strategy prioritizes timeliness and clever filtering to improve prediction accuracy. T. N. Prabu and colleagues [25] describe a hybrid deep learning model that uses CNN and RF to detect social media cyberbullying. They demonstrate how CNN's feature extraction and RF's classification power help ProTect outperform traditional and deep learning models. The proposed approach outperforms existing methods on Instagram and Twitter real-time datasets in accuracy and speed.

3. COMPARISON STUDY OF CYBERBULLYING INTELLIGENCE TOOLS:

In this section, Table 1 (Comparison of Intelligence Tools in Various Cyberbullying) will explain the comparison among previous studies.

Table 1. Comparison of Intelligence Tools in Various Cyberbullying

No. Ref	The study's goal	Method Employed	Type of Dataset	Extracting and Selecting Features	Best Result
[6]	Utilizing powerful algorithms for pattern matching, keyword classification, and detecting terrorist attacks	Aho-Corasick Automaton and Ternary Search - Assign weights to tweets. - Find matches based on keywords. - Determines time intervals and scores. - Categorizes tweets using KNN and SVM.	Create terrorism tweet datasets: I: 1,000+ tweets II: 250 tweets	Terror Attack Score Calculation Predefined keyword-weighted scores. Terror Attack, Severe Terror Attack, Normal Data.	Dataset I KNN and SVM Accuracy - KNN: 97.1% - SVM: 97.7%
[7]	- Detecting Arabic Cyberbullying with Ridge Regression and Logistic Regression - Arabic social media RR/LR evaluation. -Comparing RR and LR to Arabic text categorization models.	- Analysis of Sentiment and Machine Learning. - It uses Logistic and Ridge Regression to detect cyberbullying in social media postings' tone and sentiment.	Overview of Arabic social media data: - Labeled tweets, Content labeled bullying or not.	TF-IDF: Text to Numerical Features - Converts text into numerical features. -No extra selection algorithms. - Preprocessing ensures high-quality features.	Ridge Regression (RR): 91.17% on a larger dataset. LR: Dataset size and content slightly affect accuracy.
[8]	To detect and stop Twitter cyberbullying, the study makes use of deep learning and sentiment analysis.	HRecRCNN combines RNN, residual networks, and Inception Networks.	The study utilizes 4,556 tweets.	TF-IDF, BoW, and N-gram analysis. Dictionary of slang words and emoticons. Optimized features are chosen using the Modified Fruit Fly Algorithm (MFFA).	With 95% accuracy, HRecRCNN outperformed other models.
[9]	- Twitter Criminality Detection Machine Learning - Detects cyberbullying and threats. - Utilizes textual analysis.	Machine Learning Text Classification - Classifier comparison. - Evaluation of MNB, Decision Tree, Random Forest, Logistic Regression, SVM, KNN.	The official Twitter API is used to extract the dataset.	Text Conversion Using TF-IDF - Converts text to numerical features suitable for classification.	The LR algorithm 93%, and the RF algorithm 93%.

DOI: 10.31695/IJASRE.2025.3.3

[10]	- Aims to detect abusive message content in real time.	- Enhanced precision and decreased intricacy. Multiple machine learning classifiers are evaluated.: - Support Vector Machine (SVM) - Random Forest - Logistic Regression - K-Nearest Neighbors (KNN) - Decision Tree - Bagging Classifier - Stochastic Gradient Descent (SGD) - Adaboost Multinomial Naïve Bayes	Analysis of Hinglish Datasets and Real-time Tweets - Text mining technologies for Twitter were employed. - There are 18,307 in the Hinglish dataset. - Binary labels: 0 (non-hazardous) and -1 (toxic).	A Comparative Study of Count Vectorizer and TF-IDF	- Performance of Random Forest and Linear SVC Random Forest: F1-score (97.2%), best accuracy (97.1%). - Linear SVC/SGD is faster and more precise.
[11]	Effective Model for Detecting Cyberbullying on Twitter. Resolving the shortcomings of current models. Managing dynamic data and brief texts.	Modifies weights and biases to increase RNN accuracy and convergence rate. Hybrid Deep Learning is achieved by combining Elman RNN sequence modeling with DEA optimization. The RNN, SVM, MNB, RF, and Bi-LSTM models are compared. - Short-text classification is enhanced by DEA-RNN convergence optimization.	Twitter API gathered 10,000 manually annotated tweets using cyberbullying terms. 6,508 non-cyberbullying tweets, 3,492 cyberbullying. - SMOTE achieved class balance.	Text representation with Word2Vec and TF-IDF. The Information Gain (IG) technique is used to select features.	- DEA-RNN Model Results Best results in Scenario 3 (90:10 train-test split). 90.45% accuracy,89.52% precision,88.98% recall, F1 is 89.25%., 90.94% specific - Beat Bi-LSTM, RNN, SVM, MNB, RF.
[12]	An automated approach for detecting illegal activity; hybrid NLP feature extraction and classification. Robust conduct can be detected by social media profiles.	MLP Detects Aggressive Behaviors on Social Media - Employs GloVe, BoW, and TF-IDF embeddings. Criminal intent is identified.	Cyber-Troll Dataset : - 20,001 tweets. - 12,179 non-aggressive tweets, 7,822 hostile ones. - Human labeling and keyword collection used.	- BoW, TF-IDF, and GloVe are used for word representation. - Captures contextual information using N-gram combinations.	- TF-IDF + Bigram + Unigram: F1 Score: 87% - Bigram + Trigram + BoW + TF-IDF: F1 Score: 88%
[13]	Hybrid Deep Learning Model for Arabic Social Media Text Classification Deals with casual Arabic text problems.	CNN-BiLSTM Hybrid Model for Classification of Multiple Tasks - Uses CNN to record local patterns. Uses BiLSTM to capture long-term dependencies.	- Nine themes were identified from 9 million Arabic tweets. - Arabic text with six emotions marked. - The dataset ArSarcasm-v2 for and against the classification of sarcasm. - Sentiment datasets: Twitter, ArSarcasm-v2, SS2030, and ArSAS.	Word Embedding Overview - Word2Vec (skip-gram) using pre-trained models. - Length of 100 vectors.	-Top Results: 97.58% Subject Classification and 97% Sarcasm Detection

DOI: 10.31695/IJASRE.2025.3.3

[14]	Utilizing a hybrid approach to deep-machine learning, Developing a Crime Detection System for Arabic Tweets. Detects criminal content instantly.	- Intelligent Dictionary Creation Aho-Corasick algorithm was applied to tweets about and unrelated to crimes. -XGBoost and CNN-LSTM hybrid deep learning models were used.	- 18,493 Arabic tweets were gathered with Python tools. - Data was labeled with an intelligent dictionary. - Added 25 non-crime-related keywords and 38 crime-related keywords. Using Intelligent Dictionary	- Use TF-IDF	-Top Results: XGBoost: 100% accuracy, 99% precision, 99.63% recall, 99.36% F1-Score. Hybrid CNN-LSTM: 99.75% accuracy, 99.84% F1-Score, 99.82% MAP.
[15]	System for Detecting Cyberbullying - Suggests a smart, precise system that makes use of optical character recognition and machine learning to retrieve text.	Techniques for Detecting Cyberbullying - OCR: Takes text out of pictures or screenshots. The retrieved text is enumerated by the NLP pipeline. - TF-IDF feature extraction and Bag of Words. - Evaluation of the cyberbullying classifier: eight. AdaBoost, SGD, Linear SVC, Random Forest, Bagging Classifier, Decision Tree, Gradient Boosting, and Logistic Regression.	- 160,000 bullying and non-bullying instances were found by the study using data from YouTube, Wikipedia Talk pages, Twitter, and Kaggle.	-TF-IDF and the Bag of Words BoW: Basic word occurrence representation; TF-IDF: Indicates the word significance of a document.	Logistic Regression and Linear SVC: Highest Accuracy • 96% accuracy was attained.
[16]	Building a CNN Model for Arabic Crime Detection Creating a CNN model that is optimal. - Optimized with five hyperparameter methods: SO-Roulette Wheel, SO-Tournament, SO-Linear Rank, SO-Exponential Rank, GWO.	- Deep Learning Framework and Optimization. CNN Optimization is used to classify text., Finding optimal hyperparameters. Identifying embedding dimension, density, unit count, learning rate, and batch size.	-Two datasets used are: The first dataset is a Custom-built dataset Comprising 18,493 tweets and Includes crime-related and non-crime content and the other dataset is found online internet.	The process of feature extraction and selection Pre-trained models are used using Word Embeddings, and Snake Optimizer with Evolutionary Selection Operators is to refine feature relevance.	- Optimal results with SO-Tournament Optimization: 99.60% precision 99.44% precision, 99.44% recall, and F1-Score is 99.44%
[17]	- The study employs multiclass classification to identify and categorize cyberbullying in Bengali.	The study utilized various models, including LR, SVM, RF, NB, XGBoost, RNN, and ensemble approaches.	It uses 56,308 Bengali comments.	TF-IDF and Keras embeddings	RNN achieved the highest accuracy of 86%.
[18]	A meme-based cyberbullying detection approach employing text and image social media data is proposed using hybrid deep learning.	This work combines text- and image-processing models to generate hybrid architectures such as LSTM + ResNet, GRU + ResNet, BERT + ResNet, DISTILBERT + ResNet, ROBERTa + CNN, ROBERTa + ViT (Top Performer Model).	The study classifies cyberbullying using public and private datasets, including 7,854 word/image samples from Twitter, Reddit, Facebook, and Instagram, and 1k memes and comments.	The features that are recovered from both modalities are combined using the Late Fusion Mechanism.	The RoBERTa + ViT model excelled: Public dataset accuracy: 99.2%, private 96.1%. Public dataset F1-score: 0.992, private 0.959.

DOI: 10.31695/IJASRE.2025.3.3

[19]	Creating an Intelligent Bengali Cyberbullying Text Identification Model • Overcame traditional and sequential constraints. Uses transformer-based language models.	Hybrid Framework Implementation combines statistical models, deep learning, and transformers. - During training, context-aware transformer models are improved. The statistical models SVM, Libsvm, and SGD. Models for CNN, VDCNN, LSTM, and GRU deep learning. - The transformer models IndicBERT, XML-RoBERTa, mBERT, and BanglaBERT. The most authentic Transformer model is the BanglaBERT, which was created using Bengali design and a great deal of pre-training.	Bengali Text dataset: 34,422 bespoke corpus of tagged Bengali texts; 12.543 messages from bullies and 11,565 texts that are not bullies for testing and training. Manually gathered information from social media.	-Contextual: Transformer-based models such as BanglaBERT, mBERT, and XML-RoBERTa; non-contextual: GloVe, FastText, and Word2Vec.	BullyFilterNeT based on BanglaBERT: Highest Accuracy is 88.04% accuracy
[20]	- Developing a Cyberbullying Detection Model Categorizing cyberbullying messages in English and Hinglish.Real-time detection using NLP and machine learning.	Comparing eight algorithms to find the best model. The following models were evaluated: (SVM), LR , A RF , K-Nearest Neighbors (KNN) , Decision Tree , Bagging Classifier , SGD Classifier , and Adaboost	Twitter real-time tweets. YouTube and WhatsApp chats + comments. 12,307 English, 3,000 Hinglish rows.	TF-IDF and Count Vectorizer are compared.	Linear SVC Algorithm Overview - Earned highest accuracy and F1 score. - Algorithm: 96.1% precision, 96.1% recall, 95% accuracy. -Best results: excelled in English and Hinglish texts.
[21]	To effectively detect crime-related tweets, CNN models with Snake Optimizer for hyperparameter tuning are used. The goal is to accurately classify tweets as crime-related or not.	A Deep Learning Framework for Classifying Text in Tweets - A specially designed CNN was created for categorization and processing. - Hyperparameters are adjusted using the Snake Optimizer algorithm. Effectively adjusts CNN hyperparameters such as learning rate, kernel size, and activation functions.	-Two datasets used are The first dataset is a Custom-built dataset Comprising 18,493 tweets and Includes crime-related and non-crime content and the other dataset is found online internet.	- Word Embeddings: Textual data is represented in a dense vectorized format using pre-trained Word2Vec embeddings. - Feature Selection: By adjusting hyperparameters during training, the Snake Optimizer improves feature relevance.	Best Results: Accuracy: 99.20%, Precision: 98.91%, Recall: 99.15%, F1-Score: 99.03%

DOI: 10.31695/IJASRE.2025.3.3

[22]	. Detecting Cyberbullying using Transformer Models Used transfer learning and fine-tuning on five Transformer models. Finding the best multi-class cyberbullying severity model.	tested a variety of transformer models, including as ConvBERT, DistilBERT, ALBERT, RoBERTa, and BERT. o Transfer Learning: Applying previously learned models to a dataset on cyberbullying.	Dataset for Cyberbullying Classification -Has 47,692 tweets in it. - Divided into six categories: Age, Gender, Religion, Ethnicity, Not Bullying, and Other Types.	- Use BERT's tokenizer for special tokens like [CLS] and [SEP]. - Convert sequences into numerical vectors for input preparation.	Performance of ConvBERT and DistilBERT - DistilBERT: 93.85% F1-Score, 93.91% recall, 93.83% precision, and 94.36% accuracy. - ConvBERT: 94.31% F1-Score, 93.79% recall, 93.75% precision, and 94.30% accuracy.
[23]	Detecting Cyberbullying using Hybrid Deep Learning - Creates precise, sensitive models. - Addresses social media language issues. -CNN-LSTM integration.	- For hierarchical text features, CNNs are used. - Uses LSTMs to account for contextual subtleties and long-term interdependence. To reduce overfitting, regularization strategies, and dropout layers are used.	- Data from several social media platforms is included in the Kaggle dataset. Bullying and non-bullying classes are used to categorize the data, which show different linguistic styles and situations.	Since the hybrid design automatically extracts appropriate features from the data, no explicit feature selection is discussed.	Best Outcomes of the Model: - 95.4% accuracy - Metrics for competitive recall and precision -The precise values are not specified.
[24]	-Uses machine learning to improve the prediction of Arabic crime tweets. -Uses an artificial dictionary to filter tweets based on illegal behavior.	Tweet Data Filtering and Sorting -Utilizes genetic algorithm for data filtering based on criminal behaviors. RF, DT, and with a timeline of the tweet	-Datasets have 16779 keywords for terrorism, drug smuggling, ISIS, bullying, intentional murder, robbery, and murder. In the Arabic language	-Uses TF-IDF for feature extraction, Using a genetic algorithm to select and build a dictionary of filter	-Data time DT + TF-ID performs the best with 97% accuracy.
[25]	To identify cyberbullying on social media, a scalable hybrid deep learning model has to be developed.	-Protect is a hybrid deep-learning network that uses CNN and RF to identify cyberbullying on social media. Optimization involves regularization and parameter adjustment to reduce overfitting and enhance generalization.	Instagram and Twitter Datasets - Instagram: 10,000 bullying/non-bullying comments. - Twitter: 100,000 crowdsourced tweets. - Both require high noise treatment. - Utilized SMOTE to alleviate class imbalance.	- CNN extracts linguistic features, sentiment polarity, offensive word frequency, syntactic structures. - RF uses reduced CNN feature space for reliable classification. -TF-IDF and GloVe Embeddings in Machine Learning - Utilized for models. - Provided neural network inputs.	Performance of the CNN-RF Model - Dataset from Instagram: - Accuracy: 98.81% - Dataset from Twitter: 99.71% precision - The CNN-RF model saved 3.4 seconds during training.

4. CONCLUSION

Using artificial intelligence, deep learning, and optimization algorithms as structural analysis devices, several recent studies on identifying cyberbullying and cybercrime on social media platforms have been analyzed. Several key research themes were evidenced in these studies, including a focus on cyberbullying.

A- Literature Review on Cyberbullying Detection

1- Tweet Classification and Text Analysis using Deep Learning

DOI: 10.31695/IJASRE.2025.3.3

According to previous research, DL ranks as the most accurate method for detecting cyberbullying using diverse models like:

I. CNN (Convolutional Neural Networks) for extracting deep text patterns.
II. Sequential Data with RNN and BiLSTM.
III. Transformer Models (such as BERT, as well as AraBERT) for accurate Arabic language analysis and context comprehension.

Examples of Top-Notch Models:
a- CNN-BiLSTM model used for cyberbullying detection for Twitter, attaining 97.58% accuracy for topic classification and 97% accuracy in sarcasm detection.
b- 99.60% accuracy achieving CNN-SO-tournament model using evaluation selection optimization to classify crime and cyberbullying tweets.

2- Enhancing Model Performance Using Optimization Algorithms:
Some studies used metaheuristic optimization algorithms to get better accuracy such as:

I. Grey Wolf Optimizer (GWO), improving CNN accuracy to 99.39% in crime-related tweet classification.
II. DEA: An algorithm of Dolphin Echolocation to Optimize RNN Network Performance, Improving the rate of cyberbullying identification, reaching 90.45%.
III. Whereas, Snake Optimizer (SO) which is learned in this paper for enhancement in the neural network hyperparameters helps in better classification between aggressive and cyberbullying.
IV. Evolutionary Selection Optimization with selection methods for Roulette Wheel, Tournament, Linear Rank, and Exponential Rank selection such as: to boost the CNN performance. Evolutionary Selection Optimization Hybrid Selection Evolutionary Optimization Selection cyberbullying detection.

3- Techniques and Their Effect on the Model Accuracy
Several feature engineering techniques are developed advanced, such as:

I. TF-IDF to find the most relevant words.

II. Word Embeddings (Word2Vec, GloVe, FastText) to understand semantic relations between words.

III. Identify whether the tweet has aggressive, bullying, or neutral sentiment.

For example, TF-IDF with Decision Tree (DT) captured 97% accuracy on cyberbullying classification which improved when used with Genetic Algorithm (GA)-based optimization.

4- A Comparison of Deep Learning and Traditional Machine Learning for Detection of Cyberbullying
In comparing DL models with classical ML algorithms like:

In particular, it outperforms algorithms such as RF, LR, and SVM whereas Results indicate that deep learning drastically outperforms traditional methods:

I. CNN-BiLSTM got 97.58% accordingly, better than SVM with 93%.

II. The highest accuracy was 97% obtained from the Decision Tree (DT), followed by Logistic Regression(94.43%).

III. A few studies reported 100% accuracy with XGBoost, significantly better than any traditional model.

DOI: 10.31695/IJASRE.2025.3.3

B- Major Issues in Cyberbullying Detection:

1- Arabic Tweets and Sarcasm: Context is Key:

Cyberbullying tweets can use indirect or sarcastic language that makes it more difficult to detect. We adopt advanced Transformer models e.g. AraBERT, and CAMeLBERT, which use Self-Attention Mechanisms to handle contextual information more effectively.

2- The Effect of Imbalanced Data on the Model Performance: Most of the above studies report that having imbalanced datasets — such that cyberbullying tweets are underrepresented — leads to degraded model performance.

Proposed Solution: To balance the dataset distribution, data augmentation techniques, like re-sampling and oversampling, can be applied to the dataset.

3- Challenges in Processing Arabic Dialects and Nonstandard Language.

– Examples of Text Data: Tweets contain contractions, colloquial, and regional Arabic dialects making it difficult to analyze the text. Suggested Solution: Train with diverse Arabic dialect datasets and deploy Arabic-specific NLP models like AraBERT for better performance.

4- Potential Security and Privacy Issues

Cyberbullying detection systems that maintain user privacy are required.

- Data Encryption and Anonymous Text Analysis Solution.

C- **Cyberbullying Detection: Best Models and Results**

As per this section in Table 2 Best Models and Results in Cyberbullying Detection

Table 2. Best Models and Performance in Cyberbullying Detection

Study	Model Used	Highest Accuracy
Hybrid CNN-LSTM with XGBoost (Cybercrime and Cyberbullying Classification)	CNN-LSTM + XGBoost	99.84% - 100%
DEA-RNN (Cyberbullying Detection on Twitter)	RNN + DEA Optimization	90.45%
CNN-BiLSTM (Cyberbullying Classification in Arabic Tweets)	CNN + BiLSTM	97.58%
Grey Wolf Optimizer with CNN	CNN + GWO	99.39%
Modified Fruit Fly Algorithm with HRecRCNN	HRecRCNN + MFFA	98%
CNN-SO-tournament (Optimized Cyberbullying Detection via Evolutionary Selection)	CNN-SO-tournament	99.60%

Results show that the integrated approach with AI and deep learning alongside hyperparameter optimization techniques like GWO, DEA, and SO improved the performance of cyberbullying detection significantly. In conclusion

1- Cyberbullying detection is most accurate using deep learning.

DOI: 10.31695/IJASRE.2025.3.3

2- It shows that employing models based on the GWO and DEA yields higher performance than conventional approaches.
3- The more advanced word embedding methods (Word2Vec, GloVe) work better to acquire the semantic meaning.
4- The biggest challenge is informal Arabic, sarcasm to which the Transformer Model (BERT, AraBERT) is the strongest method of choice.

ACKNOWLEDGMENT

The authors would like to thank AL_Mustansiriyah University (www.uomusiriyah.edu.iq), Baghdad-Iraq for its support in the present work.

REFERENCES

1. N. Yuvaraj *et al.*, "Automatic detection of cyberbullying using multi-feature based artificial intelligence with deep decision tree classification," *Comput. Electr. Eng.*, vol. 92, p. 107186, Jun. 2021, doi: 10.1016/j.compeleceng.2021.107186.
2. A. Wang and K. Potika, "Cyberbullying Classification based on Social Network Analysis," in *2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService)*, May 2021, pp. 87–95. doi: 10.31979/etd.9bn7-tq9h.
3. Y. Peled, "Cyberbullying and its influence on academic, social, and emotional development of undergraduate students," *Heliyon*, vol. 5, no. 3, p. e01393, Mar. 2019, doi: 10.1016/j.heliyon.2019.e01393.
4. I. Alanazi and J. Alves-foss, "Cyber Bullying and Machine Learning: A Survey," *International Journal if Computer Science and Information Security*, vol. 18, no. 10. pp. 1–8, 2020. doi: 10.5281/zenodo.4249340.
5. B. Haidar, M. Chamoun, and F. Yamout, "Cyberbullying Detection: A Survey on Multilingual Techniques," in *Proceedings - UKSim-AMSS 2016: 10th European Modelling Symposium on Computer Modelling and Simulation*, Nov. 2017, pp. 165–171. doi: 10.1109/EMS.2016.037.
6. Sarker, A. , Chakraborty, P. , Sha, S. , Khatun, M. , Hasan, M. and Banerjee, K. (2020) Improvised Technique for Analyzing Data and Detecting Terrorist Attack Using Machine Learning Approach Based on Twitter Data. *Journal of Computer and Communications*, **8**, 50-62. doi: 10.4236/jcc.2020.87005.
7. AlHarbi, Bedoor Y., Mashael S. AlHarbi, Nouf J. AlZahrani, Meshaiel M. Alsheail, and Dina M. Ibrahim. "Using machine learning algorithms for automatic cyber bullying detection in Arabic social media." *Journal of Information Technology Management* 12, no. 2 (2020): 123-130.
8. Sherly, T., and B. Jeetha. "Sentiment analysis and deep learning based cyber bullying detection in twitter dataset." *International Journal of Recent Technology and Engineering (IJRTE)* 10, no. 4 (2021): 15-25.
9. Matias, Sheila Marie M., Jefferson A. Costales, and M. Christian. "A Framework for Cybercrime Prediction on Twitter Tweets Using Text-Based Machine Learning Algorithm." In *2022 5th International Conference on Pattern Recognition and Artificial Intelligence (PRAI)*, pp. 235-240. IEEE, 2022.
10. Shah, Karan, Chaitanya Phadtare, and Keval Rajpara. "Cyber-bullying detection in hinglish languages using machine learning." *Int J Eng Res Technol* 11 (2022).
11. Murshed, Belal Abdullah Hezam, Jemal Abawajy, Suresha Mallappa, Mufeed Ahmed Naji Saif, and Hasib Daowd Esmail Al-Ariki. "DEA-RNN: A hybrid deep learning approach for cyberbullying detection in Twitter social media platform." *IEEE Access* 10 (2022): 25857-25871.
12. Ashraf, Noorulain, Danish Mahmood, Muath A. Obaidat, Ghufran Ahmed, and Adnan Akhunzada. "Criminal Behavior Identification Using Social Media Forensics." *Electronics* 11, no. 19 (2022): 3162.
13. Jalil, Ali A., and Ahmed H. Aliwy. "Classification of Arabic Social Media Texts Based on a Deep Learning Multi-Tasks Model." *Al-Bahir Journal for Engineering and Pure Sciences* 2, no. 2 (2023): 12.
14. Abdalrdha, Z.K., Al-Bakry, A.M., Farhan, A.K. (2023). A hybrid CNN-LSTM and XGBoost approach for crime detection in tweets using an intelligent dictionary. Revue d'Intelligence Artificielle, Vol. 37, No. 6, pp. 1651-1661. https://doi.org/10.18280/ria.370630

DOI: 10.31695/IJASRE.2025.3.3

15. Sultan, Tofayet, Nusrat Jahan, Ritu Basak, Mohammed Shaheen Alam Jony, and Rashidul Hasan Nabil. "Machine learning in cyberbullying detection from social-media image or screenshot with optical character recognition." *International Journal of Intelligent Systems and Applications* 15, no. 2 (2023): 1.
16. Abdalrdha, Zainab Khyioon, Abbas Mohsin Al-Bakry, and Alaa K. Farhan. "CNN Hyper-Parameter Optimizer based on Evolutionary Selection and GOW Approach for Crimes Tweet Detection." In *2023 16th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 569-574. IEEE, 2023.
17. Sifath, Silvia, Tania Islam, Md Erfan, Samrat Kumar Dey, MD Minhaj Ul Islam, Md Samsuddoha, and Tazizur Rahman. "Recurrent neural network based multiclass cyber bullying classification." *Natural Language Processing Journal* 9 (2024): 100111.
18. Tabassum, Israt, and Vimala Nunavath. "A Hybrid Deep Learning Approach for Multi-Class Cyberbullying Classification Using Multi-Modal Social Media Data." *Applied Sciences (2076-3417)* 14, no. 24 (2024).
19. Saifullah, Khalid, Muhammad Ibrahim Khan, Suhaima Jamal, and Iqbal H. Sarker. "Cyberbullying Text Identification based on Deep Learning and Transformer-based Language Models." *EAI Endorsed Transactions on Industrial Networks and Intelligent Systems* 11, no. 1 (2024): e5-e5.
20. Nikitha, G. S., Amritasri Shenoyy, K. Chaturya, and J. C. Latha. "Detection of cyberbullying using NLP and machine learning in social networks for bi-language." *International Journal of Scientific Research & Engineering Trends* 10, no. 1 (2024).
21. Abdalrdha, Zainab Khyioon, Abbas Mohsin Al-Bakry, and Alaa K. Farhan. "Crimes Tweet Detection Based on CNN Hyperparameter Optimization Using Snake Optimizer." In *National Conference on New Trends in Information and Communications Technology Applications*, pp. 207-222. Cham: Springer Nature Switzerland, 2023.
22. Tapaopong, Wachiraporn, Atiphan Charoenphon, Jakkapong Raksasri, and Taweesak Samanchuen. "Enhancing cyberbullying detection on social media using transformer models." In *2024 5th Technology Innovation Management and Engineering Science International Conference (TIMES-iCON)*, pp. 1-5. IEEE, 2024.
23. Altayeva, Aigerim, Rustam Abdrakhmanov, Aigerim Toktarova, and Abdimukhan Tolep. "Cyberbullying Detection on Social Networks Using a Hybrid Deep Learning Architecture Based on Convolutional and Recurrent Models." *International Journal of Advanced Computer Science & Applications* 15, no. 10 (2024).
24. Abdalrdha, Zainab Khyioon, Abbas Mohsin Al-Bakry, and Alaa K. Farhan. "Arabic Crime Tweet Filtering and Prediction Using Machine Learning." *Iraqi Journal for Computers and Informatics* 50, no. 1 (2024): 73-85.
25. Nitya Harshitha, T., M. Prabu, E. Suganya, S. Sountharrajan, Durga Prasad Bavirisetti, Navya Gadde, and Lakshmi Sahithi Uppu. "ProTect: a hybrid deep learning model for proactive detection of cyberbullying on social media." *Frontiers in artificial intelligence* 7 (2024): 1269366.

DOI: 10.31695/IJASRE.2025.3.3