Journal of Information Technology Management
ISSN-OnLine: 2980-7972
University of Tehran

Artificial Intelligence-Driven Cyberbullying Detection: A Survey of Current Techniques

Kholood Alfaleh*

*Corresponding author, Department of Information Technology, College of Computer Qassim University Buraydah 51452, Saudi Arabia. E-mail: 441212457@qu.edu.sa

Abdulatif Alabdultif

Department of Computer Science, College of Computer, Qassim University, Buraydah 52571, Saudi Arabia. E-mail: ab.alabdulatif@qu.edu.sa

Suliman Aladhadh

Department of Information Technology, College of Computer Qassim University Buraydah 51452, Saudi Arabia. E-mail: s.aladhadh@qu.edu.sa

Abstract

Cyberbullying involves using hurtful or offensive language that goes against basic rules of respect and politeness. It harms the online environment and can negatively affect people by causing harassment, discrimination, or emotional pain. To combat this, it is crucial to develop automated methods for detecting and preventing the dissemination of such content. Deep learning, a branch of artificial intelligence, leverages neural networks to learn from data and perform complex tasks, effectively capturing semantic and grammatical nuances to differentiate between abusive and non-abusive language. This survey paper reviews current techniques and advancements in deep learning-based approaches for detecting cyberbullying content on online platforms, aiming to provide a comprehensive understanding of existing methodologies and identify potential avenues for future research to mitigate the spread and impact of such behaviors on the internet.

Keywords: Cyberbullying, cyber-harassment, Deep learning, Social Media

Journal of Information Technology Management, 2024, Vol. 16, Issue 4, pp. 38-63
Published by the University of Tehran, College of Management
doi: https://doi.org/10.22059/jitm.2024.99050
Article Type: Research Paper

Received: June 12, 2024
Received in revised form: July 25, 2024
Accepted: September 13, 2024
Published online: October 16, 2024

Introduction

The rapid proliferation of online communication platforms, ranging from social media networks to forums and chat applications, has fundamentally transformed how individuals interact, disseminate information, and engage in social discourse. However, it also brings to the forefront a pervasive and concerning issue—the prevalence of abusive content, encompassing language that transgresses the boundaries of civility and respect, potentially causing harm, distress, and emotional anguish to its recipients. Such content, which may include cyberbullying, harassment, and discriminatory remarks, poses a significant challenge to maintaining a healthy and inclusive online environment (Hasan et al., 2023).

Addressing the detrimental impact of abusive content necessitates the development of effective automated detection methods. In recent years, the emergence of deep learning, a subset of artificial intelligence, has offered a promising avenue for automating the detection process. This research aims to harness the potential of deep learning in combating cyberbullying and abusive content, paving the way for a more secure and respectful digital ecosystem (Daniel et al., 2023).

DL models capture semantic nuances, contextual cues, and syntactic structures, making them well-suited for tasks like sentiment analysis, language translation, and critical, abusive content detection (Alabdulwahab et al., 2023). The versatility of DL architectures in handling textual data is particularly advantageous for identifying abusive content. Techniques such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs), along with more advanced variants like long short-term memory (LSTM) networks, have demonstrated exceptional proficiency in extracting meaningful features from text (Hasan et al., 2023). However, domain adaptation, or the ability of a model to generalize across different online platforms and communication channels, presents a critical consideration (Ahmed et al., 2023).

Problem Statement and Contribution

In the digital age, cyberbullying is becoming more and more common as individuals harass others via social media and other modern technologies.

Research indicates that this kind of bullying, whether it occurs in children, adolescents, or adults, may have serious health repercussions. However, further investigation is needed to properly address this issue. Some suggestions include focusing on encouraging positive online conduct and establishing policies and guidelines to deal with cyberbullying in online groups and schools. Artificial intelligence (AI) has been of assistance and contributed to the resolution of the issue. These behaviors may be identified by words, pictures, voice, or even video.

The Contribution for This research will conduct an in-depth assessment of current mechanisms and AI techniques used for combatting abusive expressions that go undetected to recipients, to comprehend their capabilities and devise effective technical solutions. This study seeks to investigate and assess all available solutions to cyberbullying detection. Our research centers around applying and assessing deep learning/machine learning techniques together as possible measures. At our data preprocessing methods evaluation lab, we will test various data preprocessing strategies such as DL for linguistic transformations to increase the accuracy of developed models.

Background

This section provides an overview of key concepts related to cyberbullying, artificial intelligence (AI), and machine learning (ML). It explains the nature of cyberbullying and the role of cybersecurity awareness, followed by an introduction to AI and its applications in cybersecurity, and finally, a look at ML techniques used for detecting abusive behaviors online.

Cyberbullying

Understanding Bullying:

In the context of an imbalance of power, bullying is described as aggressive conduct that happens frequently against another person. Bullying is the negative behavior that occurs when someone in a position of authority engages in hostile behavior against another individual. One of the most common forms of bullying is among peers, and bullies often have more power and influence than the people they bully (Ahmed et al., 2023).

Understanding Cyberbullying

Cyberbullying on social networking sites is a unique kind of communication that is distinct from face-to-face or other sorts of cyberbullying on alternative digital communication platforms, such as email, telephone, and text messaging (Craig et al., 2007).

As stated in (McFarland et al., 2015), the following list outlines several manifestations of cyberbullying:

- Flaming: sending nasty, rude, and vulgar messages to someone or a group of individuals over the internet to upset them,
- Harassment: Sending someone negative messages repeatedly,
- Cyberstalking: Intimidation characterized by extreme fear or explicit threats of damage,
- Denigration: Spreading or disseminating defamatory, false, or harmful statements about someone to others,
- Masquerade: pretending to be someone else and releasing or distributing something that damages their reputation or puts them in danger,
- Outing and trickery: Spreading or distributing information about a person that contains sensitive, private, or embarrassing facts,

- Exclusion: Systematic and intentional measures designed to isolate a person from an online community,
- Impersonation: Impersonating the victim and using online communication to spread negative or inaccurate information to others as if it came from the victim, and
- Sexting: Sharing bad graphic images of someone without their express consent.

Cybersecurity Awareness in Preventing Cyberbullying

Implementing the ability to authenticate social media and similar accounts using an identification number would effectively mitigate cyberbullying, leading to a decrease in cyberbullying incidents (Willard et al., 2007). Combatting cyberbullying among adults may be more successful by promoting legal awareness and imposing stricter criminal sanctions (Eriş et al., 2022).

Artificial Intelligence

Understanding Artificial Intelligence

John McCarthy coined the phrase "artificial intelligence" (AI) in 1956 (Al Shamsi et al., 2019). The software can now more efficiently, effectively, and cheaply mimic human thinking, reasoning, planning, communication, and perception, all thanks to advancements in AI technology (Židová et al., 2021).

Cyber Security Approaches to AI

AI techniques such as Machine Learning (ML) and DL can be used to address cybersecurity issues like intrusion detection and prevention systems. These techniques include NLP, KRR, and rule-based Expert Systems (ES) modeling (Shapiro et al., 1992).

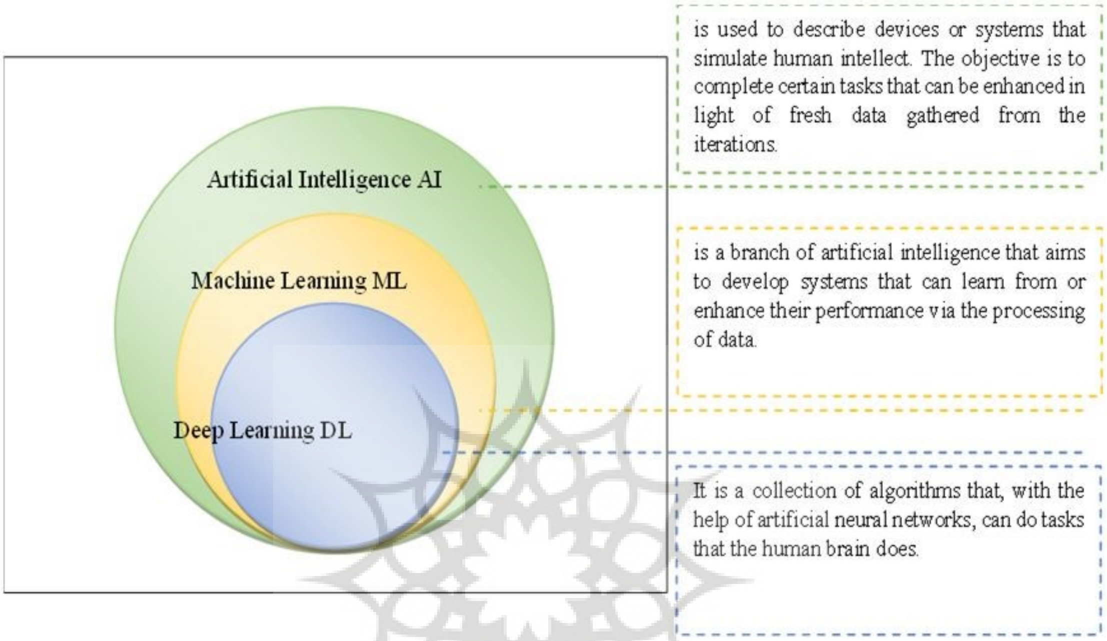
Machine Learning& Deep Learning

The capacity of a computer to educate itself on how to make decisions by utilizing the data and experiences available to it is what is meant by the term "machine learning" (ML) (PK et al., 1984). ML has several applications, with data mining (DM) being the most prominent. ML algorithms are classified into the following categories:

- Supervised Learning algorithms: These algorithms rely on labeled training data to determine outcomes (Sarker et al., 2021).
- Unsupervised Learning algorithms: These algorithms work with unlabeled data, unlike supervised learning, which depends on historical labeling (Buczak et al., 2015).
- The most popular ML algorithms used for cyberbullying and abusive language detection are:
- Naive Bayes: This algorithm uses Bayes' theorem in conjunction with rather naïve assumptions (Osisanwo et al., 2017; Khairy et al., 2021).

- Support Vector Machine (SVM): A binary classifier that presumes data samples are clearly differentiated, SVMs aim to find the best possible hyperplane to maximize the class margin (Kufel et al., 2023).

- Deep Learning (DL): DL approaches have been widely used in DM and text classification to forecast and categorize occurrences (Mahesh et al., 2020).



is used to describe devices or systems that simulate human intellect. The objective is to complete certain tasks that can be enhanced in light of fresh data gathered from the iterations.

is a branch of artificial intelligence that aims to develop systems that can learn from or enhance their performance via the processing of data.

It is a collection of algorithms that, with the help of artificial neural networks, can do tasks that the human brain does.

Figure 1. Understanding the Different Between AI, ML, and DL

Overview of Cyberbullying Detection Architecture

This section provides an overview of the typical architecture used in cyberbullying detection, outlining the key components and processes involved in developing effective detection systems. The architecture generally involves a systematic and multi-step process aimed at developing, testing, and validating the developed DL approach for identifying instances of cyberbullying. The architecture includes the following key stages:

- **Data Collection and Preparation:** The dataset was pre-processed to ensure accuracy, anonymity, and uniformity.

- **Feature Extraction and Representation:** Natural language processing techniques such as tokenization, stemming, and vectorization can be used to extract features relevant to textual content. It can also explore metadata and other multimedia elements if they are applicable.

- **Model Selection and Architecture Design:** Choose a suitable deep learning architecture for text classification tasks, such as CNNs or RNNs, and fine-tune or adapt pre-trained models for improved performance.

- **Training and Validation:** Use cross-validation for generalization and appropriate evaluation metrics to monitor performance. Divide data into training sets and validation sets.
- **Model Evaluation and Testing:** Assess the model's applicability in real life by evaluating its performance on a test set independently, analyzing its strengths, weaknesses, and areas of improvement.
- **Optimization and Deployment:** The model can be refined to achieve optimal performance by considering resource requirements and computational efficiency. It is then implemented in a real-world setting and integrated with existing communication channels or online platforms.
- **Documentation and Reporting:** Document the research in a transparent manner, including the dataset, model architecture, and parameters of training, as well as evaluation results.

By following this comprehensive methodology, the study aims to develop an effective and reliable DL approach for detecting cyberbullying content across digital communication platforms.

This diagram summarizes all these points of the method by which cyberbullying can be detected In Figure 2.

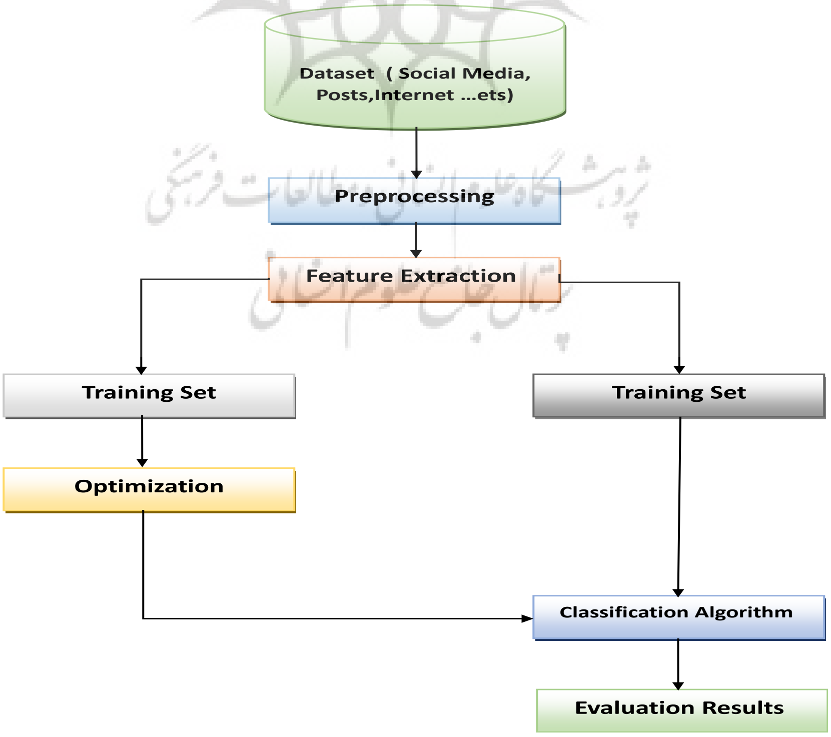


Figure 2. The proposed model for the development and improvement of cyberbullying detection

An overview of Artificial Intelligence adaptation in cyberbullying

Deep Learning Techniques

- **Transfer Learning**: A technique that allows models trained on large datasets to be applied to specific tasks (Shalev-Shwartz et al., 2014).
- **Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM)**: These technologies are known for their ability to process sequential data, such as text (Al-zaqebah et al., 2023; Zheng et al., 2018).
- **Convolutional Neural Networks (CNN)**: This technique, ideal for image analysis, can also be applied to textual data by using filters to extract keywords indicative of bullying (Mohaouchane et al., 2019; Zheng et al., 2018; AlHarbi et al., 2019; AlHarbi et al., 2020).

Machine Learning Techniques

- **Traditional ML:**
  - SVM: A powerful technology used in classifying texts with high accuracy (Akhter et al., 2021), (AlFarah et al., 2022).
  - Decision Trees: A technique that builds a model through which decisions can be made based on a series of questions or tests related to the characteristics of texts (Akhter et al., 2021).
  - Naive Bayes: A technique that uses the probabilities of words within texts to classify them based on the assumption that the presence of certain words can increase the likelihood that the text is bullying (Akhter et al., 2021).
- Feature Engineering:
  - Text Features: Extraction includes textual features such as linguistic density, emotionality, and frequency of use of certain words.
  - User Behavior: It reveals patterns such as the frequency of messages, the time of posting, and the nature of interaction with other users (AlFarah et al., 2022).

Hybrid Approaches

- **Combining ML and DL**: This approach integrates the strengths of both machine learning and deep learning, leveraging their complementary capabilities (Akhter et al., 2021; Haidar et al., 2017; Alduailaj et al., 2023).

The previous explanation can be summarized in Figure 3.

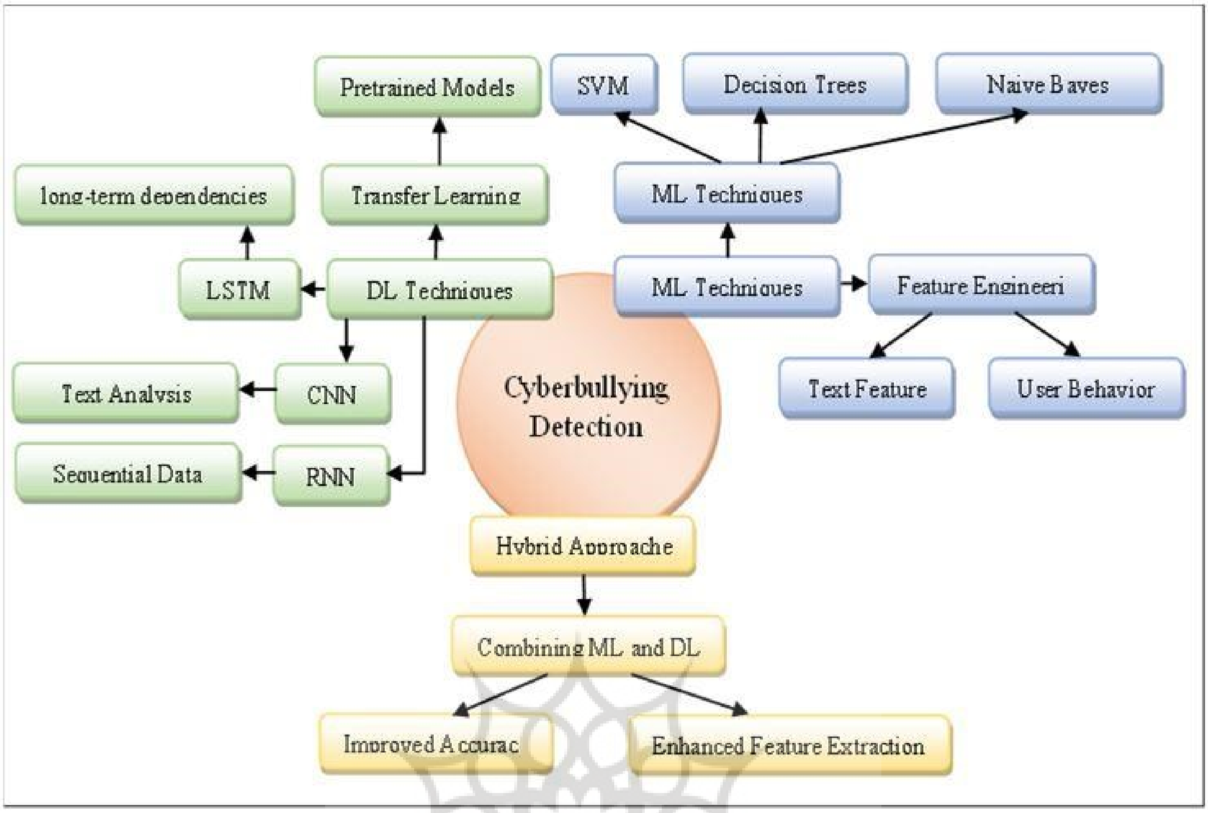


Figure 3. Deep Learning and Machine Learning Methods to Detect Cyberbullying

Historical Approaches to Cyberbullying Detection: Pre-AI Techniques

As we discussed how artificial intelligence (AI) tools can detect cyberbullying, it's essential to recognize that traditional methods were employed before this technology became widely adopted. Conventional techniques were often utilized at that time to detect cyberbullying and offensive language using whatever tools were available at that time. Here, we will go over some of these traditional ways used before AI tools became widely adopted as well as some that can still detect cyberbullying today.

Before artificial intelligence's emergence, several traditional techniques were used to detect cyberbullying using available technology. Such methods included:

- Offensive Word Lists: Lists of offensive words and phrases were developed and utilized as text search tools that detected such undesirable terms (Ahmed et al., 2021).
- Patterns and Specific Models: Systems were devised to detect specific behavioral patterns that might suggest bullying, such as repeated offensive remarks or threats (Hoque et al., 2023).

Filtering Contents: Filtering systems were used to screen messages and comments based on specific rules, such as checking for offensive words or threats in messages and comments (Sarker et al., 2022).

- Behavioral Indicators: Tools were utilized to detect unusual behavioral signals, such as repeated offensive messages or attempts at user manipulation (Reynolds et al., 2011).

Methods like these were necessary in early attempts at cyberbullying detection through technology, yet their abilities are now dwarfed by modern artificial intelligence solutions.

Literature review

Through this study, we will review previous studies about techniques that try to prevent cyberbullying. Techniques could involve employing algorithms like SVM and Naïve.

Bayes classifiers or DL models such as CNN. For this reason, many algorithms, tools, and methods have been developed that are useful in preventing cyberbullying by detecting, recording, and grouping them into lists and libraries containing offensive words, which were collected from several social networking sites such as YouTube, X platform, Instagram, and others. Some of the techniques discussed in some scientific research will be mentioned.

Deep Learning Techniques

We review recent advancements in deep learning techniques applied to cyberbullying detection, highlighting key studies and their innovative approaches to improving classification performance.

In 2019, Mohaouchane et al. (2019) explored the effectiveness of neural networks such as CNN and Bi-LSTM for cyberbullying detection. They also applied Bayesian optimization to refine hyperparameters and improve model performance through cross-validation. Similarly, in 2019, Emon et al. (2019) focused on Bengali abusive text, using ML/DL algorithms like LinearSVC with LSTM and RNN. They also implemented language-specific stemming to enhance the accuracy of text classifiers.

In 2021, Husain et al. (2021) developed the SalamNET model based on a Bi-GRU architecture. The model achieved an F1 macro score of 0.83 and provided valuable insights into detecting offensive language on Arabic social media, as part of the SemEval-2020 shared task. That same year, Kompally et al. (2021) proposed a decentralized DL approach named MaLang, featuring a novel two-level system for classifying and analyzing abusive textual content. The system used a CASE model for initial classification, followed by a cloud module called KIPP to assess the probability of toxic content.

In 2022, Roy et al. (2022) developed image-based models for cyberbullying detection using a 2D CNN and investigated transfer learning models like VGG16 and InceptionV3. The models proved both accurate and resource-efficient, with VGG16 achieving 86% accuracy and InceptionV3 89%.

In 2023, Alzaqebah et al. (2023) introduced a modified version of the simulated-annealing algorithm to enhance the performance of DL and ML techniques in Arabic text classification.

Yi et al. (2023) developed a session-based framework for detecting cyberbullying, establishing benchmarks for model comparison, and addressing dataset challenges. Finally, Al-Hashedi et al. (2023) trained emotion detection models on Wikipedia and Twitter datasets, enhancing the understanding of cyberbullying by incorporating emotional contexts. Their work highlights the complexity of online harassment and the need for integrated approaches to mitigate it.

These studies demonstrate a trajectory of innovation and interdisciplinarity, illustrating the progress made in developing more effective and nuanced methods for cyberbullying detection.

Machine learning techniques

We also summarize previous studies that have explored various machine-learning techniques for detecting cyberbullying, with a focus on Arabic language content.

Cyberbullying detection in Arabic content has been revolutionized over recent years by various research studies using machine learning (ML) and deep learning (DL) techniques to tackle this pervasive issue.

In 2017, Haidar et al. (2017) pioneered early work using machine learning techniques like Naive Bayes and SVM, tailored specifically for Arabic. Furthermore, Almutiry et al. (2021) advanced this work by employing SVM along with normalization and tokenization preprocessing techniques to enhance Arabic Sentiment Analysis (ASA) and detect cyberbullying tweets.

In 2022, AlFarah et al. (2022) performed research utilizing five artificial intelligence (AI) techniques—Naive Bayes, SVM, Logistic Regression, Random Forest, and K-Nearest Neighbor—to detect cyberbullying among Arabic messages on Twitter and YouTube. Their evaluation assessed the performance and robustness of these models in an under-researched area. Moreover, in the same year, Shannaq et al. (2022) used the Arabic Cyberbullying Corpus (ArCybC) with the SVM algorithm, achieving significant accuracy and F1 scores, demonstrating the algorithm's efficacy across diverse Twitter domains, including gaming, sports news, and celebrities.

Similarly, in 2022, Khairy et al. (2022) conducted an experimental investigation comparing single machine learning models such as K-Neighbors, Logistic Regression, and Linear SVC against ensemble methods like bagging, voting, and boosting, with the voting ensemble model showing superior performance. In contrast, Bouliche et al. (2022) proposed an innovative solution by directly feeding dynamic temporal graphs into their model, meeting the unique challenges associated with Arabic NLP and providing adequate focus on those challenges.

Meanwhile, Alduailaj et al. (2023) focused on optimizing SVM classifier performance through the integration of Farasa to better handle the complexity of Arabic text and enhance cyberbullying detection. Lastly, Muneer et al. (2023) proposed an innovative ensemble stacking learning approach utilizing a modified BERT model (BERT-M), which could easily be applied to other languages and social media platforms beyond Twitter.

Each of these studies not only contributes to an expanding body of knowledge but also showcases the diverse methodologies and technological advancements developed to combat cyberbullying in Arabic. These efforts demonstrate the evolution of detection capabilities and increasing precision in recognizing harmful content.

Integrated Deep Learning and Machine Learning Techniques (Hybrid Approaches)

We review previous studies that have explored hybrid approaches combining deep learning (DL) and machine learning (ML) techniques to enhance cyberbullying detection, highlighting their methodologies and key findings.

Research on cyberbullying detection has steadily progressed, focusing on developing more precise detection methods and gaining insights into how online harassment appears across different languages and contexts.

In 2021, Akhter et al. (2021) conducted research to investigate the rise of abusive language on social media platforms, which poses an increased risk of cyberbullying. Their investigation involved both conventional machine learning (ML) models, such as Naive Bayes, SVM, and Logistic Regression, as well as deep learning (DL) models like CNN, LSTM, and BLSTM, to detect cyberbullying activity. The results demonstrated that the CNN model outperformed conventional ML models in detecting cyberbullying incidents. Additionally, the CNN model proved more effective than conventional ML models when applied to both Urdu and Roman Urdu datasets, clearly highlighting the superior performance of DL models in cyberbullying detection.

In the same year, Cheng et al. (2021) provided an extensive overview of session-based cyberbullying detection, emphasizing the importance of multimodality analysis, temporal dynamics modeling, and user interaction modeling techniques for effective detection. Their study underscored the complexity of session-based approaches and the advanced modeling techniques required to address these challenges effectively.

Furthermore, in 2023, Alrashidi et al. (2023) conducted research on Arabic abusive content detection as a multi-class classification problem using natural language processing (NLP) techniques. They employed multiple models, including ML, DL, pre-trained language models, and multi-task learning (MTL), to expand the capabilities of automated systems in detecting and processing abusive language.

These studies represent significant advancements in cyberbullying detection, progressing from simple machine-learning techniques to more complex DL and NLP methodologies.

Moreover, they illustrate how adapting detection systems to specific linguistic contexts and online interactions contributes to creating safer and more respectful digital environments.

Table 1. Summary of previous studies on cyberbullying Detection, including applied techniques, research objectives, and identifying their limitations.

Ref#	Title	Technique	Objectives	Limitations
(Roy et al., 2022)	Cyberbullying Detection Using Deep Transfer Learning	The domain of cyberbullying detection, specifically addressing the challenge of image-based cyberbullying on social networking platforms	The main objective of the paper is to develop a model that can effectively detect image-based cyberbullying posts on social platforms. The paper explores deep learning and transfer learning frameworks to find the best-suited model for this task.	The paper mentions that cyberbullying, especially image-based cyberbullying, is a challenging task due to the various problems embedded within it. However, the specific challenges faced in the research or any limitations of the proposed approach are not explicitly discussed in the text.
(Alzaqebah et al., 2023)	Cyberbullying detection framework for short and imbalanced Arabic datasets	- Long Short-Term Memory (LSTM) - Bidirectional LSTM (Bi-LSTM)	- Address challenges such as data imbalance and expression implicitness in cyberbullying detection. - Propose a modified simulated annealing optimization algorithm to balance the training set. - Evaluate the performance of traditional machine learning and deep learning algorithms for cyberbullying classification in the context of Arabic text data	- Data imbalance -expression implicitness in cyberbullying detection - The paper does not explicitly mention potential drawbacks or shortcomings of the proposed method
(Mohaouchane et al., 2019)	Detecting Offensive Language on Arabic Social Media using Deep Learning	The paper utilizes four different neural network architectures: Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with attention mechanism, and a combined CNN-LSTM architecture	The objective of the paper is to develop an effective system for detecting offensive language on Arabic social media, particularly in Arabic YouTube comments.	The unique language characteristics and features of Arabic may pose challenges in developing effective models for offensive language detection
(Husain et al., 2020)	"Salamnet at semeval-2020 task12: Deep learning approach for arabic offensive language detection."	Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), and Long-Short Term Memory (LSTM) models with different design architectures.	The objective of the paper is to develop an offensive language detection system for Arabic social media, specifically for the SemEval-2020shared task.	The unique language characteristics and features of Arabic may pose challenges in developing accurate models for offensive language detection.
(Emon et al., 2019)	A Deep Learning Approach to Detect	Linear Support Vector Classifier (LinearSVC), Logistic Regression (Logit),	The main objective of the paper is to detect and eliminate abusive content in the Bengali language found	the complexity of detecting abusive content in Bengali text, the limited availability of labeled training data, and the need for

Ref#	Title	Technique	Objectives	Limitations
	Abusive Bengali Text	Multinomial Naïve Bayes (MNB), Random Forest (RF), Artificial Neural Network (ANN), and Recurrent Neural Network (RNN) with a Long Short-Term Memory (LSTM) cell	in social media platforms, online news portals, and blog commenting sections	effective preprocessing techniques for Bengali language text analysis.
(Yi et al., 2023)	Session-based cyberbullying detection in social media: A survey	-Convolutional Neural Networks (CNN) -Bidirectional Long Short-Term Memory (BLSTM) -Gated Recurrent Unit (GRU) -Long Short-Term Memory (LSTM) -Recurrent Neural Networks (RNN)	- Defining a framework for session-based cyberbullying detection - Understanding the challenges and progress in session-based cyberbullying detection - Proposing evidence-based criteria for best practices in dataset creation - Performing benchmark experiments -Identifying open challenges for future research.	- Mismatch between reported and published datasets - Predominantly crowdsourced annotation of datasets - Selection of social media platforms
(Akhter et al., 2021)	Abusive language detection from social media comments using conventional machine learning and deep learning approaches	The paper explores five diverse conventional machine learning (ML) models, including Naive Bayes (NB), Support Vector Machine (SVM), IBK, Logistic Regression, and JRip.	The main objective of the paper is to detect and analyze abusive language in Urdu and Roman Urdu comments from social media platforms.	natural language constructs, English-like nature of Roman Urdu script, and Nastaleeq style of Urdu for processing and classifying the comments
(AlFarah et al., 2022)	Arabic Cyberbullying Detection from Imbalanced Dataset Using Machine Learning	The paper employs five machine-learning techniques for Arabic cyberbullying detection: Naïve Bayes (NB), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest, and K-Nearest Neighbor (KNN).	The objective of the paper is to address the lack of research on Arabic cyberbullying detection and investigate the performance of different machine learning techniques for this purpose.	1. The complexity of the Arabic language, with its 30 different dialects. Limited research on Arabic cyberbullying detection, resulting in a lack of established methodologies and benchmarks.
(Haidar et al., 2017)	Multilingual Cyberbullying Detection System	The paper utilizes machine learning techniques, specifically Naïve Bayes and Support Vector Machine (SVM), for cyberbullying detection in Arabic language content.	The objective of the paper is to address the lack of research on cyberbullying detection in the Arabic language.	Limited research on Arabic cyberbullying detection, resulting in a lack of established methodologies and benchmarks.

Ref#	Title	Technique	Objectives	Limitations
(Alduailaj et al., 2023)	Detecting Arabic Cyberbullying Tweets Using Machine Learning	The paper utilizes the Support Vector Machine (SVM) classifier algorithm for cyberbullying detection. Additionally, the Farasa tool is used to overcome text limitations and improve the detection of bullying attacks in Arabic text.	The objective of the paper is to address the limited research on cyberbullying detection in the Arabic language.	Limited focus on proposing detection mechanisms for Arabic cyberbullying.
(Almutiry et al., 2021)	Arabic CyberBullying Detection Using Arabic Sentiment Analysis	The paper utilizes Machine Learning (ML) and the Support Vector Machine (SVM) algorithm for training a dataset collected automatically through ArabiTools and Twitter API.	To categorize Arabic tweets into CyberBullying and Non-CyberBullying categories.	41 Structural and morphological complexity of the Arabic language. 42 Limited availability of datasets and resources for Arabic sentiment analysis. The need to address CyberBullying detection in languages other than English.
(Alrashidi et al., 2023)	Abusive Content Detection in Arabic Tweets Using Multi-Task Learning and Transformer-Based Models	Natural language processing (NLP) techniques and develops a framework that incorporates machine learning (ML), deep learning (DL), and pretrained Arabic language models (LMs).	The objective of the paper is to develop an intelligent prediction system for automatically detecting abusive content in Arabic tweets.	The highly imbalanced and low number of samples in some attribute labels of the multi-aspect annotation dataset pose challenges for achieving strong performance in those attributes. Dealing with a multi-aspect annotation dataset is inherently challenging, requiring careful consideration of limitations and potential biases.
(Cheng et al., 2021)	Session-Based Cyberbullying Detection: Problems and Challenges	-Multi-Modality Analysis -Temporal Dynamics Analysis -Hierarchical Structure and Attention Modeling -User Interaction Modeling	- Complex Data Analysis - Core Challenges -Resource for Future Research	NA.
(Al-Hashedi et al., 2023)	Cyberbullying Detection Based on Emotion	The study focuses on cyberbullying detection, utilizing datasets from Wikipedia (toxic dataset) and Twitter (hate speech dataset) to train cyberbullying detection models (CDMs). Emotion datasets, namely Cleaned Balanced Emotional Tweets (CBET) and	The main objective of the paper is to propose cyberbullying detection models that utilize contextual, emotion, and sentiment features. The study provides a comprehensive emotion-annotated dataset for cyberbullying detection, contributing to the research in this field.	The specific challenges faced in the study are not explicitly mentioned in the text.

Ref#	Title	Technique	Objectives	Limitations
		Twitter Emotion Corpus (TEC), are combined and used to train an Emotion Detection Model (EDM) for extracting emotions from cyberbullying datasets.		
(Kompally et al., 2021)	MaLang: A Decentralized Deep Learning Approach for Detecting Abusive Textual Content	detection of abusive textual content, specifically focusing on cyberbullying in workplace communication using decentralized deep learning approach called MaLang.	The main objective of the paper is to propose the MaLang approach, a decentralized deep learning method, for detecting abusive textual content in workplace communication.	It can be inferred that some challenges may include ensuring employee compliance in installing and maintaining the CASE application on corporate and personal devices, as well as overcoming potential privacy concerns. Dataset in Arabic and accessibility: The text does not specify the language of the dataset used in the research.
(Kumari et al., 2019)	AI ML NIT Patna at HASOC 2019: Deep Learning Approach for Identification of Abusive Content.	The paper proposes a deep learning system based on Convolutional Neural Networks (CNN) to identify hate speech, offensive language, and profanity.	The main objective of the paper is to propose a deep learning system based on Convolutional Neural Networks (CNN) for the identification of hate speech, offensive language, and profanity on social media platforms.	The study highlights the challenges associated with identifying hate speech, offensive language, and profanity due to the nature of natural language structures. It mentions the difficulty of differentiating among hate speech, offensive language, and profanity, even for humans, due to fine and subtle differences.
(Bouliche et al., 2022)	Detection of cyberbullying in Arabic social media using dynamic graph neural network	detection of cyberbullying and electronic crimes using Natural Language Processing (NLP) techniques and neural networks	To propose techniques for using dynamic temporal graphs as direct inputs for detecting cybercrimes, including cyberbullying.	1.Limited research on using dynamic temporal graphs for detecting cybercrimes. 2.Insufficient focus on Natural Language Processing (NLP) for the Arabic language. Difficulties related to the unique characteristics of the Arabic language and its writing style.
(Shannaq et al., 2022)	Offensive Language Detection in Arabic Social Networks Using Evolutionary-Based Classifiers Learned From Fine-Tuned Embeddings	The study utilizes the Arabic Cyberbullying Corpus (ArCybC), which consists of tweets collected from four Twitter domains: gaming, sports, news, and celebrities.	The main objective of the paper is to address the challenges of automatically detecting offensive tweets in the Arabic language. The proposed approach's goal is to achieve accurate identification and classification of offensive and non-offensive text, with superior performance metrics on the ArCybC dataset.	The paper addresses the challenges posed by the excessive use of social networks, including the spread of offensive language, cyberbullying, and hate speech. Manual detection of cyberharassment is slow, cumbersome, and impractical for large-scale data, highlighting the need for automated methods. The specific challenge in this study is the detection of offensive tweets in the Arabic language, which has limited research compared to other languages.

Ref#	Title	Technique	Objectives	Limitations
(Khairy et al., 2022)	Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection	offensive language detection and cyberbullying, specifically focusing on Arabic text.	The main objective of the paper is to automate the detection of offensive language and cyberbullying in Arabic text.	the fallibility of machine learning algorithms in detecting offensive language and cyberbullying, potential biases in the training data, and the limited scope of the study focusing only on Arabic text.
(Muneer et al., 2013)	Cyberbullying Detection on social media Using Stacking Ensemble Learning and Enhanced BERT	The paper proposes an ensemble stacking learning approach using a combination of Deep Neural Network (DNN) methods, specifically a modified BERT model (BERT-M).	The objective of the paper is to propose an ensemble stacking learning approach for detecting cyberbullying on Twitter using Deep Neural Network (DNN) methods	Limited generalizability highlights the need for further research in this field.

Data collection

Collecting data mechanisms on cyberbullying from social media platforms:

1- **Text Analytics:** Utilize natural language processing and ML techniques to analyze text posted on social platforms.

2- **Use Application Programming Interfaces (APIs):** Utilize APIs offered by platforms like Twitter and Facebook to collect public data, as well as to analyze signs of bullying.

3- **Surveys and Questionnaires:** Administer surveys and questionnaires to collect user experiences with cyberbullying.

4- **Big Data Analytics:** Collect large volumes of social media data and examine it using advanced data analysis tools to detect patterns of bullying on these platforms.

5- **Collaboration with Organizations and Groups:** Work closely with non-profit organizations and research groups to share information, improve data collection methods, combat cyberbullying more efficiently, and promote online safety.

6- **Software and Tool Development:** Create software and tools to track interactions on social platforms and evaluate them to detect any content that might constitute cyberbullying.

We summarized the datastes, applied techniques, and social media platforms in Cyberbullying detection studies in Table 2.

Table 2. Summary of the datasets, applied techniques, and social media platforms in Cyberbullying detection studies.

Ref#	Dataset Used	Technique use	Platform	Record	NB	Bullying	TOB	Avail.	Papers
(Roy et al., 2022)	Collected images dataset	2DCNN VGG16 INCEPTION V3	Social Media (not specified)	3,000	1,542	1,458	O.L.	✗	-
(Alzaqebah et al., 2023)	Dataset 1	Machine Learning, Deep Learning (SVM, LSTM, Bi-LSTM), and Optimization (Modified Simulated Annealing)	Twitter, YouTube, and various social media platforms	15,049	9,237	Dataset 1: 5,812;	O.L.	✓	(Alakrot et al.,2018)
	Dataset 2			4,000	3,325	Dataset 2: 675;		✓	(Chowdhury et al.,2020)
	Dataset 3_test, Dataset 3_train			2,000 8,000	1,598 6,411	Dataset 3_test: 402; Dataset 3_train: 1,589		✓	(Mubarak et al.,2020)
(Mohaouchane et al., 2019)	Multi-Aspect Hate Speech Dataset	CNN B-LSTM ATTENTION B-LSTM COMBIND CNN-LSTM	Youtube	15,050	N/A	N/A	O.L.	✓	(Alakrot et al.,2018)
(Husain et al., 2020)	Arabic OffensEval 2020 dataset	ML: NB, SVM, IBK, Logistic, JRip; DL: CNN, LSTM, BLSTM, CLSTM	Twitter	10,000	8,100	1,900	O.L.	✓	(Mubarak et al., 2020)
(Emon et al.,2019)	public comment YouTube Dataset	ANN RNN SVC LSTM CNN	Twitter, YouTube	4,700	N/A	N/A	O.L.	✗	-
(Yi et al., 2023)	A balanced dataset specifically collected	Sentiment Analysis, Machine Learning (Specific techniques not detailed in the provided snippets)	Twitter & Facebook	12,000	6,000	6,000	O.L.	✗	-
(Akhter et al.,2021)	Arabic1 (Twitter),	Ensemble Machine Learning, Single ML Classifiers	Twitter, YouTube, Facebook	Arabic1: 1100,	Arabic1: 453,	Arabic1: 647,	O.L.	✓	(Balakrishnan V et al.,2020)
	Arabic2 (YouTube),			Arabic2: 8577,	Arabic2: 5332,	Arabic2: 3245,			
	Proposed Dataset (Facebook & Twitter)			Proposed: 12000	Proposed: 6000	Proposed: 6000			
(AlFarah et al., 2022)	Twitter and Facebook	Stacking Ensemble Learning, Enhanced BERT (BERT-M)	Twitter, Facebook	37,373; ~20,000	8,000	12,000	O.L.	✗	-
)Haidar et al.,2017(	A dataset consisting of 1,000 and 3,000 images	Deep Transfer Learning including VGG16 and InceptionV3 for feature	Various social platforms with images collected from	1,000 images dataset and 3,000 images dataset	For 1,000 images dataset: Not provided; For 3,000 images dataset:	For 1,000 images dataset: Not provided; For 3,000 images dataset: 1,458 bullying	O.L.	✗	-

Ref#	Dataset Used	Technique use	Platform	Record	NB	Bullying	TOB	Avail.	Papers
		extraction	Google images and MMHS150 K dataset		1,542 not bullying				
(Alduailaj et al.,2023)	Wikipedia and Twitter Datasets	Emotion Mining, BERT, Sentiment Analysis	Wikipedia, Twitter	Wikipedia: 21,830 comments, Twitter: 20,620 tweets	N/A	N/A	O.L.	✗	-
(Almutiry et al.,2021)	HASOC 2019 Dataset	Convolutional Neural Networks (CNN), GloVe, fastText, One-hot embedding	Twitter, Facebook	4700	1563	3137	O.L.	✓	(Kumari et al.,2020) (Gudumotu et al.,2023)
(Alrashidi et al.,2023)	Dataset 1, Dataset 2, Dataset 3_test, Dataset 3_train	Machine Learning Techniques	Social Media Platforms	29,049	20,571	8,478	O.L.	✓	(Husain et al.,2021()
(Cheng et al.,2021)	Arabic Cyberbullying Corpus (ArCybC)	Fine-tuned word embeddings, XGBoost, SVM, Genetic Algorithm (GA) optimization	Twitter (gaming, sports, news, and celebrities domains)	4,505	Varied (based on dataset division for NOT and HOF classes)	Offensive Language	O.L.	✓	(Alakrot et al.,2018)
(Al-Hashedi et al.,2023)	YouTube Comments, Levantine Twitter Dataset (L-HSAB), OSACT	BERT-Based Model, Transfer Learning	Various including YouTube, Twitter, News Comments	YouTube: 15,050, L-HSAB: 5,846, OSACT: 10,000	YouTube: 9,237 (Not Offensive), L-HSAB: 3,650 (Normal), OSACT: 8,100 (Not Offensive)	YouTube: 5,813 (Offensive), L-HSAB: 2,196 (Hate+Abusive), OSACT: 1,900 (Offensive)	O.L.	✗	-
(Kompally et al.,2021)	Collected from Twitter and YouTube APIs	Machine Learning with Support Vector Machine (SVM) Classifier, Natural Language Processing (NLP)	Twitter, YouTube	30,000	N/A	N/A	O.L.	✓	(Hoque et al.,2023) (Yastira et al.,2023)
	Comments from social media and online resources	ML Algorithms (LinearSVC, Logit, MNB, RF, ANN, RNN with LSTM)	Social Media (Facebook, YouTube, Prothom Alo Online)	4,700	2,022	2,678		✓	
(Kumari et al.,2019)	YouTube, Prothom Alo Online, & Facebook	Deep Learning (LinearSVC, Logistic Regression, Multinomial Naïve Bayes,	Multiple social media platforms	4,700	Positive: 1,118, Neutral: Not specified	Slang: 618, Religious Hatred: 581, Personal Attack: 624, Politically Violated: 410, Antifeminism:	O.L.	✓	(Ahmed et al.,2021) (Sarker et al.,2022)

Ref#	Dataset Used	Technique use	Platform	Record	NB	Bullying	TOB	Avail.	Papers
		Random Forest, ANN, RNN with LSTM)				445			
)Bouliche et al.,2022(	Jigsaw/Google Toxic Comment Dataset, Hate-speech Dataset	Bi-LSTM, Bi-GRU, Deep Learning, GloVe for word embeddings	Workplace messaging applications	157,000	N/A	N/A	O.L.	✓	(Kompally et al.,2021)
(Shannaq et al.,2022)	Multiple datasets reviewed including Instagram, Vine, and others	Machine Learning, Deep Learning, Rule-Based Methods, and Hybrid Approaches	FormSpring, YouTube, Twitter, Wikipedia, ASKfm	Varied across reviewed datasets	Varied	Varied	O.L.	✓	(Salawu et al.,2021) (Sari et al.,2022) (Reynolds et al.,2011)
(Muneer et al.,2013)	Dataset of Arabic YouTube comments	Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bi-LSTM with attention mechanism, Combined CNN-LSTM	YouTube	15,050 comments	62% inoffensive (approximately 9,331 comments)	38% offensive (approximately 5,719 comments)	O.L.	✗	-

Abbreviations: OL: Offensive Language; TOB: Type of Bullying; NB: Not-bully; Avail: Availability

With the use of social media platforms like Twitter, Facebook, and YouTube becoming more prevalent than ever, cyberbullying has become an increasing threat to individuals and communities alike. Researching abusive behaviors presents researchers with an immense challenge; whether text is in Arabic or English requires advanced techniques for text analysis to detect harmful content that requires precise comprehension of data as well as selection of analytical tools that fit for purpose.

As shown in Table 2, regarding the Arabic language, studies reviewed by Husain et al. (2021), Cheng et al. (2021), and Khairy et al. (2022) have employed various datasets ranging from a few thousand to tens of thousands of samples. For instance, Alzaqebah et al. (2023) utilized samples from Twitter and YouTube, with 9% classified as abusive, while one researcher alone classified 15,049 samples as abusive. These ratios highlight the challenges researchers face when working with imbalanced data, where abusive samples constitute only a small proportion of total samples, complicating the training of analytical models.

Similarly, English language texts present their own set of challenges for text analysis (Yi et al., 2023). Although English offers access to a wealth of resources and open-source libraries for text analysis, dialectal variations and different linguistic styles create barriers to model accuracy. Some studies have employed machine learning techniques, such as Support Vector

Machines (SVM) and Convolutional Neural Networks (CNN), to analyze English texts for abusive behavioral patterns.

Transformer-based models like BERT and MARBERT, as highlighted by AlFarah et al. (2022) and Kompally et al. (2021), have proven particularly effective in this regard, thanks to their focus on a deeper understanding of context, which enhances their accuracy in recognizing abusive material in both Arabic and English texts. BERT can easily adapt to multiple languages, making it an efficient tool for detecting cyberbullying across linguistic boundaries.

To effectively identify cyberbullying, an integrated approach should be utilized, starting with the collection and analysis of large, balanced datasets. When imbalanced data exists, techniques like resampling or generating new samples may be implemented to enhance model performance. Furthermore, advanced transformer models such as BERT can provide crucial support in handling complex texts while better comprehending context, whether in Arabic or English. Their accuracy can be further improved through transfer learning techniques and continuous evaluation.

Conclusion: In conclusion, detecting cyberbullying in texts written in either Arabic or English is a formidable task that demands sophisticated techniques and intelligent data processing. Enhanced transformer models are effective tools for achieving this goal; however, the primary challenge remains to ensure data balance while simultaneously developing models capable of adapting to the complexities of various languages. By employing appropriate methodologies and continuous evaluation techniques, significant progress can be made against cyberbullying on social media platforms.

Discussion

This survey provides valuable insights into the current landscape of cyberbullying detection techniques, particularly those driven by artificial intelligence (AI) and deep learning (DL). As cyberbullying incidents become more frequent due to social media platforms like Facebook and Instagram, the development of sophisticated detection systems that can accommodate different languages and contexts becomes ever more essential.

One major obstacle identified across the studies reviewed is language-specific nuances, especially in Arabic. The language's abundant morphology and various dialects present formidable challenges to both machine learning (ML) and deep learning (DL) models. While various studies have successfully implemented models such as Support Vector Machines (SVM) and Convolutional Neural Networks (CNN) for cyberbullying detection in Arabic, dialectal variance and expression limit the generalizability of these models. To address this challenge, more robust preprocessing techniques and models that better capture Arabic's rich linguistic diversity must be developed.

In contrast, English presents its own set of challenges despite extensive research and resources available, particularly regarding dialect variation and informal language use across social media platforms such as Twitter. However, transformer-based models like BERT and MARBERT have shown promising results due to their ability to quickly grasp contextual nuances. These models have been particularly effective at recognizing abusive content in Arabic texts, suggesting they may serve as universal cyberbullying detection frameworks across languages.

Moreover, studies on data imbalance have repeatedly highlighted its effects. Many datasets used for training models contain an imbalanced number of non-abusive samples compared to abusive samples, which skews model performance. Resampling or creating synthetic data has been proposed as methods to mitigate this problem; however, these techniques often introduce additional challenges, such as overfitting or noise in the datasets. Developing models capable of generalizing to real-world situations requires the use of large and balanced datasets for training.

Additionally, hybrid approaches that combine deep learning (DL) and machine learning (ML) techniques have proven highly successful at increasing the accuracy and robustness of cyberbullying detection systems. By leveraging the strengths of both methodologies—ML's expertise with structured data processing and DL's proficiency in processing unstructured text—more comprehensive detection frameworks have emerged using this hybrid model approach. These hybrid models represent an exciting avenue for future research as they address limitations encountered with current single-method approaches.

Despite considerable advances in cyberbullying detection, several challenges remain. The complexity of natural languages like Arabic, with its intricate morphological structures, and issues of data imbalance continue to hinder the emergence of universally effective models. Moving forward, research should focus on improving dataset quality while creating adaptable models capable of accommodating multiple languages and dialects, and exploring hybrid approaches that combine the advantages of machine learning (ML) and deep learning (DL). By addressing these obstacles effectively, cyberbullying detection can become more efficient and accessible. Systems designed to function across diverse linguistic and cultural environments are becoming increasingly valuable.

In the light of this discussion, we can identify the gaps as follows:

1- Problem of Data Imbalance:

Challenge: Many datasets used for training models contain an imbalanced distribution of abusive and non-abusive samples that detracts from model performance.

Gap: There is an immediate need for effective strategies that address data imbalance, such as resampling or synthesizing synthetic samples without jeopardizing model accuracy.

2- Generalizing Dialects and Linguistic Styles:

Challenge: Existing models often struggle to generalize across dialects and linguistic styles within Arabic as well as English, leading to gaps.

Gap: To ensure effective models that adapt to linguistic diversity within a single language as well as between multiple ones.

3- Limited Research into Non-English Languages:

Challenge: Cyberbullying detection research tends to focus on English; few studies exist that target other languages.

Gap: Research should focus more on non-English languages, with datasets and models tailored specifically for these.

4- Challenges of Integrating Machine Learning and Deep Learning Techniques:

Challenge: Although hybrid approaches have proven their potential, effectively combining machine learning and deep learning techniques remains an obstacle.

Gap: Further investigation should focus on ways to optimize these hybrid approaches to increase accuracy and efficiency when dealing with cyberbullying detection models.

These points summarize the key areas identified during discussion and suggest ways that future research could enhance cyberbullying detection systems.

Conclusion

In conclusion, the experiments that were evaluated show that DL and ML may be used to identify cyberbullying and abusive language on different platforms and in different languages. Problems with data quality, complicated languages, and the need for more inclusive research do persist, however. Making cyberbullying detection more effective and accessible across diverse cultures and languages might be achieved by addressing these problems via creative approaches and expanded linguistic coverage. This would considerably advance the subject.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Ahmed, M. T., Rahman, M., Nur, S., Islam, A. Z. M. T., & Das, D. (2021). Natural language processing and machine learning based cyberbullying detection for Bangla and Romanized Bangla texts. *TELKOMNIKA (Telecommunication Computing Electronics and Control), 20*(1), 89-97.

Ahmed, Md. T., Islam, A., Rahman, M., Rashed, Md. G., Urmi, A. S., & Das, D. (2023). Cyberbullying detection based on hybrid ensemble method using deep learning technique in Bangla dataset. *International Journal of Advanced Computer Science and Applications, 14*(9), 545-551.

Akhter, M. P., Jiangbin, Z., Naqvi, I. R., AbdelMajeed, M., & Zia, T. (2022). Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems, 28*(6), 1925-1940.

Al Shamsi, A. A. (2019). Effectiveness of cyber security awareness program for young children: A case study in UAE. *International Journal of Information Technology and Language Studies, 3*(2), 8-29.

Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Dataset construction for the detection of anti-social behaviour in online communication in Arabic. *Procedia Computer Science, 142*, 174-181.

Alabdulwahab, A., Haq, M. A., & Alshehri, M. (2023). Cyberbullying detection using machine learning and deep learning. *International Journal of Advanced Computer Science and Applications, 14*(10), 424–432.

Alduailaj, A. M., & Belghith, A. (2023). Detecting Arabic cyberbullying tweets using machine learning. *Machine Learning and Knowledge Extraction, 5*(1), 29-42.

AlFarah, M. E., Alharbi, B. Y., & Ibrahim, D. M. (2022). Arabic cyberbullying detection from imbalanced dataset using machine learning. *Soft Computing and Its Engineering Applications*, 397–409.

AlHarbi, B. Y., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M., Alshobaili, J. F., & Ibrahim, D. M. (2019). Automatic cyberbullying detection in Arabic social media. *International Journal of Engineering Research and Technology, 12*(12), 2330-2335.

AlHarbi, B. Y., AlHarbi, M. S., AlZahrani, N. J., Alsheail, M. M., & Ibrahim, D. M. (2020). Using machine learning algorithms for automatic cyberbullying detection in Arabic social media. *Journal of Information Technology Management, 12*(2), 123-130.

Al-Hashedi, M., Soon, L. K., Goh, H. N., Lim, A. H. L., & Siew, E. G. (2023). Cyberbullying detection based on emotion. *IEEE Access, 11*, 53907-53918.

Almutiry, S., & Abdel Fattah, M. (2021). Arabic cyberbullying detection using Arabic sentiment analysis. *The Egyptian Journal of Language Engineering, 8*(1), 39-50.

Alrashidi, B., Jamal, A., & Alkhathlan, A. (2023). Abusive content detection in Arabic tweets using multi-task learning and transformer-based models. *Applied Sciences, 13*(10), 5825.

Alzaqebah, M., Jaradat, G. M., Nassan, D., Alnasser, R., Alsmadi, M. K., Almarashdeh, I., & Alkhushayni, S. (2023). Cyberbullying detection framework for short and imbalanced Arabic datasets. *Journal of King Saud University-Computer and Information Sciences, 35*(8), 101652.

Balakrishnan, V., Khan, S., & Arabnia, H. R. (2020). Improving cyberbullying detection using Twitter users' psychological features and machine learning. *Computers & Security, 90*, 101710.

Bouliche, A., & Rezoug, A. (2022). Detection of cyberbullying in Arabic social media using dynamic graph neural network. In *TACC* (pp. 1-11).

Charbuty, B., & Abdulazeez, A. (2021). Classification based on decision tree algorithm for machine learning. *Journal of Applied Science and Technology Trends, 2*(01), 20-28.

Chen, Y., Zhou, Y., Zhu, S., & Xu, H. (2012, September). Detecting offensive language in social media to protect adolescent online safety. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Conference on Social Computing* (pp. 71-80). IEEE.

Cheng, L., Silva, Y. N., Hall, D., & Liu, H. (2020). Session-based cyberbullying detection: Problems and challenges. *IEEE Internet Computing, 25*(2), 66-72.

Chowdhury, S. A., Mubarak, H., Abdelali, A., Jung, S. G., Jansen, B. J., & Salminen, J. (2020, May). A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6203-6212).

Craig, W. M., & Pepler, D. J. (2007). Understanding bullying: From research to practice. *Canadian Psychology/Psychologie Canadienne, 48*(2), 86.

Daniel, R., Murthy, T., Kumari, C., Lydia, L., Ishak, M. K., Hadjouni, M., & Mostafa, S. M. (2023). Ensemble learning with tournament selected glowworm swarm optimization algorithm for cyberbullying detection on social media. *IEEE Access, 20,* 1–8.

Dinakar, K., et al. (2012). Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS), 2*(3), 1-30.

Emon, E. A., Rahman, S., Banarjee, J., Das, A. K., & Mittra, T. (2019, June). A deep learning approach to detect abusive Bengali text. In *2019 7th International Conference on Smart Computing & Communications (ICSCC)* (pp. 1-5). IEEE.

Gudumotu, C. E., Nukala, S. R., Reddy, K., Konduri, A., & Gireesh, C. (2023). A survey on deep learning models to detect hate speech and bullying in social media. In *Artificial Intelligence for Societal Issues* (pp. 27-44). Cham: Springer International Publishing.

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. *Neurocomputing, 187,* 27-48.

Haidar, B., Chamoun, M., & Yamout, F. (2016). Cyberbullying detection: A survey on multilingual techniques. In *UKSIM 2016*. Pisa, Italy: Publisher.

Haidar, B., Chamoun, M., & Serhrouchni, A. (2017, October). Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content. In *2017 1st Cyber Security in Networking Conference (CSNet)* (pp. 1-8). IEEE.

Hasan, M., Al, E., Hossain, M. S., Akter, A., Ahmed, M., & Islam, S. (2023). A review on deep-learning-based cyberbullying detection. *Future Internet, 15*(179), 1-47.

Hoque, M. N., Chakraborty, P., & Seddiqui, M. H. (2023). The challenges and approaches during the detection of cyberbullying text for low-resource language: A literature review. *ECTI Transactions on Computer and Information Technology (ECTI-CIT), 17*(2), 192-214.

Husain, F., & Uzuner, O. (2021). Transfer learning approach for Arabic offensive language detection system—BERT-based model. *arXiv preprint arXiv:2102.05708*.

Husain, F., Lee, J., Henry, S., & Uzuner, O. (2020). SalamNet at SemEval-2020 Task 12: Deep learning approach for Arabic offensive language detection. *arXiv preprint arXiv:2007.13974*.

Kanam, V. O. (2020). *Comparative sentiment analysis of techniques for cyberbullying detection on Twitter* (Doctoral dissertation, Strathmore University).

Kanan, T., Aldaaja, A., & Hawashin, B. (2020). Cyberbullying and cyberharassment detection using supervised machine learning techniques in Arabic social media contents. *Journal of Internet Technology, 21*(5), 1409-1421.

Khairy, M., Mahmoud, T. M., Omar, A., & Abd El-Hafeez, T. (2023). Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection. *Language Resources and Evaluation*, 1-18.

Kompally, P., Sethuraman, S. C., Walczak, S., Johnson, S., & Cruz, M. V. (2021). Malang: A decentralized deep learning approach for detecting abusive textual content. *Applied Sciences, 11*(18), 8701.

Kufel, J., et al. (2023). What is machine learning, artificial neural networks and deep learning?—Examples of practical applications in medicine. *Diagnostics, 13*(15), 2582.

Kumari, K., & Singh, J. P. (2020, May). AI_ML_NIT_Patna@TRAC-2: Deep learning approach for multilingual aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying* (pp. 113-119).

Kumari, K., & Singh, J. P. (2019). AI ML NIT Patna at HASOC 2019: Deep learning approach for identification of abusive content. *FIRE (Working Notes), 2517,* 328-335.

Mahesh, Batta. (2020). Machine learning algorithms: A review. *International Journal of Science and Research (IJSR), 9*(1), 381-386.

Mahmud, T., Ptaszynski, M., Eronen, J., & Masui, F. (2023). Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Information Processing & Management*, *60*(5), 103454.

Mat Ali, H. (2005). Censorship agent: Identifying and determining offensive words.

McFarland, L. A., & Ployhart, R. E. (2015). Social media: A contextual framework to guide research and practice. *Journal of Applied Psychology, 100*(6), 1653–1677.

Mohaouchane, H., Mourhir, A., & Nikolov, N. S. (2019, October). Detecting offensive language on Arabic social media using deep learning. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 466-471). IEEE.

Mubarak, H., et al. (2020). Arabic offensive language on Twitter: Analysis and experiments. *arXiv preprint arXiv:2004.02192*.

Muneer, A., et al. (2023). Cyberbullying detection on social media using stacking ensemble learning and enhanced BERT. *Information, 14*(8), 467.

Nandakumar, V., Kovoor, B. C., & Sreeja, M. U. (2018). Cyberbullying revelation in Twitter data using Naïve Bayes classifier algorithm. *International Journal of Advanced Research in Computer Science, 9*(1).

Narayanan, S., & Panayiotis, G. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE, 101*(5), 1203-1233.

Noviantho, S. I., & Livia, A. (2017). Cyberbullying classification using text mining. *2017 1st International Conference on Informatics and Computational Sciences (ICICoS)*.

Osisanwo, F. Y., et al. (2017). Supervised machine learning algorithms: Classification and comparison. *International Journal of Computer Trends and Technology (IJCTT), 48*(3), 128-138.

PK, F. A. (1984). What is Artificial Intelligence? In *Success is No Accident: It is Hard Work, Perseverance, Learning, Studying, Sacrifice and Most of All, Love of What You are Doing or Learning to Do* (p. 65).

Reynolds, K., Kontostathis, A., & Edwards, L. (2011, December). Using machine learning to detect cyberbullying. In *2011 10th International Conference on Machine Learning and Applications and Workshops* (Vol. 2, pp. 241-244). IEEE.

Roy, P. K., & Mali, F. U. (2022). Cyberbullying detection using deep transfer learning. *Complex & Intelligent Systems*, *8*(6), 5449–5467.

Salawu, S., Lumsden, J., & He, Y. (2021, August). A large-scale English multi-label Twitter dataset for cyberbullying and online abuse detection. In *The 5th Workshop on Online Abuse and Harms* (pp. 146-156). Association for Computational Linguistics.

Sari, T. I., Ardilla, Z. N., Hayatin, N., & Maskat, R. (2022). Abusive comment identification on Indonesian social media data using hybrid deep learning. *IAES International Journal of Artificial Intelligence*, *11*(3), 895-904.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling, and research directions. *SN Computer Science*, *2*, 1-18.

Sarker, M., Hossain, M. F., Liza, F. R., Sakib, S. N., & Al Farooq, A. (2022, February). A machine learning approach to classify anti-social Bengali comments on social media. In *2022 International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)* (pp. 1-6). IEEE.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge University Press.

Shannag, F., Hammo, B. H., & Faris, H. (2022). The design, construction, and evaluation of annotated Arabic cyberbullying corpus. *Education and Information Technologies*, *27*(8), 10977-11023.

Shannaq, F., et al. (2022). Offensive language detection in Arabic social networks using evolutionary-based classifiers learned from fine-tuned embeddings. *IEEE Access*, *10*, 75018–75039.

Shapiro, S. C. (1992). *Encyclopedia of artificial intelligence* (2nd ed.). New Jersey: Wiley Inter-science.

Willard, N. E. (2007). *Cyberbullying and cyberthreats: Responding to the challenge of online social aggression, threats, and distress*. Research Press.

Yastira, E., & Cahyono, H. D. (2023, October). Intelligent web service system for detecting cyberbullying on Twitter based on support vector machine and random forest algorithms. In *2023 International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE)* (pp. 50-55). IEEE.

Yi, P., & Zubiaga, A. (2023). Session-based cyberbullying detection in social media: A survey. *Online Social Networks and Media*, *36*, 100250.

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists*. O'Reilly Media, Inc.

Židová, M., Hollá, K., & Rybanský, Ľ. (2021). Parental control and cyberbullying. *EDULEARN21 Proceedings*. IATED.

Bibliographic information of this paper for citing:

Alafleh, Kholood; Alabdultif, Abdulatif & Aladhadh, Suliman (2024). Artificial Intelligence-Driven Cyberbullying Detection: A Survey of Current Techniques. *Journal of Information Technology Management*, 16 (4), 38-63. https://doi.org/ 10.22059/jitm.2024.99050