IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 March 2025, accepted 11 April 2025, date of publication 15 April 2025, date of current version 25 April 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3561132

TOPICAL REVIEW

Arabic Cyberbullying Detection: A Comprehensive Review of Datasets and Methodologies

HUDA ALJALAOUD[1,2], **KIA DASHTIPOUR**[1], **AND AHMED Y. AL-DUBAI**[1], **(Senior Member, IEEE)**

[1]School of Computing, Edinburgh Napier University, Merchiston Campus, EH10 5DT Edinburgh, U.K.
[2]Department of Computer Science, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Huda Aljalaoud (Huda.aljalaoud@napier.ac.uk)

ABSTRACT The freedom of speech in online spaces has substantially promoted engagement on social media platforms, where cyberbullying has emerged as a significant consequence. While extensive research has been conducted on cyberbullying detection in English, efforts in the Arabic language remain limited. To address this gap, the current study provides a comprehensive, state-of-the-art review of datasets and methodologies specifically focused on Arabic cyberbullying detection. It systematically reviews different relevant studies from six academic databases, examining their methodologies, dataset characteristics, and performance in terms of classification accuracy and limitations. The paper critically evaluates existing Arabic cyberbullying datasets according to criteria such as dataset size, dialectal diversity, annotation processes, and accessibility. Additionally, this review identifies critical limitations, including dataset scarcity, dialectal imbalance, annotation subjectivity, and methodological constraints. By synthesizing current knowledge, identifying research gaps, and suggesting future directions, this review supports the development of more robust, effective, and linguistically inclusive analytical methods. Ultimately, this work contributes significantly to natural language processing research and advances the creation of safer online environments for Arabic-speaking users.

INDEX TERMS Arabic cyberbullying detection, Arabic cyberbullying dataset, deep learning, machine learning, transformers-based.

I. INTRODUCTION

With the proliferation of online social networks, the widespread availability of information and communication technology, and the prevalent use of computers and smartphones, Internet users now have unprecedented freedom of expression in history [1]. According to a data report, the number of Internet users globally has reached 5.56 billion users, representing 67.9% of the world's population, among these users, approximately 63.9% are active on social media [2]. At countries level, 99% of the population in Saudi Arabia used the Internet in February 2025, and 99.6% of them utilized at least one social media platform [2], [3]. Globally, the predominant users of numerous social media platforms are young people, and this is global trend [3], [4].

Freedom of speech in cyberspace generates abusive behaviours, such as cyberbullying (CB) [5]. CB appeared as a novel type of bullying resulting from swift progress in Web 2.0 technologies that have transformed communications and social interactions on the Internet [6]. In addition, researchers have provided various definitions of CB, and the common interpretation is aggressive and toxic behaviour through online devices [7].

CB is a significant social issue in cyberspace, with most previous studies focusing on its psychological consequences [8]. However, this paper reviews technological approaches aimed at identifying and preventing CB incidents in Arabic. The growing awareness of CB's impact and its

The associate editor coordinating the review of this manuscript and approving it for publication was Larbi Boubchir.

potential adverse effects on victims have driven researchers to develop detection methods. Despite this, existing literature reviews on Arabic CB detection remain limited and fail to comprehensively cover all relevant aspects. To the best of our knowledge, only four surveys have examined detection methods, while no specialized dataset review has been conducted specifically for cyberbullying; instead, a single survey has focused on toxic speech datasets.

Alsunaidi et al. [6] conducted a survey on the analysis of nine empirical studies using automated Arabic CB detection approaches. Their study summarized papers and categorized them into three main classes: supervised learning, unsupervised learning and ensemble learning. All the conducted studies were either supervised learning or ensemble learning. Moreover, Khairy et al. [8] examined 27 research papers on the automatic detection of CB and abusive language in Arabic, along with related detection approaches; only 10 of them focused on CB detection. They concluded that all datasets utilized in the detection process were labelled based on individual posts, which contradicts the definition of CB as repeated behaviour. Additionally, the majority of datasets used were imbalanced, impacting the performance of the classifiers.

Albayari et al. [9] conducts a systematic review of nine research papers on Arabic CB detection, focusing on the machine learning (ML) and deep learning (DL) models utilized, as well as their corresponding performance metrics. The findings indicate that Support Vector Machines (SVM) is the most frequently employed traditional classifier, followed by Random Forest (RF) and Naïve Bayes (NB). In contrast, among deep learning methodologies, Convolutional Neural Networks (CNNs) emerge as the most widely adopted architecture for CB classification. However, a significant challenge identified in this domain is the lack of a standardized benchmark dataset, which obstructs the direct comparison of classification models and their effectiveness across different studies.

Additionally, Arif [10] conducted a systematic review of research on CB detection in both English and Arabic using machine learning techniques. The study examined research directions and theoretical underpinnings in this field while also exploring future directions and challenges. However, despite analyzing 45 research papers, only four were dedicated to Arabic CB detection, highlighting the limited focus on this language. In terms of datasets, Bensalem et al. [11] conducted a comprehensive investigation of 54 Arabic toxic speech datasets and their corresponding studies. The analysis was structured around 18 criteria spanning four key dimensions: annotation processes, availability details, reusability, and content. However, despite the extensive dataset evaluation, only two datasets were explicitly designated for CB detection [12], [13], underscoring the scarcity of Arabic-specific resources in this field.

The necessity of an updated and comprehensive survey on Arabic CB detection research and publicly available datasets arises from several critical factors. Rapid advancements in ML and natural language processing (NLP) have introduced sophisticated techniques capable of addressing the complexities inherent in analyzing Arabic CB. Additionally, the growing volume of Arabic digital content emphasizes the need to reassess existing datasets and detection methodologies, especially given the linguistic diversity and numerous dialects within the Arabic language. Despite its global significance, Arabic remains underrepresented in NLP research, further limiting resources and specialized datasets available for effective CB detection.

The study explicitly addresses several clear objectives. Initially, it seeks to survey recent advancements in Arabic CB detection by critically reviewing state-of-the-art research, methodologies, and analytical tools. Moreover, it thoroughly evaluates publicly available datasets, analyzing their utility and limitations to guide future research in the domain. The study also highlights successful techniques from other languages that could potentially be adapted to Arabic, thereby expanding the range of methodological approaches available for researchers.

Additionally, this research identifies critical challenges currently hindering Arabic CB detection, particularly dataset limitations, dialectal variations, the lack of real-time detection capabilities, and the absence of multimodal analysis methods. It proposes strategic research directions, such as exploring multimodal detection approaches, leveraging large language models (LLMs) and pre-trained language models (PLMs) to manage linguistic diversity, and adopting ensemble and transformer-based architectures for more accurate and context-sensitive outcomes.

It distinguishes itself from prior research through three key contributions. First, it consolidates all available datasets related to Arabic CB and conducts an in-depth analysis of their characteristics. Second, it presents a comprehensive literature review on Arabic CB detection, systematically tracing the evolution of research in this area from its inception. Third, it addresses challenges in the methodologies and datasets used in previous studies, providing critical insights to support future advancements in Arabic CB detection. By identifying these limitations and proposing directions for improvement, this work serves as a foundational resource for researchers seeking to enhance CB detection techniques in Arabic.

Following the introduction, the remainder of this paper is structured as follows: Section II provides background information, including a definition of CB, an overview of the Arabic language, and Arabic CB datasets. Section III outlines the research methodology. Section IV presents a comprehensive review of Arabic CB detection techniques. Section V discusses the advancements, challenges, and future directions in Arabic CB detection. Finally, Section VI concludes the paper by summarizing key findings and research contributions.

II. ARABIC CB DETECTION BACKGROUND

A. CB DETECTION

Bullying is characterized by aggression, in which an imbalance of power exists between the bully and the victim [14].

Moreover, bullying can manifest in various forms, including physical, verbal and psychological aggression. Bullying has historically been confined to in-person interactions, but with the rise of technology, it has also become prevalent on online platforms, referred to as CB [15]. The term cyberbullying first appeared in the New York Times articles in 1995 and in the Canberra Times in 1998 [16]. There is ongoing debate among researchers regarding the definition of CB. Some consider this type of behaviour an extension of traditional bullying, adapted to the digital age, through the use of electronic communications. However, others believe that it should be recognized as a distinct and separate form of mistreatment that differs fundamentally from conventional bullying practices [17].

According to Mouheb et al. [18], 'the consequences of cyberbullying are profound, encompassing adverse effects, such as diminished self-esteem, depression, and anxiety, as well as eliciting feelings of anger, fear, and frustration. In some extreme instances, it tragically culminates in suicide'. Due to the negative consequences that may affect individuals, it is essential to monitor and prevent occurrences of CB to reduce this form of online harassment [19].

Cyberspace hosts numerous forms of CB that impact users on various Internet platforms. Several studies, including those conducted by Sheri Bauman [16], Teng et al. [17], and Haidar et al. [20], classify CB into various types based on its nature and impact, including:

- Flaming: Initiating an online argument or conflict.
- Masquerade: Pretending to be someone else with malicious intentions.
- Denigration: Spreading rumors to harm reputation.
- Impersonation: Pretend to be someone to cause harm or damage to their reputation.
- Harassment: Sending repeated offensive messages.
- Outing: Revealing humiliating information or secrets about someone.
- Trickery: Deceiving someone into sharing sensitive information for online sharing.
- Exclusion: Intentionally and harshly leaving someone out of online group activities.
- Cyberstalking: Repeated and intense harassment, often involving threats and instilling fear.

It is appropriate to assert that our survey provides a comprehensive examination of all forms of CB research conducted in the Arabic language.

B. ARABIC CB STUDIES

Globally, the Arabic language ranks fifth in terms of the most spoken languages, and its utilization on the Internet is expanding significantly [8]. It is the official language in over 20 countries, contributing to a growing research interest in various Arabic language domains. Furthermore, Arabic remains one of the most widely used languages both among Internet users and speakers worldwide [2].

Arabic exhibits a rich diversity of dialects, which can be broadly categorized into two primary forms: Modern Standard Arabic (MSA) and Spoken Dialects. MSA serves as the standardized, literary form of the language, utilized in formal contexts such as writing, media, and official discourse. It is important to note that MSA is not acquired as a first language by any native Arabic speaker. In contrast, spoken Arabic exhibits significant regional variation, with dialects often being partially mutually intelligible across neighboring geographic areas. These spoken dialects can be classified into five major regional groups: Levantine, Gulf, Egyptian, Maghrebi, and Iraqi, each of which contains numerous sub-dialects. The exact number of Arabic dialects remains a topic of scholarly discussion, as classification methodologies differ. However, some linguists estimate that more than 30 distinct dialects exist, reflecting the extensive geographical reach and historical evolution of the language [21]. This linguistic diversity highlights the complexity and cultural richness of Arabic as a global language.

TABLE 1. Distribution of Cyberbullying detection studies: Arabic vs. other languages.

Academic databases	Arabic	Other language
ScienceDirect	2	49
MDPI	2	41
ACM	1	37
Springer link	3	98
IEEE Xplore	9	115
Google Scholar	36	862

However, the efforts in Arabic CB research are still modest when compared with other languages, as shown in Table 1. The table indicates that, while a significant volume of research on CB detection is available in other languages—primarily English, with some studies in Hindi, Bengali, and Turkey—the number of studies focused on Arabic remains comparatively lower. This underscores the importance of promoting and supporting research initiatives in Arabic to better detect and combat CB.

Table 1 highlights a significant disparity in the number of studies focused on CB detection in the Arabic language compared to other languages, spanning the period from 2017 to February 2025 in five academic databases plus Google Scholar. This contrast underscores the relatively limited research attention devoted to Arabic text-based CB detection within the broader academic landscape. As depicted in Table 1, a total of 53 studies focusing on Arabic were identified across various academic databases, compared to 1202 studies in other languages. Specifically, ScienceDirect has 2 studies in Arabic versus 49 in other language; MDPI shows 2 versus 41; ACM has 1 versus 37; Springer Link presents 3 versus 98; and IEEE Xplore features 9 versus 115. This comparison highlights the underrepresentation of Arabic in CB research and underscores the need for increased focus and investigation in this area. On the temporal scale, six studies were conducted before 2022, while 21 studies were published afterward, highlighting the increasing academic attention toward the psychological consequences of CB on social media users' mental health [22], [23], [24],

[25]. This trend underscores the growing recognition of CB as a significant societal issue, prompting intensified research efforts to understand its impact and develop effective mitigation strategies.

C. ARABIC DATASETS

As far as we know, current efforts in the field of Arabic CB datasets have focused on developing new benchmarks, despite the absence of standardised datasets for Arabic CB detection [12], [26]. Many researchers have created experimental CB datasets by extracting data from social media platforms, making it difficult to establish a fair comparison of the effectiveness of existing methods for detecting Arabic CB. However, despite the absence of a standard corpus for Arabic CB detection, ongoing efforts are being made to create one. Nowadays, four Arabic CB datasets are available and can be reached online, as presented in Table 2.

Currently, most studies construct their datasets using data from social networking platforms such as X (formerly Twitter), Facebook, and YouTube [12]. X serves as the primary source for nearly all datasets, as shown in Table 2, primarily due to its easily accessible API for data collection. Table 2 provides a summary of Arabic datasets utilized in CB research. These datasets differ in terms of their sources, sizes, included dialects, balance ratios, availability, and annotation methods. They reflect the diverse approaches and platforms employed in gathering Arabic CB data, underscoring the challenges associated with achieving balance and manual annotation to ensure data quality and representativeness. The online availability of these datasets enhances their utility for broader research and facilitates verification by other scholars.

From Table 2, it is evident that the datasets span multiple platforms, including X, YouTube, Facebook, and Instagram, and encompass various Arabic dialects, such as MSA, Egyptian, Gulf, and Levantine. The balance ratio reflects the proportion of CB instances relative to the total dataset size, underscoring challenges related to dataset representativeness. Most datasets are manually annotated, with only one employing a combination of manual and automated methods. Additionally, while some datasets are publicly available, others require special access upon request. This information is essential for researchers seeking to advance Arabic CB detection, as it provides insight into the scope and characteristics of each dataset, enabling more informed selection and application in future studies.

Building on previous analyses, several key deficiencies in Arabic CB datasets have been identified. These include a predominant class imbalance, where the number of CB instances is significantly lower than non-CB cases, limiting the effectiveness of classification models. Furthermore, dataset sources are primarily concentrated on popular social media platforms, reducing diversity and generalizability. Additionally, restricted public access to some datasets hampers collaborative research efforts. Another major limitation is the scarcity of datasets that offer multi-class categorization of CB incidents, which is crucial for developing more nuanced and context-aware detection systems. Lastly, the heavy reliance on manual annotation raises concerns regarding data consistency and objectivity, as Arabic encompasses a wide range of dialects [27]. When annotators speak a dialect different from the one used in the text, misinterpretations and incorrect annotations may occur. This issue underscores the need for standardized and potentially automated annotation processes to enhance reliability. Addressing these challenges is essential for improving the effectiveness and applicability of CB detection models across various digital platforms.

D. FEATURE REPRESENTATION

The authors in [10] and [32] classified CB detection features into five distinct categories. The first is content features, which are extracted from user texts or derived from emojis, video, images or audio content. NLP techniques are employed. This method is widely utilised in CB detection studies due to its effectiveness in revealing offensive terms. Common models for word representation include word embedding (WE), bag of words (BoW) and term frequency-inverse document frequency (TF-IDF). A second feature is network features, which are significant in CB detection, as they reflect the level of social engagement of the cyberbully. These features may include the number of user followers and the number of accounts followed by the user. This feature is rarely used due to regulations aimed at protecting user privacy on social media platforms.

Third are activity features that capture the user's communication activities, such as the number of posts disliked or liked and engagement with hashtags. Fourth, user profile features are extracted from the user's social media profile and may include demographic information such as gender, age and social group affiliations. The user's age is particularly relevant in determining the severity of CB, as it influences the dynamics between the victim and the cyberbully. Additionally, knowledge of the user's social group types can aid in predicting their personality and behaviour, as there is a correlation between personality traits and hostile behaviour in users. Finally, sentiment features are employed to detect hostile behaviours exhibited by users. This can be achieved through two methods: using a well-trained classifier or utilising a dictionary of sentiment-related words to determine the sentiment of user posts.

E. PERFORMANCE MEASURES

The criteria employed to evaluate the efficacy of individual CB model classifiers consist of accuracy, precision, recall, F1 score, AUC-PR and AUC-ROC. The metrics are described as follows:

Accuracy denotes the proportion of correctly classified instances and is calculated using the following formula:

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

TABLE 2. Cyberbullying Arabic datasets.

Ref.	Dataset source	Dataset size	Dataset name	Dialect included	Balanced ratio	Available public online	Annotation
Alduailaj et al. [13]	X, YouTube	30,000	Arabic-cyberbullying-tweets	MSA	0.3	Yes https://www.kaggle.com/datasets/alanoudaldealij/arabic-cyberbullying-tweets	Manually
Khairy et al. [12]	Facebook, X	12,000	Cyber_2021	MSA	0.5	Yes https://github.com/omammar167/Arabic-Abusive-Datasets/tree/main	Manually
Haidar et al.[20]	X	37,000, 2,196 bullying	NA	Levantine, Gulf and Egypt	Imbalanced 0.062	No	Manually
Shannag et al.[28]	X	4,505	ArCyC	Egyptian, Gulf, and Levantine	0.38	Yes https://data.mendeley.com/datasets/z2dfgrzx47/1	Manually
Albayari et al.[26]	Instagram	46,898	Instagram Arabic corpus	MSA, Egyptian, Gulf, and Levantine	Multi class	Yes https://bit.ly/3Md8mj3	Manually
Almutiry and Fattah [29]	X	17,748	AraBully-Tweets	MSA	Imbalanced	No	Manual and Automatic
Alrougi et al. [30]	X	10,000	ArCBDs	Egyptian, Gulf, and Levantine.	0.38	Upon request.	Manually
Musleh et al.[31]	X	9,000	NA	Saudi Arabia	NA	Upon request.	Manually

where TP = T*rue Positive*, TN = *True Negatives*, FP = False Positive, FN = False Negative.

Precision (P) represents the proportion of accurately identified positive instances relative to the total number of positive predictions and its formula is:

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

Recall (R) indicates the percentage of actual positive instances that are correctly identified by the classifier. It is calculated as follows:

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

The F1 score is known as the F-measure, and it provides a balance between precision and recall in assessing test accuracy. It achieves its optimal value at 1 and its lowest value at 0, computed by

$$F1\ Score = \frac{2 * P * R}{P + R}$$

AUC-ROC (Area Under the Receiver Operating Characteristic Curve) represents a model's ability to distinguish between classes at various thresholds. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across threshold levels, with TPR defined as

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

and FPR as

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

AUC-ROC is the area under this curve, ranging from 0 to 1, where higher values indicate better performance, with 1 representing a perfect classifier and 0.5 indicating random performance. Values below 0.5 suggest that the classifier performs worse than random.

AUC-PR (Area Under the Precision-Recall Curve) is a metric particularly useful for imbalanced datasets, focusing on the model's ability to correctly identify the positive class. The Precision-Recall(PR) curve plots Precision against Recall across thresholds. AUC-PR is the area under this curve, with higher values indicating better detection of positive cases.

In this survey, many studies utilized the accuracy measure [12], [13], [26], [28].

III. RESEARCH METHODOLOGY

In their pioneering study on the challenges associated with detecting CB in Arabic text, Alduailej and Khan [33]

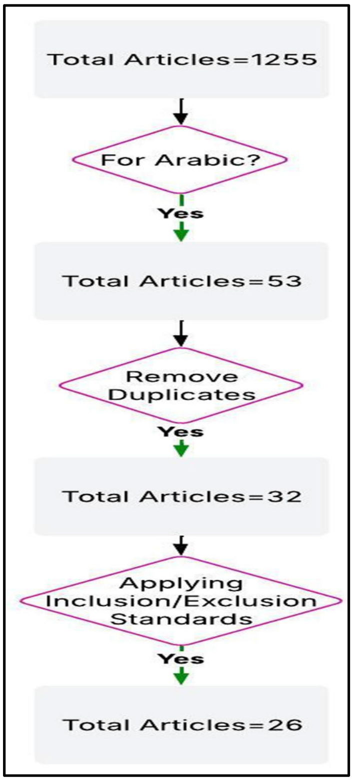


FIGURE 1. A Systematic approach to article selection.

highlighted the absence of prior research specifically focused on this issue within their literature review. In the same year, Haider et al. [20] introduced the first detection model for Arabic CB [34]. Consequently, this systematic review encompasses studies published from 2017 through February 2025, aiming to capture the full spectrum of research efforts in the domain of Arabic CB detection. A total of 26 studies have been included in this review, providing a comprehensive overview of the advancements and methodologies employed in this field.

A. SELECTION STRATEGY

A comprehensive literature search was undertaken using a systematic approach and advanced search queries. The initial query employed was [cyberbullying OR cyberbully] AND [detection], with publication dates restricted from 2017 to 2025. To specifically target Arabic studies, an additional parameter ([Arabic]) was appended. Searches were conducted across five academic databases—ScienceDirect, MDPI, ACM, Springer Link, and IEEE Xplore—in addition to Google Scholar. The search process was completed in February 2025, as detailed in Table 1. This approach was designed to identify and analyse all pertinent studies published within the specified timeframe.

B. INCLUSION AND EXCLUSION STANDARDS

This section outlines the criteria established during the selection process to ensure that the studies included in this analysis are relevant, significant, and aligned with the primary objectives of the research. The criteria were designed to systematically identify and evaluate studies that contribute meaningfully to the field of CB classification in Arabic texts.

Studies were included if they met the following conditions:

- The study must focus on the classification of CB specifically in Arabic texts.
- The study must utilize artificial intelligence (AI) techniques to perform text classification tasks.
- The study must report the model's performance metrics, such as accuracy, precision, recall, F1-score, or other relevant evaluation measures.
- The study must have been published in or after the year 2017 to ensure the inclusion of recent and up-to-date research.

Studies were excluded based on the following criteria:

- Studies that did not address the classification of CB in Arabic texts were excluded.
- Studies that were not published in peer-reviewed journals or conference proceedings were excluded.
- Studies that failed to describe the classification methodology in sufficient detail were excluded.

These criteria were applied to ensure the selection of high-quality, relevant, and recent studies that contribute to the advancement of knowledge in the domain of Arabic text-based CB classification.

IV. ARABIC CB DETECTION METHODS

A. MACHINE LEARNING MODELS

Haidar et al. performed the first work to detect CB in the Arabic language [20]. They have made significant contributions to the field of Arabic CB detection. In their first study, Haider et al. [20] attempted to address the issue of CB in both the Arabic and English languages. While there has been considerable progress in CB detection for English, relatively no work has been done in Arabic. To implement ML techniques, a dataset must be prepared for both training and testing the system, and the WEKA toolkit was chosen for this purpose, as it supports the Arabic language. The dataset consisted of only 2,196 instances of bullying content out of a total of 35,273 instances. Two models, SVM and NB, were selected to detect CB. However, the results obtained by this system are not as accurate as those achieved by previous work on English CB detection systems, while the goal of the paper was to demonstrate the detection of CB in Arabic. In their second work [35], Haider et al. further enhanced their approach by employing DL techniques, specifically by training a feed-forward neural network (FFNN) on an Arabic dataset. More recently, they introduced ensemble ML as an additional method to enhance Arabic CB detection, refining their previous efforts [36].

Alduailaj et al. [13] suggested employing ML techniques to automatically detect CB in Arabic. Specifically, they used an SVM classifier algorithm utilising a real dataset sourced from YouTube and X to identify CB. The Farasa tool was also included to address textual limitations and enhance the detection of bullying incidents on Arabic social media platforms. The findings indicated that SVM continued to surpass NB in the detection of CB content.

Shannag et al. [28] discussed creating and evaluating ArCybC, a multi-dialect annotated Arabic CB corpus, by outlining the compilation process, exploring CB tweets from different X groups, and highlighting the susceptibility of certain groups to harassment. The corpus was filtered using a harassment lexicon and annotated by five annotators. Five ML approaches were then compared for CB detection: logistic regression (LR), RF, decision tree (DT), the extreme gradient boosting algorithm and SVMs. The combination of SVM and word embedding techniques yielded encouraging outcomes. The obstacles included the scarcity of publicly accessible Arabic dataset on CB and the challenges associated with their annotation.

Albayari et al. [26] in their first study, introduce the first Instagram-based Arabic corpus specifically designed for CB and offensive language detection, establishing the largest CB dataset with 46,898 manually annotated comments. The study further evaluates the performance of four classical machine learning classifiers—LR, SVM, RFC, and Multinomial Naïve Bayes (MNB)—on the dataset. Among these, the SVM classifier achieved the highest performance, highlighting its effectiveness in identifying CB content in Arabic social media. By publicly releasing the dataset, the study provides a foundational benchmark for further research in Arabic NLP, facilitating advancements in automated detection of CB language, sentiment classification, and dialectal analysis. Musleh et al. [31] utilized a meticulously curated dataset that captures specific characteristics of the Arabic geographic region, predominantly focusing on Saudi Arabia.. The collected tweets were subjected to extensive pre-processing—including cleaning, normalization, tokenization, stop-word removal, and stemming—prior to feature extraction using both TF-IDF and N-gram models. The study evaluated a suite of nine ML algorithms (SVM, NB, RF, LR, AdaBoost, CatBoost, LightGBM, Bagging, and XGBoost), with the combination of XGBoost and TF-IDF emerging as the superior model, achieving an accuracy of 89.95%. Despite these promising results, the authors acknowledge limitations related to the relatively small and potentially dialect-biased dataset, underscoring the need for further data augmentation and more balanced sampling to enhance the generalizability of the findings.

Saadi et al. [43] aimed to identify negative Arabic comments indicative of CB through the application of an SVM algorithm. Feature extraction was conducted using both the term frequency-inverse document frequency (TF-IDF) vectorizer and count vectorizer (CV) methods. Subsequently, a cuckoo search optimisation algorithm was employed for optimisation, yielding favourable outcomes.

Mursi et al. [44] introduced ArCyb, a machine-learning model designed to identify CB in social media, leveraging a manually annotated Arabic dataset. They implemented both SVM and Multi-layer Perceptron (MLP) classifiers and conducted classification tests on the dataset after preprocessing, yielding accuracy rates of 89% for MLP and 92% for SVM.

B. DEEP LEARNING MODELS

Bashir and Bouguessa [39] introduce a data mining approach for the detection of CB and harassment in Arabic texts, addressing the unique challenges posed by the complexity and resource scarcity of Arabic language processing. The key contributions of the study include the development and evaluation of various learning strategies—ranging from traditional ML classifiers (e.g., Multinomial Naïve Bayes MNB, Linear Support Vector Classification LSVM, LR, Ridge Classifier, Bernoulli Naïve Bayes BNB, RF, and KNN) to DL techniques (with a particular focus on LSTM networks). The authors also compare the effectiveness of different word embedding methods, demonstrating that the Continuous Bag of Words (CBOW) model outperforms the Skip-gram approach in capturing semantic relationships in the context of CB. The dataset, constructed by collecting 36,056 tweets from Twitter via a stream API using a curated list of CB keywords, was meticulously preprocessed to remove noise, and standardized to address linguistic variabilities. Experimental results indicate that the LSTM model achieves the highest overall accuracy of 72%, thereby validating the potential of deep learning approaches for robust CB detection in Arabic social media texts.

Bouliche and Rezoug [40] introduces a novel dynamic graph neural network (DGNN) approach for detecting CB in Arabic social media by modelling entire comment sessions as dynamic temporal graphs. In these graphs, each node represents a comment and edges denote reply to relationships, allowing the model to capture both spatial and temporal dependencies without converting the data into static graphs. A custom message passing algorithm is used to update node embeddings, addressing challenges such as node deletion and index shifting. The method was evaluated on a dataset of 11,268 Arabic comments [41], from which 753 dynamic graphs were generated, demonstrating promising improvements in learning despite challenges like class imbalance and memory constraints.

Alzaqebah et al. [45] introduced an automated framework for CB detection tailored to addressing the challenges posed by imbalanced short text and various Arabic dialects present in dataset, called MSAUS. A novel approach was proposed within this framework to address the imbalance issue by employing a modified simulated annealing optimisation algorithm aimed at identifying an optimal subset of samples from the dominant class to equalise the training dataset. The evaluation of this method involved tradition ML algorithms, such as SVM, as well as DL algorithms, such as long short-term memory (LSTM) and bidirectional LSTM (BiLSTM). The researchers used three different offensive datasets from [27], [46], and [47]. Key performance metrics, including recall, accuracy, specificity, mean squared error and

TABLE 3. Summary of CB detection models, their performance on Arabic datasets, and associated limitations.

Year	Reference	Classifier Type	Classifier Used	Size	Platform	Features	Best Metrices	Limitation
2017	Haider et al. [20]	ML	NB, SVM	35,273	X	SentiStrength feature vector	SVM Recall 94.1%	Extremely imbalanced dataset, unavailable, with no details about the annotators.
2018	Haider et al. [35]	DL	FFNN	4,913	X	Word embedding	Accuracy 94.5%	Small imbalanced dataset
2019	Haider et al. [36]	Ensemble model	Bagging, boosting (KNN, SVM, NB)	34,890	X	N/A	Accuracy 93.3%	Extremely imbalanced dataset, unavailable, with no details about the annotators.
	Alharbi et al. [37]	Lexicon-based	PMI, chi-square, entropy	100,327	X, Microsoft-Flow, YouTube	N/A	PMI $F_{Avg.}$ 81%	Unavailable dataset, with no details about its balance.
	Mouheb et al. [38]	ML	NB	25,000	YouTube, X	TF-IDF	Accuracy 95.9%	Unavailable dataset, with no details about its balance or annotation process.
2021	Bashir & Bouguessa [39]	ML and DL	MNB, LSVC, LR, Ridge, BNB, RF, KNN and LSTM	36,056	X	TF and TF-IDF	LSTM Accuracy 72%	Unavailable dataset, using a sentiment analysis dataset.
2022	Shannag et al. [28]	ML	SVM, RF, XGBoost, DT, LR.	4,505	X	Word embedding, TF-IDF	SVM with word embedding Accuracy 86.3%	Small imbalanced dataset
	Albayari et al. [26]	ML	LR, SVM, RF, MNB	46,898	Instagram	N/A	SVM F1 69%	Dialectal imbalanced dataset
	Bouliche & Rezoug [40]	Graph Neural Network (GNN)	Dynamic Graph Neural Network (DGNN)	11,268 [41]	YouTube	WordPiece segmentation	Accuracy 90%	Imbalanced and abusive dataset
	Alhashmi et al. [42]	Consensus-based ensemble model	SDL, RF, SVM, ANN, XGBoost (XGB).	34,244	X, WhatsApp, Vine, Instagram Packet	N/A	Accuracy 88.5%	Translated imbalanced dataset.
2023	Khairy et al. [12]	ML, ensemble ML	Linear SVC, LR, K-neighbors. bagging-RF, voting, boosting-Adaboost	12,000	Facebook, X and YouTube	TF-IDF	LR Accuracy 65.5%	No details about the annotators.
	Alduailaj et al. [13]	ML	SVM, NB	30,000	X, YouTube	TF-IDF, BoW	SVM with TF-IDF Accuracy 95.4%	Imbalanced, with no details about the annotators.
	Saadi et al. [43]	ML	SVM with Cuckoo search	17,748[29]	X	TF-IDF Count vectorizer	F1 91.3%	Unavailable, imbalanced dataset
	Mursi et al. [44]	ML	MLP, SVM	4,140	X	TF-IDF	SVM Accuracy 92%	Small, unavailable dataset.

sensitivity, were utilised to indicate the effectiveness of the framework on communication platforms. The findings suggest that the proposed framework enhances the effectiveness of the algorithms tested, with BiLSTM emerging as the most effective method for CB classification.

AlMutawa and Faisal [48] introduces a DL framework for detecting CB in Arabic YouTube comments using a Convolutional Long Short-Term Memory (CLSTM) model. The key contribution lies in its tailored approach to address the unique challenges posed by the Arabic language by

TABLE 3. *(Continued.)* Summary of CB detection models, their performance on Arabic datasets, and associated limitations.

Year	Reference	Classifier Type	Classifier Used	Size	Platform	Features	Best Metrices	Limitation
	Alzaqebah et al.[45]	ML, DL	SVM, NB, RF, KNN, LR, LSTM, BiLSTM, CNN	Dataset 1: 15,049 [27] Dataset 2: 4,000 [46] Dataset 3: 10, 000 [47]	YouTube, Facebook, X YouTube, X	TF-IDF	MSAUS Accuracy 88.00%	All offensive dataset
	Almutawa & Faisal [48]	DL	Convolutional LSTM (CLSTM)	365 comments from 16 YouTube videos	YouTube	N/A	Accuracy 90%	Small, unavailable dataset, with no details about its balance or annotation process.
	Alfarah et al. [49]	Transformer-based	BERT	14,708	YouTube, X	N/A	AraBERTv2 large F1 84.58%	Imbalanced dataset
	Sadek et al. [50]	ML, Transformer-based, and LLM-based	NB, AraBERT, ChatGPT	46,743	X and YouTube	TF-IDF/BOW for NB;	AraBERT Accuracy 91%	Unavailable, dialectal imbalanced dataset
	Hussain et al. [51]	ML, ensemble ML	Decision tree, K-nearest neighbors, bagging ensemble, adaptive boosting, XGBoost, LR, RF, multinomial NB, SVM and NN	Used three datasets [12] [13][26]	Facebook, X, Instagram	Count vectorizer, TF-IDF	SVC with TF-IDF Accuracy 98.49%	Imbalanced datasets, use traditional feature extraction methods, and lack of DL models
2024	Musleh et al. [31]	ML	SVM, NB, RF, LR, AdaBoost, CatBoost, LightGBM, Bagging, XGBoost	9,000	X	Pre-processed text with TF-IDF and N-gram features	XGBoost with TF-IDF Accuracy 89.95%	Unavailable dataset, with no details about its balance or annotation process.
	Benaissa et al. [34]	PLM-based and DL	BERT, RoBERTa, DistilBERT	6,186 [52]	Aljazeera	Contextualized word embeddings	RoBERTa, Recall 90%	Obscene dataset
	Albayari et al. [53]	DL /Hybrid Model	CNN–BLSTM–GRU	AA-MCU (46,898) AA-BCU (46,898) AA-BCB (57,438) DataSet2 [27]: 15,050	Instagram	N/A	CNN–BLSTM–GRU in AA-BCU/AA-BCB Accuracy ~85%	Dialectal imbalanced dataset, limited hyperparameter optimization
	Albayari et al. [54]	DL /Ensemble	CNN, Bi-LSTM, Bi-GRU, Hybrid Bi-LSTM-LSTM, CNN-Bi-GRU, CNN-Bi-LSTM, Bi-LSTM-Bi-GRU AND Ensemble Stacking (Random Forest, MLP, LR, NB, SVM, KNN as meta-learners)	46,898 AA-MCU [26]	Instagram	Word embeddings	Stacking DL with RF Accuracy 94.73%	Dialectal imbalanced dataset, computational complexity of Ensemble Models
	Daraghmi et al. [55]	DL/ Hybrid	CNN, Bi-LSTM, GRU	105,371	Facebook, Twitter, Instagram	Stacked Word Embeddings (GloVe and FastText)	CNN-BiLSTM-GRU Accuracy 98.83%	Potential bias in annotation & lexicon creation.
	Azzeh et al. [56]	DL /Hybrid Model	AraCB (integrating CNN, multi-head self-attention, ResNet)	4,505ArCyC [28]	X	skip-gram, word2vec embeddings, and ResNet	Accuracy 82.3%	Limited dataset scope & generalization issues
	Mahdi et al. [57]	Transformer-based	E-BERT	30,000 [13]	X	WordPiece tokenizer	Accuracy 98.45%	Applies pre-trained BERT (English) models

leveraging DL techniques that capture both local (via convolutional layers) and sequential (via LSTM layers) textual patterns. The study is supported by a meticulously curated dataset comprising 365 comments extracted from 16 YouTube videos. These comments were collected using Python scripts interfacing with the YouTube API and then rigorously pre-processed to standardize the text for analysis. Experimental evaluation revealed that the proposed CLSTM model achieved an accuracy of 90%, demonstrating its robust capability in distinguishing between CB and non-CB content.

C. HYBRID AND ENSEMBLE MODELS

Khairy et al. [12] examined the effectiveness of detecting CB and offensive language in Arabic text by employing three single-learner ML methods (LR, linear SVC and K-neighbors) and three ensemble ML methods (boosting with AdaBoost), bagging with RF and voting. Ensemble ML is described as a meta-learning strategy that combines predictions from multiple single-learner classifiers to enhance performance over those of any individual classifier. Ensemble ML approaches generally outperform single-learner methods, particularly the voting ensemble classifier.

Alhashmi et al. [42] presented a CB detection model known as a consensus-based ensemble. A variety of diverse classifiers, including both traditional techniques of ML and DL, were trained using a translated labelled Arabic CB dataset sourced from across five distinct platforms. Various ML algorithms were examined in the comparison, including SVM, Sequential Deep Learning (SDL), Artificial Neural Network (ANN), RF and XGBoost (XGB). The outputs from these classifiers were merged through a consensus-based decision-making process, which used the F1 score of each classifier to determine their ranking. The proposed model achieved an overall improvement of 1.3% compared to the best-trained classifier.

Albayari et al. [53] in their second study introduces a novel hybrid DL model (CNN–BLSTM–GRU) for Arabic CB detection, rigorously evaluated against several architectures including LSTM, GRU, CNN-LSTM, CNN-BLSTM, LSTM-ATT, and LSTM-TCN using two datasets. The primary dataset [26] was utilized in three distinct scenarios: (1) the original multiclass categorization (Positive, Bullying, Neutral, Toxic), designated as AA-MCU (Authors' Names-Multiclass-Unbalanced); (2) a binary classification, where comments were labelled as either bullying or non-bullying, designated as AA-BCU (Authors' Names-Binary Class-Unbalanced); and (3) a balanced version of the binary dataset, obtained via an oversampling method, designated as AA-BCB (Authors' Names-Binary Class-Balanced). The models were assessed using standard evaluation metrics—accuracy, precision, recall, and F1-score—with the proposed hybrid model consistently outperforming the other architectures. Conspicuously, the CNN–BLSTM–GRU achieved the highest overall accuracy (up to 83% on certain datasets) and delivered competitive precision, recall, and F1-scores, with an F1-score of approximately 91% in multiclass classification scenarios. These findings underscore the model's capability to effectively capture the complex linguistic nuances of Arabic and its potential for real-world integration into social media platforms to mitigate CB.

In their third study, Albayari et al. [54] tackle the escalating issue of CB in Arabic digital communications by proposing an advanced DL-based solution. The authors systematically evaluate seven individual DL models for Arabic CB detection, including CNN, Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional GRU (Bi-GRU), Hybrid Bi-LSTM-LSTM, CNN-Bi-GRU, CNN-Bi-LSTM, and Bidirectional LSTM-Bidirectional GRU (Bi-LSTM-Bi-GRU), assessing each model using standard performance metrics such as accuracy, precision, recall, and F1-score. To further enhance detection performance, the study implements an ensemble stacking approach, integrating predictions from the top-performing models through a meta-learner classifier. Specifically, three ensemble configurations are tested: one combining the two best models (Bi-LSTM and Bi-LSTM-Bi-GRU), another incorporating the top four models, and a final ensemble stacking all seven models. The results reveal that the ensemble model utilizing a Random Forest meta-learner achieves the highest accuracy of 94.73%, surpassing both individual deep learning models and other stacking ensembles employing meta-learners such as MLP, LR, NB, SVM, and KNN. These findings underscore the effectiveness of combining diverse DL models through ensemble stacking to significantly improve the accuracy of CB detection in Arabic textual data.

Daraghmi et al. [55] introduces an integrated hybrid DL model tailored for detecting CB in Arabic textual data. The authors propose a novel approach by combining three robust neural network architectures—CNN, Bi-LSTM, and Gated Recurrent Units (GRU)—alongside stacked word embeddings (GloVe and FastText) to effectively capture both localized textual patterns and long-term contextual dependencies. The model is trained and evaluated on a comprehensive dataset, which integrates six benchmark datasets sourced from platforms such as Facebook, Twitter, and Instagram [26], supplemented by a custom-developed Arabic CB lexicon. Among these datasets, the first and fourth datasets, referenced from GitHub, do not explicitly mention their specific content or purpose. The second dataset refers to the Arabic Sentiment Twitter Corpus, which contains labelled positive and negative tweets. The third dataset pertains to the Arabic Levantine Hate Speech Dataset, focusing on hate speech and abusive language detection. The sixth dataset, referred to as the Offensive Dataset, contains manually labelled offensive and non-offensive comments. Remarkably, only one of these datasets is explicitly designed for CB detection, while the others are adapted for this purpose through pre-processing and labelling. The hybrid architecture employs CNN layers for extracting local features, BiLSTM layers to model bidirectional long-term dependencies, and GRU layers for efficient sequence modelling, all enhanced by stacked word embeddings to enrich semantic representation. The proposed CNN-BiLSTM-GRU model achieves remarkable performance, with an accuracy of 98.83%, alongside high scores in recall, precision, specificity, and F1-measure. The findings underscore that the integration of multiple DL architectures with advanced text representation techniques significantly enhances the detection of CB in Arabic online communications, offering a robust solution to this growing challenge.

D. PRETRAINED TRANSFORMER MODELS

Benaissa et al. [34] propose a class imbalance-sensitive approach that leverages fine-tuned PLMs for the detection of CB in both English and Arabic datasets. The authors address the critical challenge of class imbalance by employing a multi-faceted strategy comprising text-level data augmentation via Easy Data Augmentation, cost-sensitive learning, and the novel application of focal loss—a loss function typically utilized in computer vision—to improve the detection of the minority CB class. In this framework, fine-tuned Transformer-based models (BERT, RoBERTa, and DistilBERT) are used as feature extractors, generating contextual embeddings that are subsequently processed by multi-channel convolutional neural networks and bidirectional LSTMs for classification. The approach was rigorously evaluated on two English datasets (Formspring and Impermium) and one Arabic dataset (Aljazeera) [52]. Experimental results indicate that the proposed models achieve significant improvements in key performance metrics, with recall values ranging from 78% to 96% and notable enhancements in accuracy, precision, F1-score, and AUC compared to conventional methods. These findings underscore the effectiveness of integrating PLMs with advanced imbalance-sensitive techniques for robust CB detection across diverse linguistic contexts.

Alfarah et al. [49] presented a methodology utilising Arabic Bidirectional Encoder Representations from Transformers (BERT) models to label the issue of CB in Arabic online social networks. Comparative analysis revealed that Arabic BERT models outperformed conventional ML and DL methods. Fine-tuning of the Arabic BERT models was conducted, followed by three rounds of experiments. The AraBERT v2 large model achieved the best AUC and F1 scores, reaching 85.94% and 84.58%, respectively.

Sadek et al. [50] investigate the detection of CB in Arabic by utilizing a multi-dialect dataset consolidated from various social media platforms, including Twitter and YouTube. Notably, this dataset is not publicly available. The study addresses the challenges inherent in Arabic CB detection by comparing the performance of three classification approaches: a classical Naïve Bayes model, the transformer based AraBERT model, and ChatGPT. Experimental results demonstrate that AraBERT consistently outperforms both Naïve Bayes and ChatGPT, achieving superior accuracy, precision, recall, and F1-scores—particularly on the Twitter dataset, where AraBERT reached an accuracy of 91%. Nonetheless, the study also emphasizes that the multi-dialect dataset introduces additional challenges, suggesting that ensemble methods tailored to individual dialects may be necessary to enhance generalizability and performance in real-world applications.

The principal contribution of Azzeh et al. [56] study is the development of AraCB, an Arabic CB detection system that combines CNN, multi-head self-attention, and ResNet architectures to effectively capture the semantic and contextual subtleties of Arabic text. AraCB utilizes a custom-trained skip-gram word2vec model in conjunction with positional encoding to produce dense word embeddings that are subsequently refined through multi-head attention layers. The system was evaluated using the publicly available ArCybC dataset [28], comprising 4,505 tweets collected from X. The experimental results reveal that AraCB significantly surpasses conventional classification methods such as SVM, LR, RF, RNN, and XGB; specifically, it achieves improvements of 16.5% in average accuracy, 29.89% in recall, 18.4% in precision, 26.93% in F1-score, and 16.66% in AUC. For example, with a word2vec representation of vector size 100 and a flatten aggregation approach, AraCB attained an accuracy of 82.3%, recall of 77.3%, precision of 78.2%, F1-score of 77.7%, and an AUC of 89.2%. These outcomes underscore the robustness of AraCB and its considerable promise for the accurate detection of CB in Arabic text, despite the complex challenges posed by the language.

Mahdi et al. [57] introduce an end-to-end transformer-based model, referred to as E-BERT, designed to detect CB in Arabic social media content. Acknowledging the distinct linguistic challenges inherent to the Arabic language, such as dialectal diversity, intricate morphology, and script complexity, the authors adapt an English-pretrained BERT model by fine-tuning it on an Arabic-specific dataset. To address morphological complexities, the model utilizes the WordPiece tokenizer, customized for Arabic, which decomposes words into meaningful subword units. Additionally, special tokens ([CLS] and [SEP]) are incorporated to align with BERT's input format, and input sequences are standardized through padding or truncation to a fixed length. The enhanced model demonstrates exceptional performance, achieving an accuracy of 98.45%, precision of 99.17%, recall of 99.10%, and an F1-score of 99.14%. These results indicate a substantial improvement over baseline models and underscore the efficacy of cross-lingual transfer learning for Arabic text analysis. The study also addresses challenges such as vocabulary mismatch and dialectal variations when applying English-pretrained models to Arabic, highlighting how tailored tokenization and fine-tuning strategies can effectively mitigate these issues.

E. SENTIMENT AND REAL-TIME MODELS

Sentiment analysis is a natural language processing technique used to determine the emotional tone of text, categorizing it as positive, negative, or neutral. This method enables automated systems to detect emotions or hostility, which is particularly valuable in identifying CB behaviour. Alharbi et al. [37] proposed an automated method for detecting CB through sentiment analysis and lexicon-based techniques. Data were collected from YouTube comments, X API and Microsoft Flow, totalling approximately 100,327 tweets and comments gathered into a single file. Following data cleaning and pre-processing, the data were categorised into bullying and non-bullying categories, with classification performed by three individuals to ensure consideration of the majority opinion. After data preparation and configuration for lexicon generation, PMI, chi-square and entropy methods were implemented. The results indicate that the PMI method outperformed the entropy and chi-square methods in detecting CB.

Real-time filtering refers to the automated, immediate processing and categorization of content as it is generated, allowing for prompt intervention. In CB detection, real-time filtering can help prevent exposure to harmful content by instantly identifying and managing offensive messages based on predefined rules. Mouheb et al. [18] introduced a method for identifying CB within Arabic X streams. The proposed approach involves real-time monitoring of X messages, followed by pre-processing and noise reduction of tweets. Offensive messages are then detected and categorised based on their strength, with weights assigned to each offensive message. The proposed tool functions as a real-time filter designed to automatically detect and categorise CB messages. This filter can be implemented either on the user's device or integrated into the social media platform. Similar to antivirus software or spam filters, the tool operates according to user-defined rules. Upon detection of a CB message, the tool can take appropriate actions, such as deleting or hiding the message, issuing a warning notification or notifying third parties for professional intervention, depending on the severity of the bullying message. The experiments demonstrated that the proposed system successfully identified CB messages with near real-time efficiency.

Table 3 offers a detailed overview of various methodologies and classifiers used in the detection of CB within Arabic text datasets. It encompasses a range of ML, DL, and transformer based techniques applied across different platforms and datasets, highlighting the features utilized and the best performance metrics achieved by each approach. The table 3 underscores the diverse approaches in the detection of CB within Arabic texts. ML techniques, such as SVM and LR, have shown significant efficacy, particularly when combined with features like TF-IDF and word embeddings [12], [13], [28], [26]. DL models, especially LSTM, have also demonstrated high performance in capturing complex linguistic patterns [39]. Ensemble methods further enhance accuracy by integrating multiple classifiers [12], [42], [51]. These methodologies collectively contribute to the advancement of NLP in detecting CB, highlighting the importance of tailored approaches for Arabic language datasets.

V. RECENT CB DETECTION METHODS IN ENGLISH

Teng et al. [5] created an automated system to detect CB using two methods: traditional ML and transfer learning. Utilising the AMiCA dataset, which contains a substantial amount of CB context and structured annotations, was integral to the research. Various features such as textual analysis, static and contextual word embeddings, sentiment and emotional analysis, psycholinguistics, toxicity features and term lists were incorporated into the traditional ML approach. The utilisation of the transfer learning approach involved fine-tuning optimised versions of pre-trained language models, specifically DistilBert, Electra-small and DistilRoBerta. These models were chosen due to their faster training computation compared to their base forms. After fine-tuning, DistilBert outperformed traditional ML in achieving the highest F-measure. The study determined that transfer learning was the most effective method for improving performance with minimal effort, as it eliminated the need for feature engineering and resampling.

Mathur et al. [58] introduced a real-time system designed to detect CB on X by leveraging natural language processing and ML techniques. The system undergoes training on a dataset of CB-related tweets using various ML algorithms to compare their performances. Through tuning, RF emerged as the most effective algorithm. To enable real-time analysis, Selenium was utilised to extract tweets from specified X accounts while recording the timestamps of previously analysed tweets. Furthermore, an image captioning model is integrated to describe images posted on the account, contrasting them with user-provided captions to sift out spam content. The primary goal is to combat CB by offering a valuable tool for online platforms to identify and eliminate harmful content. The findings underscore the significance of ML algorithm selection and pre-processing techniques in enhancing CB detection performance on X.

Murshed et al. [59] introduced an efficient model for English CB classification and detection, integrating fuzzy adaptive equilibrium optimisation (FAEO) clustering-based topic modelling with an ensemble convolutional neural network (ECNN) to enhance the accuracy of CB detection. The study began with data cleansing in the pre-processing phase. Then, features were extracted using a TF-IDF, followed by the creation of word clusters by FAEO, based on the text data. Finally, ECNN was employed for classification across various CB categories, including sexism, insult, aggression and racism. The performance of the proposed ECNN model was assessed using two datasets from social media: the RW-CB-Twitter and CB Mendeley (CB-MNDLY) datasets. The proposed FAEO-ECNN model was compared to LSTM, BiLSTM, RNN and CNN-LSTM and demonstrated superior performance in comparison. However, there were limitations to this research. First, the implementation of ECNN for CB detection is currently restricted to the English language. Additionally, the FAEO-ECNN model is specifically designed to detect CB through text analysis, excluding other media forms, such as audio, images and videos, from its scope of detection.

Almomanu et al. [60] proposed a novel hybrid method employing DL models for feature extraction combined with ML classifiers to enhance the detection of CB in image-based content. Utilising pre-trained DL models such as InceptionV3, ResNet50 and VGG16 to extract features and then applying these features to classifiers like logistic regression and SVMs improves the ability to understand the intricate contexts in which CB occurs. This approach leverages the strengths of advanced visual recognition technologies to enhance the accuracy and efficiency of CB detection in

TABLE 4. Summary of techniques for detecting Arabic Cyberbullying.

Methods	Classifiers	References used
ML	SVM	[12] [13] [20][26][28] [31][36] [39][42][43][44][45][51][54]
	Ridge	[39]
	Naïve Bayes	[13] [20][31][36][38][45] [50][51][54]
	MNB	[26][39][51]
	BNB	[39]
	Logistic Regression	[12] [26][28][31][39][45][51][54]
	K-neighbors	[12] [39][51]
	Random Forest	[26][28][31][39][42][45][51][54]
	XGBoost	[28][31][42][51]
	Decision tree	[28][51]
Ensemble	Bagging	[12] [31][36][51]
	Voting	[12]
	Boosting-Adaboost	[12] [31][36][51]
DL	LSTM	[39][45][48][51][54]
	Bi-LSTM	[45][53][54][55]
	CLSTM	[46]
	KNN	[36][39][45][54]
	CNN	[45][53][54][55][56]
	ANN	[42]
	SDL	[42]
	FFNN	[35]
	Hybrid	[53][54][55][56]
	Dynamic Graph Neural Network	[40]
Transformer Based	BERT	[34][49][57]
	RoBERTa	[34]
	DistilBERT	[34]
	AraBERT	[50]
	ChatGPT	[50]

images. Although the hybrid technique causes additional computational expenses, these costs are justified by proven accuracy enhancements. The importance of efficiency is recognised as a key aspect that requires enhancement. With optimised implementations and infrastructure, handling the increased resource demands would be feasible.

Alam et al. [61] developed single and dual ensemble-based voting models to classify content as either 'offensive' or 'not-offensive'. To accomplish this objective, four ML classifiers and three ensemble models were employed in combination with two distinct feature extraction methods and various n-gram analyses. These techniques were applied to a dataset obtained from Twitter. The findings indicated that logistic regression and the bagging ensemble model achieved the best individually in detecting CB. Furthermore, the proposed single-layer ensemble (SLE) and double-layer ensemble (DLE) models achieved the highest performance of 96% accuracy when TF-IDF (Unigram) was employed as feature extraction in conjunction with K-fold cross-validation.

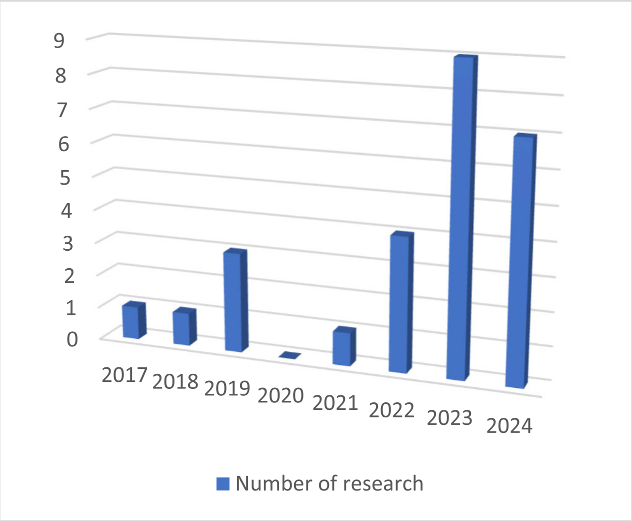


FIGURE 2. Number of Arabic Cyberbullying studies.

Rasool et al. [62] introduced a novel algorithm, AVOAGNN-CBDC, for CB detection that combines an African vulture optimisation algorithm with a graph neural network. The AVOAGNN-CBDC method involves several stages of data pre-processing and employs a FastText-based word embedding process to enhance its CB detection capabilities. The AVOA technique is then utilised for optimal parameter selection in the GNN model to improve classification performance. Experimental results conducted on a CB dataset demonstrated the superiority of the AVOAGNN-CBDC model against the LSTM, RNN, CNN and GNN methods.

Hussein et al. [63] introduced a method for identifying CB instances on X by employing graph mining and ML methodologies. Alongside traditional text-based features, such as TF.IDF and BOW, the study incorporated centrality measures extracted from the graph structure of interactions. Supervised learning techniques were applied, utilising ML algorithms trained on labelled data. Experiments conducted on a substantial X dataset validated the efficacy of the approach. The findings demonstrate a notable enhancement in the efficiency and accuracy of detecting CB on X through the integration of graph mining techniques with ML. In particular, employing the random forest model with positive features yielded promising results.

Plaza-Del-Arco et al. [64] investigated zero-shot learning with prompting as a method of detecting hate speech. This method explores the impact of zero-shot learning in detecting hate speech across three languages when there is a limitation of labelled data. The researchers conducted experiments using a variety of large language models and verbalisers across eight benchmark datasets. The research results underscore the potential of prompting for hate speech detection and demonstrate the significant impact of both prompt selection and the choice of language model to enhance the accuracy of the predictions in this task.

VI. DISCUSSION: CHALLENGES IN ARABIC CB DETECTION

Since the first research study on CB detection in Arabic was conducted in 2017 [20], scholarly attention to this issue has significantly increased. Between 2017 and 2022, only six studies were published on Arabic CB detection. However, in the past two years (2023–2024), nearly twenty studies have emerged, reflecting a rapid surge in interest, see Figure 2 for illustration. This growing attention is largely driven by the increasing recognition of the adverse mental health effects of CB, highlighting the urgent need for more effective detection mechanisms [23], [24], [25].

A. ADVANCEMENTS IN CB DETECTION MODELS

In recent years, researchers have developed various CB detection models that primarily leverage ML, DL, and, more recently, transformer-based techniques to enhance detection accuracy in Arabic texts, as summarized in Table 3. Among ML techniques, SVM remains the most commonly used due to its strong performance, followed by LR, while NB and RF are employed with similar frequency. In the domain of DL, LSTM and KNN have been widely utilized. Additionally, hybrid DL approaches have gained considerable traction due to their ability to achieve higher accuracy through the combination of multiple models. Ensemble learning techniques, which integrate multiple classifiers to enhance predictive performance, have also seen a marked increase in application. With the advent of transformer-based architectures, researchers have increasingly adopted BERT-based models for CB detection [49]. These models have demonstrated superior performance, whether utilized the pre-trained BERT model or fine-tuned PLMs on non-CB datasets [45]. Moreover, the features employed in Arabic CB detection models vary widely but commonly include TF-IDF, bag of words, word embeddings, and count vectorizers. These techniques are instrumental in transforming textual data into analysable formats, facilitating more effective classification. Furthermore, the performance of these models is typically evaluated using accuracy, F1-score, and recall, with numerous studies reporting high effectiveness based on accuracy measures. Some models have achieved accuracy and F1 scores around or above 90%, demonstrating their robustness in CB detection.

Additionally, these models have been trained on Arabic language datasets collected from diverse social media platforms, including Facebook, X, YouTube, and Instagram. The inclusion of multiple platforms underscores the widespread prevalence of CB across various digital communication channels, highlighting the need for robust detection methods. Currently, there are four publicly available datasets, all of which are manually annotated and specifically created for CB detection in Arabic. The first dataset, introduced in 2022, is relatively small, containing approximately 4,500 instances [28]. However, it is imbalanced, with a 0.38 class ratio, and includes data from three dialects: Egyptian, Gulf, and Levantine. The second dataset is also imbalanced, with a 0.3 class ratio, and is exclusively composed of Modern Standard Arabic (MSA) despite MSA not being the official spoken language of any Arabic-speaking country [13]. The third dataset is balanced but only features MSA dialect, limiting its applicability to informal Arabic texts [12]. The fourth and most extensive dataset contains approximately 50,000 instances and encompasses multiple Arabic dialects [26]. Due to its size and linguistic diversity, this dataset has the potential to serve as a future benchmark for evaluating Arabic CB detection methods, contributing to the advancement of Arabic NLP research.

B. CHALLENGES IN ARABIC CB DETECTION

Despite advancements in CB detection models, several key challenges persist, particularly regarding data availability, linguistic complexity, and sociocultural constraints. One of the primary challenges in developing Arabic CB detection models is the absence of a standardized benchmark dataset that comprehensively covers all Arabic dialects. As of 2022, the first publicly available CB dataset was introduced. Subsequently, three more datasets became publicly available, two of which primarily focus on MSA. The cost associated with manual dataset labelling and classification for CB detection significantly intensifies this challenge [65]. This resource limitation restricts the development of high-quality annotated datasets, which are essential for capturing the linguistic nuances necessary for accurate CB detection in the Arabic language. As a result, some CB detection models are trained on offensive or abusive language datasets rather than datasets specifically curated for CB detection [40], [45]. This reliance on non-specialized datasets may negatively impact classification accuracy and limit the generalizability of the models.

Another significant challenge in Arabic CB detection arises from the linguistic complexity of Arabic, as detection tools must effectively capture both formal (MSA) and informal (dialectal) language variations to ensure accurate analysis [66]. Arabic dialects vary considerably across regions, often leading to differences in word meanings. Some words that are neutral or positive in one country may be perceived as offensive in another. For instance, the term ‘‘Yetqalash’’ is used as a compliment in Yemen, whereas in Morocco, it is regarded as offensive [20]. Accordingly, manual annotation for dialectal datasets poses additional challenges, as annotators from diverse linguistic backgrounds may misinterpret words and expressions, leading to inconsistencies in labelling CB instances [27]. Another critical issue is bias in dialect annotation—since one dialect may dominate, subjective biases may be introduced, increasing the likelihood of misclassification, and reducing model reliability [21]. The lack of comprehensive, dialect-inclusive datasets remains a major limitation in advancing Arabic CB detection research. Furthermore, in Arabic NLP, pre-processing techniques such as removing diacritical marks are often essential. However, Arabic word meanings are influenced by diacritical, making it challenging to standardize pre-processing without losing linguistic nuances [67].

Beyond technical and linguistic challenges, social and cultural sensitivities also impact research on CB in

Arabic-speaking societies. Topics such as harassment and bullying are often considered sensitive issues, which may contribute to limited public discussions and fewer research initiatives [68]. These sociocultural factors, combined with linguistic complexities and dataset limitations, significantly hinder the development of effective CB detection tools for Arabic content.

C. FUTURE DIRECTIONS AND RESEARCH GAPS

Despite recent advancements in CB detection, several research gaps remain, presenting opportunities for further exploration. One notable gap is the limited integration of multimodal data in CB detection models. Current research primarily focuses on text-based detection, often overlooking other modes of communication, such as images, videos, and audio. Given that CB can occur through multiple media types, there is a pressing need to develop multimodal detection models that can effectively analyse diverse online interactions, as demonstrated by Zeng et al. [69]. A comprehensive review of existing literature indicates that, as of February 2025, no studies have explored multimodal approaches in Arabic CB detection, highlighting a significant research gap. MohammedJany et al. employed DL techniques, utilizing the VGG16 pre-trained model for bullying detection in images and XLM-RoBERTa with a BiGRU model for text-based bullying detection [70]. Their approach achieved reasonable accuracy, suggesting that adapting this methodology to the Arabic context could yield promising results.

Another critical limitation is the restricted generalizability of CB detection models across Arabic dialects and informal online language. Most existing datasets are not uniformly representative of the wide range of Arabic dialects spoken across different regions, nor do they fully capture the informal variations used in online interactions. This lack of dialectal inclusivity makes it challenging to develop robust CB detection models that can effectively analyse the full spectrum of Arabic CB content. Addressing this gap requires curating dialect-inclusive datasets and designing models capable of handling both MSA and diverse dialectal variations or utilizing the Dial2MSA-Verified dataset as a standardized resource [71]. Another promising area for future research involves utilizing weakly supervised models by leveraging LLMs and PLMs to reduce dependence on large, labelled datasets, thereby enhancing the adaptability and scalability of CB detection models [72].

Additionally, a notable gap identified in Table 4 is the underutilization of advanced DL techniques, such as CNNs and Arabic LLMs, despite their superior feature extraction capabilities [55]. Moreover, the concentration of research on traditional ML techniques suggests that integrating DL and transformer-based advancements could further enhance CB detection performance. Moreover, ensemble methods have been applied in a limited number of studies, indicating an opportunity for broader evaluation across different CB detection contexts. Given that LLMs improve contextual understanding, they can enhance classification accuracy by capturing dialectal variations and nuanced language usage [73], [74]. The adoption of transformer-based approaches represents a paradigm shift in CB detection, particularly through Arabic PLMs. These models leverage context-aware representations and deep contextual embeddings, which would significantly improve classification accuracy in Arabic texts [75]. Future research should focus on exploring underutilized DL methods, assessing their effectiveness in diverse CB detection scenarios, and integrating them with emerging NLP advancements to enhance model robustness and generalizability [76], [77].

Additionally, there is a lack of real-time detection capabilities in current CB detection systems. Most existing models are trained offline and rarely focus on real-time detection, limiting their effectiveness in dynamic social media environments where CB occurs in rapid and evolving contexts. Real-time monitoring is crucial for implementing proactive intervention strategies and mitigating the negative impact of CB incidents before they escalate [78]. The development of efficient real-time CB detection systems is a key priority for future research, ensuring timely identification and intervention in CB incidents.

Addressing these gaps could involve developing more sophisticated models that can analyse multimodal content, enhancing dialectal and colloquial language recognition in text processing and focusing on the real-time application of these models. Such advancements would significantly bolster the robustness and applicability of CB detection systems, making them more effective across diverse and rapidly changing online platforms.

VII. CONCLUSION

This study examined the advancements and challenges in Arabic CB detection, offering suggestions for future research based on superior models that have been successfully applied in CB detection for other languages. This comprehensive review analysed four datasets and 26 research papers specifically dedicated to Arabic CB detection techniques. While significant progress has been made using ML and DL to address CB in Arabic texts, several gaps remain, limiting the effectiveness and applicability of current models. The primary challenges identified include the need for models to handle multimodal data, enabling detection across various media types beyond text, thereby providing a more comprehensive understanding of CB patterns. Additionally, current models lack generalizability due to the diverse dialects and informal language variations present in the Arabic online sphere, which are not uniformly addressed by existing datasets. Another critical issue is the lack of real-time detection capabilities, which are essential for immediate intervention in CB incidents. Furthermore, the underutilization of advanced DL techniques, such as CNNs and LLMs, suggests an opportunity to integrate more sophisticated technologies to enhance detection accuracy and processing capabilities.

To address these limitations, future research should focus on developing more robust CB detection systems that can analyse multimodal content, adapt to linguistic variations, and operate in real time. Leveraging LLMs, PLMs, and embedding techniques will be crucial in enhancing model performance [79]. These improvements are essential for creating safer online environments and protecting users from the harmful effects of CB, thereby advancing the field of Arabic CB detection towards dynamic, adaptive, and responsive solutions that keep pace with the evolving nature of online interactions.

REFERENCES

[1] L. Pokhun and Y. M. Chuttur, ''Can machine learning really detect cyber-bullying?'' *Int. J. Bullying Prevention*, vol. 2023, pp. 1–20, Aug. 2023, doi: 10.1007/s42380-023-00191-9.

[2] S. Kemp. (2025). *Digital 2025: Global Overview Report*. Meltwater. Accessed: Mar. 4, 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-global-overview-report

[3] S. Kemp. (2025). *Digital 2025: Saudi Arabia*. Meltwater. Accessed: Mar. 4, 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-saudi-arabia

[4] A. I. Al-Ghadir and A. M. Azmi, ''A study of Arabic social media users—Posting behavior and author's gender prediction,'' *Cognit. Comput.*, vol. 11, no. 1, pp. 71–86, Feb. 2019, doi: 10.1007/s12559-018-9592-7.

[5] T. H. Teng and K. D. Varathan, ''Cyberbullying detection in social networks: A comparison between machine learning and transfer learning approaches,'' *IEEE Access*, vol. 11, pp. 55533–55560, 2023, doi: 10.1109/ACCESS.2023.3275130.

[6] N. Alsunaidi, S. Aljbali, Y. Yasin, and H. Aljamaan, ''Arabic cyberbullying detection using machine learning: State of the art survey,'' in *Proc. 27th Int. Conf. Eval. Assessment Softw. Eng.*, Jun. 2023, pp. 499–504, doi: 10.1145/3593434.3593968.

[7] M. Dadvar, ''Experts and machines united against cyberbullying,'' Ph.D. thesis, University of Twente, Enschede, The Netherlands, 2014, doi: 10.3990/1.9789036537391.

[8] M. Khairy, T. M. Mahmoud, and T. Abd-El-Hafeez, ''Automatic detection of cyberbullying and abusive language in Arabic content on social networks: A survey,'' *Proc. Comput. Sci.*, vol. 189, pp. 156–166, Jan. 2021, doi: 10.1016/j.procs.2021.05.080.

[9] R. Bayari, S. Abdullah, and S. A. Salloum, ''Cyberbullying classification methods for Arabic: A systematic review,'' in *Proc. Int. Conf. Artif. Intell. Comput. Vis. (AICV).*, in Advances in Intelligent Systems and Computing, Cham, Switzerland, A. E. Hassanien, Ed. Springer, Jan. 2021, pp. 375–385, doi: 10.1007/978-3-030-76346-6_35.

[10] M. Arif, ''A systematic review of machine learning algorithms in cyberbullying detection: Future directions and challenges,'' *J. Inf. Secur. Cybercrimes Res.*, vol. 4, no. 1, pp. 01–26, Jun. 2021, doi: 10.26735/gbtv9013.

[11] I. Bensalem, P. Rosso, and H. Zitouni, ''Toxic language detection: A systematic review of Arabic datasets,'' *Expert Syst.*, vol. 41, no. 8, pp. 1–30, Aug. 2024, doi: 10.1111/exsy.13551.

[12] M. Khairy, T. M. Mahmoud, A. Omar, and T. Abd El-Hafeez, ''Comparative performance of ensemble machine learning for Arabic cyberbullying and offensive language detection,'' *Lang. Resour. Eval.*, vol. 58, no. 2, pp. 695–712, Jun. 2024, doi: 10.1007/s10579-023-09683-y.

[13] A. M. Alduailaj and A. Belghith, ''Detecting Arabic cyberbullying tweets using machine learning,'' *Mach. Learn. Knowl. Extraction*, vol. 5, no. 1, pp. 29–42, Jan. 2023, doi: 10.3390/make5010003.

[14] D. J. Pepler and W. Craig, ''Making a difference in bullying,'' LaMarsh, Tech. Rep. 59, 2000.

[15] J. Sui, ''Understanding and fighting bullying with machine learning,'' Ph.D thesis, Dept. Comput. Sci., University of Wisconsin-Madison, Madison, WI, USA, 2015.

[16] S. Bauman, ''Types of cyberbullying,'' in *Cyberbullying: What Counselors Need To Know*. The American Counseling Association, 2015, pp. 53–58.

[17] T. H. Teng, K. D. Varathan, and F. Crestani, ''A comprehensive review of cyberbullying-related content classification in online social media,'' *Expert Syst. Appl.*, vol. 244, Jun. 2024, Art. no. 122644, doi: 10.1016/j.eswa.2023.122644.

[18] D. Mouheb, M. H. Abushamleh, M. H. Abushamleh, Z. A. Aghbari, and I. Kamel, ''Real-time detection of cyberbullying in Arabic Twitter streams,'' in *Proc. 10th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS)*, Jun. 2019, pp. 1–5.

[19] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, and I. Trancoso, ''Automatic cyberbullying detection: A systematic review,'' *Comput. Hum. Behav.*, vol. 93, pp. 333–345, Apr. 2019, doi: 10.1016/j.chb.2018.12.021.

[20] B. Haidar, M. Chamoun, and A. Serhrouchni, ''Multilingual cyberbullying detection system: Detecting cyberbullying in Arabic content,'' in *Proc. 1st Cyber Secur. Netw. Conf. (CSNet)*, Oct. 2017, pp. 1–8.

[21] H. Bouamor, N. Habash, M. Salameh, W. Zaghouani,O. Rambow, D. Abdulrahim, O. Obeid, S. Khalifa, F. Eryani, A. Erdmann, and K. Oflazer. (2018). *The MADAR Arabic Dialect Corpus and Lexicon*. [Online]. Available: http://www.ustar-consortium.com/

[22] X. Xu, B. Yao, Y. Dong, S. Gabriel, H. Yu, J. Hendler, M. Ghassemi, A. K. Dey, and D. Wang, ''Mental-LLM: Leveraging large language models for mental health prediction via online text data,'' in *Proc. ACM Interact. Mob Wearable Ubiquitous Technol.*, Mar. 2024, vol. 8, no. 1, pp. 1–32, doi: 10.1145/3643540.

[23] S. Bansal, N. Garg, J. Singh, and F. Van Der Walt, ''Cyberbullying and mental health: Past, present and future,'' *Frontiers Psychol.*, vol. 14, pp. 1–19, Jan. 2024, doi: 10.3389/fpsyg.2023.1279234.

[24] N. Alrasheed, S. Nishat, A. Bin Shihah, A. Alalwan, and H. Jradi, ''Prevalence and risk factors of cyberbullying and its association with mental health among adolescents in Saudi Arabia,'' *Cureus*, vol. 14, no. 12, pp. 1–10, Dec. 2022, doi: 10.7759/cureus.32806.

[25] F. H. A. Shibly and U. Sharma, ''Detection of cyberbullying in social media to control users' mental health issues using recurrent neural network architectures,'' *J Pharm Negat Results*, vol. 13, no. 3, pp. 1–12, Jan. 2022, doi: 10.47750/pnr.2022.13.s03.072.

[26] R. Albayari and S. Abdallah, ''Instagram-based benchmark dataset for cyberbullying detection in Arabic text,'' *Data*, vol. 7, no. 7, p. 83, Jun. 2022, doi: 10.3390/data7070083.

[27] A. Alakrot, L. Murray, and N. S. Nikolov, ''Dataset construction for the detection of anti-social behaviour in online communication in Arabic,'' *Proc. Comput. Sci.*, vol. 142, pp. 174–181, Jan. 2018, doi: 10.1016/j.procs.2018.10.473.

[28] F. Shannag, B. H. Hammo, and H. Faris, ''The design, construction and evaluation of annotated Arabic cyberbullying corpus,'' *Educ. Inf. Technol.*, vol. 27, no. 8, pp. 10977–11023, Sep. 2022, doi: 10.1007/s10639-022-11056-x.

[29] S. Almutiry and M. A. Fattah, ''Arabic CyberBullying detection using Arabic sentiment analysis,'' *Egyptian J. Lang. Eng.*, vol. 8, no. 1, pp. 39–50, Apr. 2021.

[30] M. Alrougi, G. Alamoudi, and H. Algamdi, ''ArCBDs: A corpus for cyberbullying detection of Arabic tweets,'' *IJARCCE*, vol. 13, no. 1, pp. 146–154, Jan. 2024, doi: 10.17148/ijarcce.2024.13121.

[31] D. Musleh, A. Rahman, M. A. Alkherallah, M. K. Al-Bohassan, M. M. Alawami, H. A. Alsebaa, J. A. Alnemer, G. F. Al-Mutairi, M. I. Aldossary, D. A. Aldowaihi, and F. Alhaidari, ''A machine learning approach to cyberbullying detection in Arabic tweets,'' *Comput., Mater. Continua*, vol. 80, no. 1, pp. 1033–1054, 2024, doi: 10.32604/cmc.2024.048003.

[32] A. Dewani, M. A. Memon, and S. Bhatti, ''Cyberbullying detection: Advanced preprocessing techniques & deep learning architecture for Roman Urdu data,'' *J. Big Data*, vol. 8, no. 1, pp. 1–20, Dec. 2021, doi: 10.1186/s40537-021-00550-7.

[33] A. H. Alduailej and M. B. Khan, ''The challenge of cyberbullying and its automatic detection in Arabic text,'' in *Proc. Int. Conf. Comput. Appl. (ICCA)*, Sep. 2017, pp. 389–394.

[34] A. R. Benaissa, A. Harbaoui, and H. H. Ben Ghezala, ''Class imbalance-sensitive approach based on PLMs for the detection of cyberbullying in English and Arabic datasets,'' *Behav. Inf. Technol.*, vol. 2024, pp. 1–18, Feb. 2024, doi: 10.1080/0144929x.2024.2313142.

[35] B. Haidar, M. Chamoun, and A. Serhrouchni, ''Arabic cyberbullying detection: Using deep learning,'' in *Proc. 7th Int. Conf. Comput. Commun. Eng. (ICCCE)*, Kuala Lumpur, Malaysia, Sep. 2018, pp. 284–289.

[36] B. Haidar, M. Chamoun, and A. Serhrouchni, ''Arabic cyberbullying detection: Enhancing performance by using ensemble machine learning,'' in *Proc. Int. Conf. Internet Things (iThings) IEEE Green Comput. Commun. (GreenCom) IEEE Cyber, Phys. Social Comput. (CPSCom) IEEE Smart Data (SmartData)*, Jul. 2019, pp. 323–327.

[37] B. Y. Alharbi, M. S. Alharbi, N. J. Alzahrani, M. M. Alsheail, J. F. Alshobaili, and D. M. Ibrahim, "Automatic cyber bullying detection in Arabic social media," *Int. J. Eng. Res. Technol.*, vol. 12, no. 12, pp. 2330–2335, Dec. 2019.

[38] D. Mouheb, R. Albarghash, M. F. Mowakeh, Z. A. Aghbari, and I. Kamel, "Detection of Arabic cyberbullying on social networks using machine learning," in *Proc. IEEE/ACS 16th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2019, pp. 1–5.

[39] E. Bashir and M. Bouguessa, "Data mining for cyberbullying and harassment detection in Arabic texts," *Int. J. Inf. Technol. Comput. Sci.*, vol. 13, no. 5, pp. 41–50, Oct. 2021, doi: 10.5815/ijitcs.2021.05.04.

[40] A. Bouliche and A. Rezoug, "Detection of cyberbullying in Arabic social media using dynamic graph neural network," in *Proc. Tunisian-Algerian Joint Conf. Appl. Comput., Constantine, Algeria*, Constantine, Algeria, 2022, pp. 1–11.

[41] H. Mubarak, K. Darwish, and W. Magdy, "Abusive language detection on Arabic social media," in *Proc. 1st Workshop Abusive Lang. Online*, Aug. 2017, pp. 52–56.

[42] A. A. Alhashmi and A. A. Darem, "Consensus-based ensemble model for Arabic cyberbullying detection," *Comput. Syst. Sci. Eng.*, vol. 41, no. 1, pp. 241–254, 2022, doi: 10.32604/csse.2022.020023.

[43] M. Q. Saadi and B. N. Dhannoon, "Arabic cyberbullying detection using support vector machine with cuckoo search," *Iraqi J. Sci.*, vol. 64, no. 10, pp. 5322–5330, Oct. 2023, doi: 10.24996/ijs.2023.64.10.37.

[44] K. T. Mursi, A. Y. Almalki, M. M. Alshangiti, F. S. Alsubaei, and A. A. Alghamdi, "ArCyb: A robust machine-learning model for Arabic cyberbullying tweets in Saudi Arabia," *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, pp. 1059–1067, 2023.

[45] M. Alzaqebah, G. M. Jaradat, D. Nassan, R. Alnasser, M. K. Alsmadi, I. Almarashdeh, S. Jawarneh, M. Alwohaibi, N. A. Al-Mulla, N. Alshehab, and S. Alkhushayni, "Cyberbullying detection framework for short and imbalanced Arabic datasets," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 8, Sep. 2023, Art. no. 101652, doi: 10.1016/j.jksuci.2023.101652.

[46] S. A. Chowdhury, H. Mubarak, A. Abdelalí, S. Jung, B. J. Jansen, and J. Salminen, "A multi-platform Arabic news comment dataset for offensive language detection," in *Proc. 12th Lang. Resour. Eval. Conf., Jul.*, May 2020, pp. 6203–6212.

[47] H. Mubarak, A. Rashed, K. Darwish, Y. Samih, and A. Abdelalí, "Arabic offensive language on Twitter: Analysis and experiments," in *Proc. ofthe 6th Arabic Natural Lang. Process. Workshop, Kyiv, Ukraine, Apr.*, Kyiv, Kyiv, Ukraine, Jan. 2020, pp. 126–135.

[48] S. S. AlMutawa and M. Faisal, "Deep learning for digital safety: Cyberbullying detection in Arabic YouTube content," in *Proc. 10th Int. Conf. Social Netw. Anal., Manage. Secur. (SNAMS)*, Nov. 2023, pp. 1–6, doi: 10.1109/snams60348.2023.10375399.

[49] M. E. AlFarah, I. Kamel, and Z. Al Aghbari, "Toward detection of Arabic cyberbullying on online social networks using Arabic BERT models," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2023, pp. 1–6, doi: 10.1109/isncc58260.2023.10323808.

[50] A. Sadek, M. I. Khalil, and C. Salama, "Detection of cyberbullying in Arabic using machine learning and ChatGPT," in *Proc. 5th Novel Intell. Lead. Emerg. Sci. Conf. (NILES)*, Oct. 2023, pp. 242–245, doi: 10.1109/niles59815.2023.10296682.

[51] O. K. E. Hussien, A. E. Aboutabl, and R. M. Y. Haggag, "Comparative performance of data mining techniques for cyberbullying detection of Arabic social media text," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 11s, pp. 392–400, Oct. 2023, doi: 10.17762/ijritcc.v11i11s.8167.

[52] J. W. Patchin and S. Hinduja, "Bullies move beyond the schoolyard: A preliminary look at cyberbullying," *Youth Violence Juvenile Justice*, vol. 4, no. 2, pp. 148–169, Apr. 2006, doi: 10.1177/1541204006286288.

[53] R. Albayari, S. Abdallah, and K. Shaalan, "Cyberbullying detection model for Arabic text using deep learning," *J. Inf. Knowl. Manage.*, vol. 2024, Jan. 2024, Art. no. 2450016, doi: 10.1142/s0219649224500163.

[54] R. Albayari, A. A. Al Shamsi, and M. Alrammal, "Enhancing cyberbullying detection in Arabic text through ensemble stacking models," *Engineered Sci.*, vol. 31, p. 1272, Apr. 2024, doi: 10.30919/es1272.

[55] E.-Y. Daraghmi, S. Qadan, Y.-A. Daraghmi, R. Yousuf, O. Cheikhrouhou, and M. Baz, "From text to insight: An integrated CNN-BiLSTM-GRU model for Arabic cyberbullying detection," *IEEE Access*, vol. 12, pp. 103504–103519, 2024, doi: 10.1109/ACCESS.2024.3431939.

[56] M. Azzeh, B. Alhijawi, A. Tabbaza, O. Alabboshi, N. Hamdan, and D. Jaser, "Arabic cyberbullying detection system using convolutional neural network and multi-head attention," *Int. J. Speech Technol.*, vol. 27, no. 3, pp. 521–537, Sep. 2024, doi: 10.1007/s10772-024-10118-4.

[57] M. A. Mahdi, S. M. Fati, M. A. G. Hazber, S. Ahamad, and S. A. Saad, "Enhancing Arabic cyberbullying detection with end-to-end transformer model," *Comput. Model. Eng. Sci.*, vol. 141, no. 2, pp. 1651–1671, 2024, doi: 10.32604/cmes.2024.052291.

[58] S. A. Mathur, S. Isarka, B. Dharmasivam, and C. D. Jaidhar, "Analysis of tweets for cyberbullying detection," in *Proc. 3rd Int. Conf. Secure Cyber Comput. Commun. (ICSCCC)*, May 2023, pp. 269–274, doi: 10.1109/ICSCCC58608.2023.10176416.

[59] B. A. H. Murshed, Suresha, J. Abawajy, M. A. N. Saif, H. M. Abdulwahab, and F. A. Ghanem, "FAEO-ECNN: Cyberbullying detection in social media platforms using topic modelling and deep learning," *Multimedia Tools Appl.*, vol. 82, no. 30, pp. 46611–46650, Dec. 2023, doi: 10.1007/s11042-023-15372-3.

[60] A. Almomani, K. Nahar, M. Alauthman, M. A. Al-Betar, Q. Yaseen, and B. B. Gupta, "Image cyberbullying detection and recognition using transfer deep machine learning," *Int. J. Cognit. Comput. Eng.*, vol. 5, pp. 14–26, Jan. 2024, doi: 10.1016/j.ijcce.2023.11.002.

[61] K. S. Alam, S. Bhowmik, and P. R. K. Prosun, "Cyberbullying detection: An ensemble based machine learning approach," in *Proc. 3rd Int. Conf. Intell. Commun. Technol. Virtual Mobile Netw. (ICICV)*, Feb. 2021, pp. 710–715, doi: 10.1109/ICICV50876.2021.9388499.

[62] H. A. Rasool, F. Aldolaimy, F. F. Hasan, A. H. Alsalamy, M. Saleem, A. H. Alkhayyat, and M. Sharma, "Evolutionary algorithm with graph neural network driven cyberbullying detection on low resource Asian languages," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 2023, pp. 1–16, Jul. 2023, doi: 10.1145/3609799.

[63] F. N. A. Hussein and H. J. Aleqabie, "Cyberbullying detection in Twitter conduction graph mining and machine learning," *J. Univ. Kerbala*, vol. 20, no. 2, pp. 1–14, 2023.

[64] F. M. Plaza-Del-Arco, D. Nozza, and D. Hovy, "Respectful or toxic? Using zero-shot learning with language models to detect hate speech," in *Proc. 7th Workshop Online Abuse Harms (WOAH)*, 2023, pp. 60–68.

[65] A. Alghamdi, X. Duan, W. Jiang, Z. Wang, Y. Wu, Q. Xia, Z. Wang, Y. Zheng, M. Rezagholizadeh, B. Huai, P. Cheng, and A. Ghaddar, "AraMUS: Pushing the limits of data and model scale for Arabic natural language processing," *Assoc. for Comput. Linguistics*, vol. 2023, pp. 2883–2894, Jan. 2023. [Online]. Available: https://commoncrawl.org

[66] B. S. Alzaidi, Y. Abushark, and A. I. Khan, "Arabic location named entity recognition for tweets using a deep learning approach," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 12, pp. 76–83, 2022, doi: 10.14569/ijacsa.2022.0131211.

[67] B. Hamzaoui, D. Bouchiha, and A. Bouziane, "A comprehensive survey on Arabic text classification: Progress, challenges, and techniques," *Brazilian J. Technol.*, vol. 8, no. 1, Feb. 2025, Art. no. e77611, doi: 10.38152/bjtv8n1-022.

[68] X. Liang, "The cause and influence of cyberbullying," *J. Educ., Humanities Social Sci.*, vol. 26, pp. 661–668, Mar. 2024.

[69] T. Li, Z. Zeng, Q. Li, and S. Sun, "Integrating GIN-based multimodal feature transformation and multi-feature combination voting for irony-aware cyberbullying detection," *Inf. Process. Manage.*, vol. 61, no. 3, May 2024, Art. no. 103651, doi: 10.1016/j.ipm.2024.103651.

[70] S. MohammedJany, C. B. R. Killi, S. Rafi, and S. Rizwana, "Detecting multimodal cyber-bullying behaviour in social-media using deep learning techniques," *J. Supercomput.*, vol. 81, no. 1, pp. 1–19, Jan. 2025, doi: 10.1007/s11227-024-06772-9.

[71] A. Khered, Y. Benkhedda, and R. Batista-Navarro, "Dial2MSA-Verified: A multi-dialect Arabic social media dataset for neural machine translation to modern standard Arabic," in *Proc. The 4th Workshop Arabic Corpus Linguistics (WACL)*, 2025, pp. 50–62. [Online]. Available: https://github.com/khered20/

[72] A. Alotaibi, F. Nadeem, and M. Hamdy, "Weakly supervised deep learning for Arabic tweet sentiment analysis on education reforms: Leveraging pre-trained models and LLMs with snorkel," *IEEE Access*, vol. 13, pp. 30523–30542, 2025, doi: 10.1109/ACCESS.2025.3541154.

[73] B. AlKhamissi, M. ElNokrashy, M. Alkhamissi, and M. Diab, "Investigating cultural alignment of large language models," in *Proc. 62nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, Aug. 2024, pp. 12404–12422. [Online]. Available: https://github.com/b

[74] B. Ogunleye and B. Dharmaraj, "Use of large language model for cyberbullying detection," *Analytics*, vol. 3no. 2, pp. 694–707, 2023, doi: 10.20944/preprints202306.1075.v1.

[75] L. Wang, N. Yang, X. Huang, L. Yang, R. Majumder, and F. Wei, "Improving text embeddings with large language models," 2023, *arXiv:2401.00368*.

[76] M. Mashaabi, S. Al-Khalifa, and H. Al-Khalifa, "A survey of large language models for Arabic language and its dialects," 2024, *arXiv:2410.20238*.

[77] H. Huang, F. Yu, J. Zhu, X. Sun, H. Cheng, D. Song, Z. Chen, A. Alharthi, B. An, J. He, Z. Liu, Z. Zhang, J. Chen, J. Li, B. Wang, L. Zhang, R. Sun, X. Wan, H. Li, and J. Xu, "AceGPT, localizing large language models in Arabic," 2023, *arXiv:2309.12053*.

[78] R. Shrestha and R. Dave, "Machine learning for identifying harmful online behavior: A cyberbullying overview," *J. Comput. Commun.*, vol. 13, no. 1, pp. 26–40, 2025, doi: 10.4236/jcc.2025.131003.

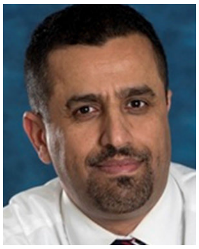
[79] D. Ottosson, "Cyberbullying detection on social platforms using large language models," Mid Sweden University, Sweden, U.K., Final Project, 2023.

HUDA ALJALAOUD received the master's degree (Hons.) in computer science from King Abdulaziz University, Jeddah, Saudi Arabia, in 2017. She is currently pursuing the Ph.D. degree in computer science with Edinburgh Napier University, Edinburgh, U.K.

Previously, she was a Teaching Assistant and later as a Lecturer, where she supported students and collaborated on projects related to natural language processing. She aspires to drive innovation in Arabic AI technologies, aiming to bridge linguistic gaps and promote inclusivity in digital spaces. She has published two peer-reviewed journal articles. Her research focuses on leveraging artificial intelligence in cybersecurity, specifically developing techniques for detecting cyberbullying in Arabic texts.

KIA DASHTIPOUR received the Ph.D. degree in computing science from the University of Stirling, Stirling, U.K., in 2019.

Since 2021, he has been a Lecturer in computing with Edinburgh Napier University, Edinburgh, U.K. He has authored or co-authored more than 120 peer-reviewed research articles, including numerous highly cited works in the areas of his research interests, which include multimodal signal and image processing, with a focus on audio-visual speech enhancement and sentiment and opinion mining.



AHMED Y. AL-DUBAI (Senior Member, IEEE) received the Ph.D. degree in computing from The University of Glasgow, Glasgow, U.K., in 2004.

In 2004, he joined the University of West London, London, U.K. In 2005, he joined Edinburgh Napier University, Edinburgh, U.K., where he became a Professor and the Research Lead of the Cybersecurity and IoT Group. He published widely in world-leading journals and prestigious international conferences. He has been involved with research in the areas of group communication algorithms, smart spaces, high-performance networks, and NLP. He served on several editorial boards of scholarly journals. He is a fellow of the Higher Academy, U.K. He was a recipient of the several academic awards and recognitions. He has served as a guest editor for more than 30 special issues in scholarly journals and chaired and co-chaired more than 40 international conferences/workshops.