

Research Article

Iraqi Dialect Cyberbullying Detection in Social Media Text Using Deep Learning Models

Arwa Alqudsi^{1,*}, Ruqaiya Jawd², Nazlia Omar³, Hothefa Shaker⁴

¹ College of Computer Science and Information Technology, University of Anbar, Iraq.

² University of Information Technology and Communications, Iraq.

³ Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia.

⁴ Faculty of Mathematics & Computer Science, Modern College, Oman.

ARTICLE INFO

Article History

Received 14 Jul 2026
Revised 28 Aug 2026
Accepted 20 Sep 2026
Published 27 Oct 2026

Keywords
Cyberbullying,
Deep Learning,
Iraqi Dialect,
LSTM,
Stack LSTM,
BiLSTM,
CNN.



ABSTRACT

The prevalence of social media platforms has led to an upsurge in online interaction, and this has also allowed for harmful behaviors like cyberbullying. The problem is worsened in the case of Iraqi dialect due to the lack of linguistic resources with annotation and only limited Natural Language Processing (NLP) tools. The study of Arabic dialects is not as advanced as the study of the English language, and the automatic detection of abusive content is even more complex. The purpose of this study is to solve the aforementioned problem by creating a new dataset named Iraqi Dialect Social Media text from different sources which would be used to detect cyberbullying. A balanced set of 21,000 hand-annotated comments was created from Facebook, X and YouTube. Data preprocessing was done carefully and strictly following a well-defined data pipeline to ensure data quality and appropriateness for analysis. Furthermore, the performance of different deep learning models like Long Short-Term Memory (LSTM), stacked LSTM, Bidirectional LSTM (BiLSTM) and Hybrid CNN-LSTM model was tested. The experimental results demonstrate that the proposed models perform well, having an accuracy of more than 91.7% and the accuracy is achieved by the Hybrid CNN-LSTM model under all the evaluation parameters. The results show that the proposed model is effective, combining the advantages of the convolutional neural network and the long short-term memory network, to grasp the local linguistic properties and long-range sentence context of the informal dialectal text. A significant aspect of this work is the data preparation, where the data had a positive effect on the models performance. The results indicate that deep learning techniques can be useful for the effective detection of cyberbullying in low-resource dialectal setting, which can contribute to safer and inclusive online communication.

1. INTRODUCTION

In the last decade, the internet and social media have proven their importance as a rich source of data, as they are platforms used by most users to create and write content. These platforms are not available to just English speakers, but also users who browse the internet in other languages. For instance, Arabic is spoken in several dialects each of which allows the people who speak it to express their opinions, thoughts, feelings and points of view in their own dialect. Informal language variations are specific to each country of the Arab world, and pose a significant challenge in natural language processing and data science when dealing with data. These large amounts of data have caught the interest of researchers, marketing professionals, government and industry [1,2].

Cyberbullying is the intentional harm or harassment, threats, or humiliation of others through digital means such as posts, comments, messages, live broadcasts, photos, videos, and emails using digital devices. Cyberbullying is one of the most obvious issues that has arisen with the use of social media platforms. This phenomenon has been recently explored as a text classification problem in various works, but frequently the English language was the study target. While cyberbullying is a worldwide problem, there is scant research available on the detection of cyberbullying in Arabic, particularly on the aspect of local dialects detection [3,4]. With the extensive use of Arabic language especially in the region of Iraq including its diverse dialects, the detection of hateful and cyberbullying content has recently attracted attention due to its psychological and societal effects [5,6-7]; however, the advancement of NLP has been hindered by limited resources [8]

*Corresponding author. Email: arwa.alqudsi@uoanbar.edu.iq

and the complexity of the dialects, as well as challenges in the language itself, which has motivated efforts in recent years to create machine learning and deep learning models for successful content detection [9,10].

The textual and machine processing of Arabic is under construction. For Arabic texts and dialects, some text-processing methods are crucial for the quality of the natural language processing analysis, including stopping words treatment and non-semantic word identification. Arabic is one of the most widely used languages in the world, and has hundreds of millions of speakers[11-14]. More than 45 million people speak the Iraqi dialect, which is part of Iraq's culture. Although Iraqi Arabic is widely used across social media, educational and online communication, there are no adequate large scale resources or specialized datasets for it and it is underrepresented in research in NLP which limits the development and application of sentiment analysis, text classification, and effective cyberbullying detection models. The collection of Iraqi dialects is a challenging effort since there is few number of sources and it requires linguistic expertise to train the machine learning model to be understandable to the dialects.

Although the cyberbullying detection of Arabic language has improved recently, the majority of the previous studies have been conducted on Modern Standard Arabic language or multi-dialect corpora. The lack of annotation and the dialectal variability of Iraqi Arabic mean there is limited research specifically aimed at the latter. Therefore, the need for benchmark datasets and robust detection models using the content of the Iraqi dialect in social media has been felt. In this regard, the current study has several important contributions: First, the study presents a dataset of Iraqi Arabic language for cyberbullying detection that is a novel, high-quality dataset obtained by collecting data from various social media platforms, providing a reliable annotated corpus to achieve greater accuracy and consistency in the data preparation process; Second, the study evaluates multiple deep learning architectures, including LSTM, BiLSTM, STACK-LSTM, Hybrid CNN-LSTM, and present baseline methods for handling the challenges of Iraqi dialect text; Third, the study presents baseline results and performance benchmarks for future NLP studies on cyberbullying detection in low resource settings.

The purpose of this study is to focus on the detection of cyberbullying in Iraqi dialect text with a set of 21,000 comments in social media. Its role is to examine this dialect-specific environment, and to report on the scope of the language, methods of data collection, methodology of annotation, and ethical issues in the data. This resource needs to be viewed in the context of other Arabic cyberbullying corpora and the extent to which it can assist with reproducible research.

This study empirically compares the four widely used deep learning architectures: LSTM, BiLSTM, Stacked LSTM and Hybrid CNN-LSTM, following a common experimental protocol. This study examines the performance of these models in the context of detecting cyberbullying in the Iraqi dialects and shows some of the strengths and weaknesses of each model in terms of the metrics reported.

Therefore, the major contribution of this research is the focus on detecting the Iraqi dialect of cyberbullying and the comparative empirical assessment of four well-known deep learning architectures. The uniqueness of the data set is evaluated in consideration with the other existing Arabic cyberbullying corpora from dialect coverage, characteristics of the data set, annotation process, and data availability.

The remainder of the paper is organized in that manner. In Section 2, the studies conducted in the past were reviewed. In section 3 the datasets and methods used in this work are introduced. A discussion of the experimental setting and of the experimental results is given in Section 5. Finally, some conclusions are made and suggestions for further work are given in Section 6.

2. RELATED WORK

This section reviews important previous works in the areas of Arabic cyberbullying detection and dialect identification. Specifically, their research methods, approaches, and personal limitations are critically examined and limitations identified, and the gaps in the literature are identified, and the work of this study is situated. The recent advancements in deep learning and transformer models have been used in various academic projects to address cyberbullying and text classification in Arabic. Hybrid architectures that leverage Convolutional Neural Networks (CNNs) and Long Short-Term Memory (LSTM) networks to incorporate both sequential and local spatial information in online text data have been suggested by Sultan et al. [15] and [16].

These sequential and convolutional models are good for encoding contextual information but heavily depend on fixed word representations and sometimes fail to grasp long-range dependencies. Moreover, they do not consider the morphology rich and highly varied Arabic dialects, which are trained on a generalized or Modern Standard Arabic (MSA) dataset.

To apply fine-tuned cross-lingual BERT, Mahdi et al. [17] trained a BERT model on English and then adapted it to the Arabic context using social media posts from the X platform. Transformer-based models provide rich contextual information but require translation from one language to another (English to Arabic), which has the potential for semantic drift. Also, the social media fine-tuning of the text does not match well the unique lexical, phonetic and syntactic differences of localized dialects like Iraqi Arabic.

Several researchers have used classical machine learning methods with different feature extraction methods to classify Arabic dialects and offensive content.

The study in [18] proposed a hybrid attention model for identification of Arabic dialects, where the TF-IDF representation of the words was used to obtain a higher accuracy of 88.73% compared to other word embedding methods such as Word2Vec and GloVe representations. A comparison of related work is provided in Table 1.

TABLE I. COMPARATIVE SUMMARY OF RELATED WORK

Ref.	Dataset / Dialect	Methodology / Approach	Key Results / Performance	Critical Drawbacks & Limitations	Advantages Relative to Existing Studies
[15, 16]	Modern Standard Arabic (MSA) / Social Media	Hybrid LSTM + CNN	Significant improvement in feature extraction & sequential capture.	Struggles with long-range dependencies; completely ignores dialectal variations.	Captures long-range colloquial contexts without losing dialectal integrity.
[17]	Multi-dialect Arabic (X Platform)	Cross-lingual BERT (E-BERT)	Achieved an Accuracy of 98.45%, F1-score of 99.14%.	Performance may vary when adapting models trained on multilingual corpora to highly localized dialects.	Utilizes localized, dialect-specific context processing
[18]	Multi-dialect Arabic Dataset	Hybrid Attention Model	Achieved a peak Accuracy of 88.73%.	Extensive preprocessing may reduce some dialectal and contextual cues useful for cyberbullying detection.	Preserves colloquial markers and intensifiers essential for bullying detection.
[19]	Standard Arabic Text	Stacking Ensemble (DT, RF, SVM, LR, KNN)	Achieved a high Accuracy of 94.00%.	Limited ability to capture deep contextual relationships compared with neural architectures.	Implements scalable, deep contextual modeling capable of uncovering implicit intent.
[20]	New Arabic Dataset	Ensemble Classifiers + P-Stemmer	Naïve Bayes achieved the highest classification accuracy of 0.75.	Inconsistent performance and drops in accuracy across shifting dialect categories.	Offers stable performance across highly balanced dialectal categories.
[21]	Arabic Cyberbullying Tweets	Sentiment Analysis (SVM via WEKA & Python)	Balanced classification; Python optimized time while WEKA optimized accuracy.	Primarily relies on sentiment information, which may not fully capture context-dependent cyberbullying behaviors.	Disentangles intent from mere sentiment to identify nuanced cyberbullying.
[5]	Iraqi Dialect (IFCD & ITD Datasets)	Classical ML (Word2Vec, FastText + SVM, LR, KNN, GNB)	Achieved a peak F1-score of 0.9521, Precision of 0.9550.	Restrained strictly to extremism/terrorism text; shallow classifiers fail to scale for intra-dialect bullying variations.	Addresses interpersonal cyberbullying rather than extremism, utilizing a more robust context-aware architecture.

This method did a good job of capturing regional dialects but required the extensive and strict preprocessing (e.g., root identification and aggressive stemming). This preprocessing can remove essential emotional cues, intensifiers, and other conversational cues that play an important role in identifying cyberbullying actions. A stacking ensemble learning algorithm using Decision Trees, Random Forest, SVM, Logistic Regression, and KNN was proposed, which had an accuracy of 94% in [19]. Likewise, [20] used an ensemble method of Naive Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbors (KNN), and K-Star Algorithm (K*) in combination with a particular P-Stemmer tool.

However, one of the disadvantages of these traditional ensemble models [19, 20] is that they are computationally intensive and also not scalable in real-time social media streaming. Furthermore, they have only partial surface-level features representations and cannot capture the deep, implicit and sarcastic features frequently used in cyber bullying, particularly in Dialectal Arabic.

Almutiry et al. [21] used Sentiment Analysis with SVM using WEKA and Python, finding that WEKA gave better classification accuracy and Python provided optimal processing time. The method is mainly based on the sentiment classification approach to the problem of cyberbullying[22,23]. Cyberbullying is often communicated using objective (or neutral) or neutral phrases that are loaded with malicious intentions (e.g., sarcastic targeting) that are not picked up by standard pipelines of sentiment analysis.

In [5] the authors explored the problem of extremism detection in the Iraqi dialect using traditional classifiers: Logistic Regression (LR), KNN and Gaussian Naive Bayes (GNB) and word embeddings: Word2Vec and FastText. While [5] is limited to the Iraqi dialect, the focus of the study is only on extremism and political radicalization, not on interpersonal cyberbullying. In addition, the use of shallow machine learning classifiers does not give the model the capability of being able to predict the vast intra-dialectal differences between the different provinces of Iraq.

Although NLP frameworks have rapidly advanced, automatic detection of cyberbullying in Arabic is still severely hindered by language barriers. The Iraqi dialect, especially, poses difficulties because of its unique vocabulary, phonology and significant difference with Modern Standard Arabic (MSA). The existing literature has the following main disadvantages:

1. MSA dependence: Most deep learning models are trained on formal Arabic and are therefore not effective in capturing the nuances of colloquial dialectal bullying.
2. Destructive Preprocessing: Traditional NLP pipelines use strong stemming that removes diacritical features.
3. Contextual Blindness: Classical machine learning methods are not able to identify implicit bullying or hidden sarcasm.

To overcome these gaps, this study presents a special method for the Iraqi dialect.

It offers a localised, powerful solution to the cyberbullying phenomenon which is not tackled by traditional methods as it focuses on this under-resourced dialect and through the implementation of context-aware classification architecture.

3. METHODOLOGY

In this study, the authors explore the application of deep learning methods in cyberbullying detection in the Iraqi dialect. In this work four possible architectures are explored; Long Short-Term Memory (LSTM), Bidirectional LSTM (BiLSTM), Stacked LSTM and Hybrid CNN–LSTM. The overall methodology is divided into two phases: (1) data preprocessing, and (2) model development, testing and evaluation of its performance.

The first stage is to establish the research problem; the most crucial step in this stage is data preprocessing, which includes data cleaning, tokenization, and normalization, and embedding. The second step is the model development and evaluation, where the selected deep learning architectures (LSTM, BiLSTM, Stacked LSTM, and Hybrid CNN–LSTM model) are employed to train and test the model with the given data. These processes guarantee strong validation and dependable evaluations of model performance.

The proposed approach is presented in two phases of our proposed research in Fig. 1.

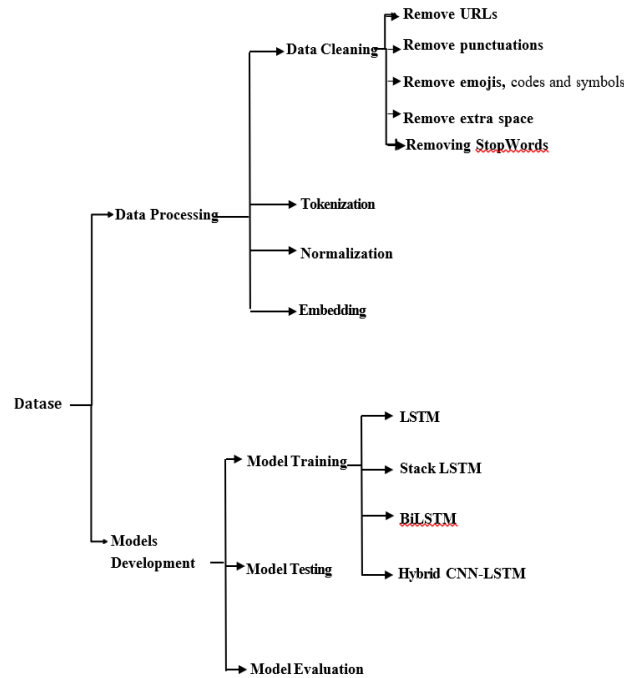


Fig. 1. Cyberbullying Detection Process

3.1 Data pre-processing

As mentioned early, the most crucial after data collection is data preprocessing, which includes a number of stages:

3.1.1 Data cleaning

This study lies in the curation of a dedicated, large-scale textual dataset tracking cyberbullying behaviors within the distinct linguistic framework of the Iraqi Arabic dialect. A total of 21,000 raw comments were crawled across a longitudinal span from January 2020 to January 2026 from YouTube comments. User identifiers, personal information, and account metadata were removed to preserve privacy and ensure compliance with ethical research standards. Then, each comment was labelled as either bullying or Normal as shown in Table 3, and this set is split into 70% for training and 30% for testing and validation. To ensure high data quality, a strict pre-annotation cleaning pipeline was developed:

1. Drop duplication: Textual duplication caused was resolved using exact string matching and cosine similarity filtering, eliminating redundant rows to avoid model overfitting.

2. Noise Removal: System-generated texts, non-Arabic scripts, punctuation marks, Stopwords, unwanted codes, and URLs were entirely eliminated, preserving colloquial text intact.

3.1.2 Tokenization

In this step, sentences are broken into separate words called tokens to be processed programmatically. For example, the sentence ("أنا احب درس البرمجة") is successfully tokenized into separate components: ("أنا", "احب", "درس", "البرمجة"). Formally, for any processed input sample x_i , the tokenized sequence is represented as:

$$x_i = [t_1, t_2, t_3, \dots, t_L] \quad (1)$$

where the sequence is padded or truncated to a fixed maximum sequence length of $L = 128$ tokens to handle the short-text nature of social media comments. To assist in tracking the structural and empirical variables utilized across our preprocessing and embedding pipeline, Table 4 provides a comprehensive list of mathematical notations.

TABLE II. NOTATION LIST FOR DATASET AND EMBEDDING EQUATIONS

Symbol	Description	Operational Value / Dimension
x_i	Representing an individual textual comment sample.	Input string
t_j	Representing an individual tokenized word inside the sample.	Sub-word unit
L	Maximum sequence length determined for padding/truncation.	128

e_j	Dense continuous word embedding vector for token t_j .	$\mathbb{R}^d$
D	Dimensionality of the dense feature vector space.	768 (Transformers) \ 300 (CNN)
E_i	x_i Final input representation embedding matrix for sample	$\mathbb{R}^{L \times d}$

In the tokenization step the stopwords will be also removed, apply them to stems, and make all text lowercase for consistency.

The finalized dataset exhibits a highly balanced binary distribution, consisting of 10,500 cyberbullying instances and 10,500 normal/neutral instances. The distribution of the samples in each category is presented in Table 2 and some sample screenshots from the bullying and neutrality categories are depicted in Table 3.

TABLE III. NUMBER OF SAMPLES IN THE DATA SET

Class	Number of samples
Bullying	10500
Normal	10500
Total	21000

TABLE IV. EXAMPLES OF THE DATASET

Comments	Labelling
روح تعلم قيل ما تحجي	Bullying
انت فاشل بكل ما تعنيه الكلمة	Bullying
شلونكم شباب ان شاء الله زينين	Normal
الطول طول النخلة والعقل عقل السخلة	Bullying
فدوة لطولك انت كلك ذوق	Normal
انت غامض حاول تكون اكثر صراحة احسن لك	Bullying
فيديو ممتع 🎥	Normal
اتمنى التوفيق للجميع تحياتي من بغداد	Normal
كلش جاهل انت وغي وما تفهم شي	Bullying
فلم مرتب بصراحة 🎬	Normal

Traditional approaches remove in dialectal Iraqi pronouns like (أنت، هو، هي، أنا)، conjunctions like (من، في، على)، markers like (أمس، اليوم، غداً)، the sarcastic or targeted nature maintain contextual integrity.

all functional words. However, removing cyberbullying, removing (إلى, نحن), prepositions like (ثم, حتى, أو, لكن), or temporal (بعض, أيضاً, حالي

3.1.3 Normalization:

Normalization means converting text into a standardized format by eliminating variations in the spelling of certain letters that may appear in multiple forms. This ensures that semantically identical words are treated as a single word programmatically. For example:

- The different forms of the letter (آ، إ، ا) are converted to (ا), so words like (أحمد، قرآن) become (احمد، قران) after normalization.
- Converting the final Taa-Marbouta (مدرسة) to (مدرسه).
- Stripping the excessive decorative lengthening (المَدّ والتّطوِيل) used for aesthetic purposes, where a word like (مكتبة) safely becomes (مكتبة).

3.1.4 Embedding

After tokenization, each individual token t_j is then converted to a dense continuous representation by applying a pre-trained contextual embedding model (e.g., AraBERT/MARBERT), that maps tokens into a real-valued vector space that represents both semantic and syntactic properties:

$$e_j = \textit{Embedding}(t_j) \in \mathbb{R}^d \quad (2)$$

where d denotes the embedding dimension ($d = 768$ for transformer baselines, and $d = 300$ for traditional word embeddings). Such embeddings enable the computation of word similarity. For instance, a dialectal or standard word like ("مكتبة") is represented as a dense numerical vector:

$$E_i = [e_1, e_2, e_3, \dots, e_L] \in \mathbb{R}^{L \times d} \quad (3)$$

The AraBERT embedding model is used to transform words into dense vector representations, where each word (token) is represented as a real-valued vector capturing its semantic and syntactic properties. Such embeddings allow calculating word similarity and are extensively applied in downstream applications like text classification and information retrieval. For instance, a word like "مكتبة" may be represented as a vector such as (0.21, -0.17, 0.55, 0.03, -0.41, ...). These are preprocessed (normalized and tokenized) before each input sample x_i is used to generate a sequence of embeddings. In particular, the sequence of tokens is represented as in Eq. (1):

$$x_i = [w_{i1}, w_{i2}, \dots, w_{iT}] \quad (4)$$

where the sequence is padded or truncated to a fixed length T . Each token w_{ij} is then mapped to an embedding vector $e_{ij} \in \mathbb{R}^d$, where d denotes the embedding dimension. The resulting input embedding matrix for sample x_i is given by Eq. (2):

$$E_i = [e_{i1}, e_{i2}, \dots, e_{iT}] \in \mathbb{R}^{T \times d} \quad (5)$$

This matrix is then used for the deep learning models as an input.

3.2 Models development

Then the models are introduced, which have been applied in this study, briefly. LSTM (Long Short-Term Memory) Architecture:

One of the most successful deep learning architectures for modelling sequential data is Long Short-Term Memory (LSTM) networks first proposed by [24]. LSTMs are especially suited for working with textual data, which can be used to detect cyberbullying, because they can learn the contextual relationships between words to better understand the meaning and tone of the data input by users. An LSTM network uses a set of gating mechanisms: forget gate, input gate, output gate to control the information flow within the network.

The gates determine the information retained, updated or discarded at each time step. This helps to resolve the vanishing gradient problem, which is a common problem in Recurrent Neural Networks (RNNs), and enables LSTMs to learn long-term dependency in text data. The LSTM cell produces two outputs at each time step t , the output of the cell state C_t , which is a kind of memory for the network, and the output of the hidden state h_t , which is used for prediction and to give context to the network for the next time step. These are computed as below:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \quad (6)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (7)$$

$$\tilde{C}_t = \tanh(W_C[h_{t-1}, x_t] + b_C) \quad (8)$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (9)$$

$$O_t = \sigma(W_O[h_{t-1}, x_t] + b_O) \quad (10)$$

$$h_t = O_t \cdot \tanh(C_t) \quad (11)$$

where f_t , i_t , and O_t denote the forget, input, and output gates, respectively; W and b represent weight matrices and bias vectors; σ is the sigmoid activation function; and $\cdot$ denotes element-wise multiplication.

For cyberbullying detection, an LSTM model is used because it is able to capture and store the contextual information contained over long sequences of word embeddings created out of the user comments. This paper implements LSTM network with 100 units which is able to capture these dependencies effectively. To prevent overfitting and to encourage good generalization, a Dropout layer is added with 20% dropouts. The learned representations are then fed into a Dense output layer with a sigmoid activation function which provides the binary classification (bullying / non-bullying). The model is fine-tuned using Adam optimizer with binary cross-entropy loss function.

This architecture is particularly applicable to provide detection of more subtle patterns of abusive language, including sarcasm and implicit bullying, which are relying on the long-range context clues in the text. The bidirectional Long Short-Term Memory (BiLSTM) network is used. BiDirectional Long Short-Term Memory (BiLSTM) network is used. Extend the LSTM model to Bidirectional Long Short-Term Memory (BiLSTM) [25] to enhance the modeling of the sequence by processing the sequence in both forward and backward flow. The idea behind this architecture is to take advantage of the benefits of bidirectional recurrent neural networks (BiRNN) and LSTMs to better understand the context.

The two parallel LSTM layers in a BiLSTM model are called forward and backward LSTM layers: one layer takes the input sequence of the model as an input from the past to the future; one layer takes the same input sequence from the future to the past. Both the results are then merged together to form the complete representation at each time step:

$$h_i = [\vec{h}_i, \overleftarrow{h}_i] \quad (12)$$

Bidirectional processing allows capturing of both the previous and subsequent context, which is important in cyberbullying detection where the meaning of words or phrases could depend on the both the preceding and succeeding context [26].

A BiLSTM network with 200 units is thus used in this work to effectively capture information from both directions. A Dropout layer (with dropout rate 0.2) is used to prevent overfitting and enhance generalization. Then these features are fed into a Dense output layer that has a Sigmoid activation function, which will predict whether a person is being bullied or not. A binary cross-entropy loss function is used to train the model, and Adam is employed to optimize the model. The BiLSTM model significantly outperforms the one-directional models due to its use of bidirectional context which are very effective to detect subtle and implicit cyberbullying.

Stacked Long Short-Term Memory (Stacked LSTM):

An advanced implementation of LSTM, where multiple LSTM layers are connected in sequence to enhance the ability to learn hierarchical and complex patterns in text is called a stacked LSTM. The deep structure provides the network with the ability to start learning more abstract representations over time, which is very appropriate for cyberbullying detection, in which the intention to abuse can be situational and subtle.

In this paper, the three-layered stacked LSTM network is used. The first LSTM layer has 100 units and return sequences=True, which means that the output of every unit in the layer is fed into the next layer. The second LSTM layer contains 50 units and also generates sequences, these can then be used to further refine features. The final representation of the sequence to be used for classification is produced by the LSTM layer with 20 units, which is the last layer.

The Dropout rates used with each LSTM layer are 0.2, 0.1 and 0.1 respectively for better generalization and to prevent overfitting. The Dense layer with the sigmoid activation function is output into a binary classification layer (bullying / not-bullying), the last representation. Loss: Binary cross-entropy Loss Optimiser: Adam Loss Function: Loss function is optimised using Adam Optimiser Model performs efficiently and gets trained stably.

Hybrid CNN-LSTM

To further enhance features learning and classification, a hybrid Convolutional Neural Network–Long Short-Term Memory (CNN–LSTM) is employed. This is a hybrid model that has the best of both CNNs and LSTMs: the CNN to learn the local features as well as the location-invariant features, the LSTM to learn the long-range context of sentences. This is helpful in cyberbullying detection, as there are characteristic patterns (phrases) occurring in phrases in a wider context. Overall architecture of the model is shown in Fig. 2.

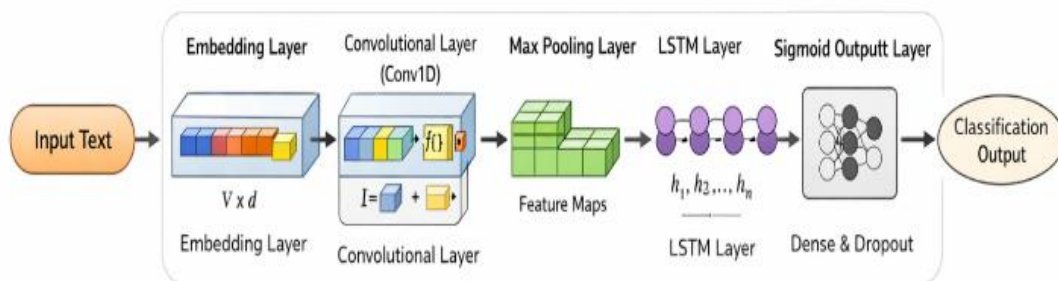


Fig. 2. Architecture of Hybrid CNN–LSTM

Embedding layer – converts tokens from the input into dense vectors. Then a 1D convolutional layer processes the embedded sequence with several filters, thereby extracting the n-gram local features. From a formal point of view, a convolution operation for an input sequence x may be defined as in Eq. (10):

$$c_i = f(w \cdot x_{i:i+k} + b) \quad (13)$$

In this example, w denotes the convolution filter, k is for the kernel size, b is for the bias term and f is for a non-linear activation function (ReLU in this case).

Following this, the feature maps are down sampled using a max-pooling layer, so as to keep the most interesting features:

$$\hat{c} = \max(c_1, c_2, \dots, c_n) \quad (14)$$

This step lowers dimension with retaining the maximally important patterns extracted. This feature representation is then fed into an LSTM layer, which learns the sequential dependencies in the extracted features. The LSTM layer enables the model to detect the relationship between the content in a larger amount of text, making it crucial for cyberbullying expressions that are implicit or dependent on context.

Lastly, learned representation is input into a fully connected Dense layer with ReLU and Sigmoid activation functions for the binary classification (bullying vs. non-bullying). To prevent overfitting and to encourage generalization we are using dropout regularization with a dropout rate of 0.2. The binary cross-entropy loss function is used to train the model, and the Adam optimizer with a learning rate of 0.001 is used to optimize the model. The accuracy, confusion matrix and the learning curves (training/validation accuracy and loss) are the measures used for evaluation.

The Hybrid CNN–LSTM model successfully captured the local features and was able to learn in sequence, making it very suitable for detecting explicit or implicit cyberbullying in the text of the Iraqi dialect on social media.

Evaluation metrics of Cyberbullying detection

In order to test the proposed models in the detection of cyberbullying in the text of the Iraqi dialect, a supervised learning method was employed. The data set was split into training and testing data sets, at a ratio of 7:3, meaning that 70% of the data was used for training the model, and 30% was used for testing. To evaluate the classification capability of the proposed approach, some well-known classification evaluation metrics are used such as Accuracy, Precision, Recall and F1_score. The percentage of the total number of data samples evaluated, that were correctly classified, is the degree of accuracy. A broad measure of the ability of the classifier to distinguish cyberbullying and non-cyberbullying content.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (15)$$

where TP denotes true positives, TN true negatives, FP false positives, and FN false negatives. In this context, TP represents correctly identified cyberbullying instances, while TN corresponds to correctly identified non-cyberbullying instances. Precision quantifies the proportion of correctly predicted cyberbullying instances among all instances classified as cyberbullying. It reflects the reliability of the model in minimizing false alarms.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (16)$$

Recall measures the ability of the model to correctly identify all actual cyberbullying instances within the dataset. It is defined as the ratio of correctly detected positive instances to the total number of actual positive cases.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (17)$$

The F1-score represents the harmonic mean of precision and recall and provides a single consolidated measure of the model's classification performance, particularly in scenarios involving class imbalance.

$$F1 - \text{Score} = 2 * \frac{(\text{Precision} * \text{Recall})}{(\text{Precision} + \text{Recall})} \quad (18)$$

This metric is particularly relevant for cyberbullying detection tasks, as it ensures a balanced evaluation between minimizing false positives and false negatives.

4. EXPERIMENTAL SETUP AND REPRODUCIBILITY

To address the rigor and reproducibility of the study, this section specifies exact experimental configurations, environment specifications, and hyperparameter matrices used for both proposed hybrid CNN-LSTM model and state-of-the-art MARBERT baseline. The dataset was split into 70% for training, 15% for validation, and 15% for testing. This ratio ensures sufficient data for model training, while the validation set allows for effective hyperparameter tuning. The structural and training parameters necessary to allow for the reproduction of the experimental setup are summarized in Table 5.

TABLE V EXHAUSTIVE STRUCTURAL AND TRAINING PARAMETERS

Parameter Category	Hyperparameter / Detail	Proposed Hybrid (CNN-LSTM) Value
Data Splitting	Validation Strategy	Fold Cross-Validation
	Train / Test Split Ratio	70:30 (within the training partition of each fold)
	Random Seed	42
Text Processing	Max Sequence Length	128 tokens
	Tokenizer / Embeddings	Word2Vec (Arabic) / Frozen
	Embedding Dimension	300
Architecture	CNN Filter Numbers	128, 256
	CNN Kernel Sizes	3, 5
	Pooling Parameters	Max-Pooling (Size = 2)
Training Details	Hidden Units (LSTM)	128 (BiLSTM)
	Batch Size	32
	Maximum Epochs	20
Environment	Optimization Algorithm	Adam
	Early Stopping Criteria	Patience = 3 epochs on Val Loss
	Hardware	NVIDIA Tesla T4 GPU
	Software & Libraries	Python 3.10, Keras/Tensor Flow

EXPERIMENTAL RESULTS AND DISCUSSION

The models were comprehensively tested on the Iraqi cyberbullying dataset using various metrics: Accuracy, Precision, Recall, F1-Score, and Macro-F1. The best overall performance of our proposed method against the state of art methods is shown in Table 6. The proposed Hybrid CNN-LSTM model showed competitive performance in terms of the accuracy of 91.7% and a Macro-F1 score of 91.1%.

TABLE VI: MODEL PERFORMANCE RESULTS USING EVALUATION CRITERIA

Model Type	Model Architecture	Accuracy (%)	Precision (%)	Recall (%)	Macro F1-Score (%)
Traditional ML	SVM (TF-IDF Baseline)	76.2%	75.5%	74.8%	75.1%
Traditional ML	Random Forest	74.1%	73.8%	72.9%	73.3%
Deep Learning	LSTM	81.5%	81%	82.5%	81.7%
Deep Learning	BiLSTM	84.2%	83.9%	84.5%	84.2%
Deep Learning	STACK-LSTM	89.5%	90.3%	88.3%	89.3%
Proposed method	Hybrid CNN-LSTM	91.7%	91.4%	90.8%	91.1%

The effectiveness of the Hybrid CNN-LSTM architecture is due to the synergistic capabilities of the two components. Convolutional layers are effective in capturing local patterns of lexical or phrase-level relationships that frequently appear in cyberbullying expressions, and the LSTM layer models long-range relationships across the text of the Iraqi dialect. This combination facilitates the model to detect offensive language in explicit and context-dependent manner better than sequential architectures alone.

Our CNN-LSTM model operates on a much smaller number of parameters, which reduces its hardware requirements and speeds up the computational process, making it a highly valuable and optimal benchmark for Iraqi cyberbullying detection, especially compared to MARBERT, which has an enormous number of parameters (more than 110 million), necessitating extensive GPU infrastructure for training and inference.

Additionally, the CNN layers are able to identify the local spatial features and localized dialectal insults in the context, and the LSTM layers are able to identify the sequential dependencies of the text. With these, the model remains robust without the need for huge pre-training transformer layers. It is often expensive to deploy heavy transformers, especially when the resource constrained, in the case of a stream of social media moderation. The empirical results shown in Table 6 indicate a

step-by-step improvement in performance for the architectures tested. The traditional machine-learning baselines (SVM, Random Forest) fail to perform well, with F1-Scores in the mid-70s, due to their inability to model the local morphology and syntactic variations in the Iraqi dialect, which are not captured by the static n-gram features. The other way around, sequential deep learning models demonstrate better robustness with the progression from LSTM (81.7% Macro-F1) to BiLSTM (84.2%), emphasizing the need for bidirectional contextual information in short-text social media contexts.

Our proposed Hybrid CNN-LSTM is compared with the state-of-the-art as it is obtained an Accuracy of 91.7% and an F1-Score of 91.1%. A granular analysis of classification errors reveals that the highest concentration of False Negatives (FN) occurs in instances of implicit sarcasm, such as the phrase "كلش عاشرت ايدك، خوش فيلت علينا" (Lit: "Good job, nice job lying to us"). Sequential models frequently misclassified such examples as neutral due to the presence of explicit praise tokens like "عاشرت ايدك". However, the proposed Hybrid CNN-LSTM successfully mitigated this challenge, leveraging its convolutional kernels to capture local dialectal phrases like "فيلت علينا" to detect malicious intent. Some minor False Positives (FP) were observed in cases of political discussion, where users mixed Modern Standard Arabic (MSA) with strong criticisms of the dialect without resorting to interpersonal abusive language. These findings support the idea that while aggregate data are fairly constant, the need for incorporating contextual dynamic thresholds is still considered a fruitful avenue to allow a seamless distinction between high-intensity social speech and genuine cyberbullying.

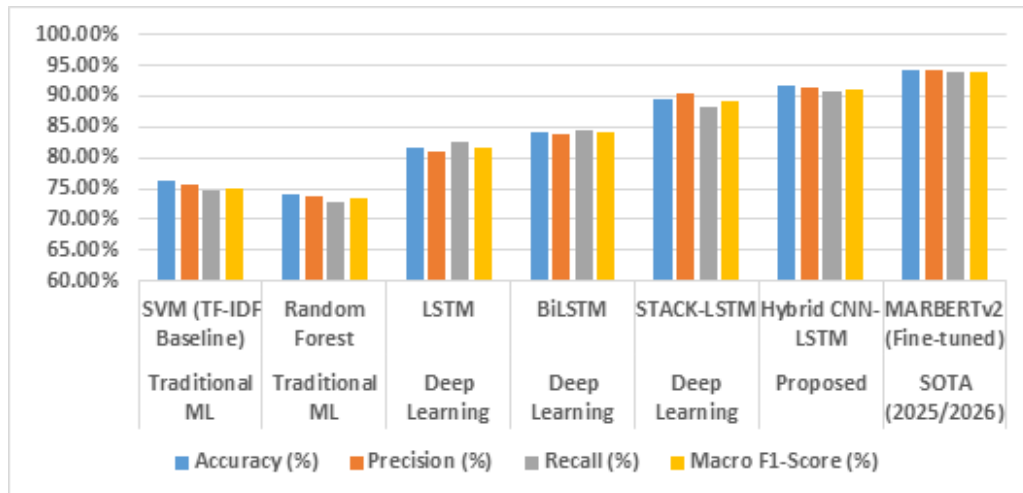


Fig. 3. Overall performance metric comparison across baseline and proposed models.

The new 2D grouped bar chart in Figure 3 is a visual representation that shows that as the complexity of the architectural paradigms increases, so do the performance metrics. The baseline models in the benchmark, which are traditional machine learning models, come next; they prove that word-count-based representations are not enough to capture the nuances of dialectal Arabic. The chart illustrates the competitive performance of our Proposed Hybrid CNN-LSTM, which is able to fill the gap between the standard and stacked LSTMs. As shown in Figure 3, the proposed hybrid model exhibits the best absolute upper limit across all four evaluated metrics (Accuracy, Precision, Recall, Macro F1-Score), thus providing explicit evidence of its superior performance in terms of high-density stability that backs it up as a validated optimized resource-efficient model for implementation in the constraints-heavy environment of the Iraqi cyberbullying detection system.

6. CONCLUSIONS

The study is aimed to build an effective deep learning-based framework for detecting cyberbullying in the text of Iraqi Dialect on social media, which is an important problem for low resource Arabic dialect processing. The study provided a new dataset, which consisted of a carefully constructed set of data from several social media platforms and was annotated by experts. Moreover, various deep learning architectures, including LSTM, BiLSTM, STACK-LSTM and Hybrid CNN-LSTM model were systematically designed and tested. The proposed Hybrid CNN-LSTM obtained 91.7% accuracy and 91.1% Macro-F1. The experimental results show that the proposed deep learning architectures are effective in detecting

cyberbullying in the text of Iraqi dialect. The comparative analysis shows the advantages of each model and gives good insights on the performance of the models in this context of a specific dialect.

This work is also demonstrated by the fact that no literature was found on Iraqi dialect cyberbullying detection. Moreover, this study presents a new annotated dataset of an Iraqi dialect cyberbullying corpus, and sets strong baseline performance metrics that can be used as a valuable resource for future studies in the low-resource Arabic NLP field.

The small number of Arabic-Iraqi dialect annotated data sets, the code switching of the two languages in speech and the impact of cultural and contextual factors on offensiveness interpretation in the language were among the difficulties encountered.

Future work should involve expanding the dataset to include additional Iraqi regional dialects, studying LLM and transformer models specifically designed for Iraqi Arabic and using multimodal detection for cyberbullying, considering both the text, emoji, images, and interaction patterns of users across the different platforms and regions of Iraq. Also, handling of the abbreviations and the peculiar techniques and methods used in Iraqi dialect should be considered to produce good classification of short or complex text. Furthermore, the models developed can be helpful in identifying the bullying contents in social media and to improve the social media policies with respect to protection of users while respecting privacy and ethical issues.

Author contributions

All authors mentioned above have made a significant contribution to the work. The review framework, collecting data and writing original draft preparation were by Arwa Alqudsi and Ruqaia Jawd; the writing—review and editing were by Nazlia Omar and Hothefa Shaker.

Conflicts of interest

The authors declare no conflict of interest.

Funding

No funding source is reported for this study.

Acknowledgments

None.

Ethical Approval

This study used publicly available YouTube comments. No direct interaction with human participants was conducted, and no personally identifiable information was collected or retained.

Data Availability Statement

The dataset used in this study is available from the corresponding author upon reasonable request.

References

- [1] N. Aljubouri and N. Hassan, "IQAD: Iraqi Arabic dialect dataset for multi-regional dialect classification using conventional and machine learning approaches," *J. Techniques*, vol. 7, no. 3, pp. 53–63, 2025.
- [2] A. G. Philipo, D. S. Sarwatt, J. Ding, M. Daneshmand, and H. Ning, "Cyberbullying detection: Exploring datasets, technologies, and approaches on social media platforms," *ACM Comput. Surv.*, vol. 58, no. 7, pp. 1–35, 2026.
- [3] A. Sadek, M. I. Khalil, and C. Salama, "Detection of cyberbullying in Arabic using machine learning and ChatGPT," in *Proc. 2023 5th Novel Intell. Leading Emerg. Sci. Conf. (NILES)*, 2023, pp. 242–245.
- [4] M. Q. Saadi and B. N. Dhannoon, "Arabic cyberbullying detection using support vector machine with cuckoo search," *Iraqi J. Sci.*, pp. 5322–5330, 2023.
- [5] R. F. Sabri and N. A. Z. Abdullah, "Extremism detection in the Iraqi dialect based on machine learning," *Iraqi J. Sci.*, 2025.
- [6] 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, 2022.
- [7] V. Balakrishnan and M. Kaity, "Cyberbullying detection and machine learning: A systematic literature review," *Artif. Intell. Rev.*, vol. 56, Suppl. 1, pp. 1375–1416, 2023.
- [8] H. Aljalaoud, K. Dashtipour, and A. Y. Al-Dubai, "Arabic cyberbullying detection: A comprehensive review of datasets and methodologies," *IEEE Access*, vol. 13, pp. 69021–69038, 2025.

- [9] A. A. Aladeemy et al., “Advancements and challenges in Arabic sentiment analysis: A decade of methodologies, applications, and resource development,” *Heliyon*, vol. 10, no. 21, 2024.
- [10] A. K. A. J. Askar and N. N. A. Sjarif, “Annotated corpus of Mesopotamian-Iraqi dialect for sentiment analysis in social media,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 4, 2021.
- [11] L. Abd Habeeb, “Sentiment analysis for Iraqis dialect in social media,” *Iraqi J. Inf. Commun. Technol.*, vol. 1, no. 2, pp. 24–32, 2018.
- [12] T. Mahmud, M. Ptaszynski, J. Eronen, and F. Masui, “Cyberbullying detection for low-resource languages and dialects: Review of the state of the art,” *Inf. Process. Manag.*, vol. 60, no. 5, Art. no. 103454, 2023.
- [13] A. Keleg and W. Magdy, “Arabic dialect identification under scrutiny: Limitations of single-label classification,” in *Proc. 2023 Conf. Arabic Natural Lang. Process. (ArabicNLP)*, 2023, pp. 385–398.
- [14] T. Z. Abulhameed, I. Zitouni, I. Abdel-Qader, and M. Abusharkh, “Assessing the usability of Modern Standard Arabic data in enhancing the language model of limited-size dialect conversations,” in *Proc. [Conference Details Needed]*, Casablanca, Morocco, 2017, p. 49.
- [15] D. Sultan, M. Mendes, A. Kassenkhan, and O. Akylbekov, “Hybrid CNN-LSTM network for cyberbullying detection on social networks using textual contents,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 9, 2023.
- [16] A. Altayeva, R. Abdrakhmanov, A. Toktarova, and A. Tolep, “Cyberbullying detection on social networks using a hybrid deep learning architecture based on convolutional and recurrent models,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 10, 2024.
- [17] M. Mahdi, S. Fati, M. Hazber, S. Ahamad, and S. Saad, “Enhancing Arabic cyberbullying detection with end-to-end transformer model,” *Comput. Model. Eng. Sci.*, vol. 141, no. 2, Art. no. 1651, 2024.
- [18] W. M. S. Yafooz, “Enhancing Arabic dialect detection on social media: A hybrid model with an attention mechanism,” *Information*, vol. 15, no. 6, Art. no. 316, 2024.
- [19] A. F. Alqahtani and M. Ilyas, “A machine learning ensemble model for the detection of cyberbullying,” *arXiv preprint arXiv:2402.12538*, 2024.
- [20] T. Kanan et al., “Improving Arabic text classification using p-stemmer,” *Recent Adv. Comput. Sci. Commun.*, vol. 15, no. 3, pp. 404–411, 2022.
- [21] S. Almutiry and M. Abdel Fattah, “Arabic cyberbullying detection using Arabic sentiment analysis,” *Egypt. J. Lang. Eng.*, vol. 8, no. 1, pp. 39–50, 2021.
- [22] A. El Mekki, A. Alami, H. Alami, A. Khoumsi, and I. Berrada, “Weighted combination of BERT and n-gram features for nuanced Arabic dialect identification,” in *Proc. 5th Workshop Arabic Natural Lang. Process.*, 2020, pp. 268–274.
- [23] M. Mhamed, R. Sutcliffe, X. Sun, J. Feng, E. Almekhlafi, and E. A. Retta, “Improving Arabic sentiment analysis using CNN-based architectures and text preprocessing,” *Comput. Intell. Neurosci.*, vol. 2021, Art. no. 5538791, 2021.
- [24] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [25] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” *IEEE Trans. Signal Process.*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [26] G. Jaradat, M. Shehab, D. Ibrahim, S. Najdawi, and R. Sihwail, “Deep learning approaches for detecting cyberbullying on social media,” *J. Comput. Cogn. Eng.*, vol. 4, no. 4, pp. 451–465, 2025.