Jordanian Journal of Computers and Information Technology (JJCIT), Vol. 11, No. 03, September 2025.

ADVANCED DEEP-LEARNING TECHNIQUES FOR IMPROVED CYBERBULLYING DETECTION IN ARABIC TWEETS

Marah Hawa, Thani Kmail and Ahmad Hasasneh

(Received: 1-Mar.-2025, Revised: 11-May-2025 and 15-Jun.-2025, Accepted: 19-Jun.-2025)

ABSTRACT

Cyberbullying has emerged as a pressing issue in the digital era, particularly within Arabic-speaking communities, where research remains limited. This study investigates the detection of Arabic cyberbullying on social media using both traditional machine learning (ML) and deep learning (DL) techniques. A publicly available dataset of Arabic tweets was used to train and evaluate several ML models (SVM, NB, LR and XGBoost), alongside a recurrent neural network (RNN). The results demonstrate that the RNN significantly outperforms classical ML models, highlighting the efficacy of DL in accurately identifying abusive content in Arabic text. These results emphasize the necessity of incorporating linguistically rich data and advanced neural architectures to improve cyberbullying-detection systems in low-resource languages such as Arabic.

KEYWORDS

Machine learning algorithms, Arabic tweets, Deep-learning techniques, Recurrent neural network, Cyberbullying.

1. INTRODUCTION

Cyberbullying involves the use of digital platforms—such as smartphones and social media—to inflict harm through behaviors like verbal abuse, offensive language and harassment. Its psychological impact can be profound, especially among teenagers, leading to issues, such as low self-esteem, anxiety and identity-related concerns. The problem has intensified globally with the growing popularity of platforms, like Twitter (now X), where anonymity enables harmful behavior without accountability [1].

Recent reports highlighted the scale of the issue: in 2024, 28% of adolescents experienced cyberbullying and over 42% of youth aged 13–24 years in the MENA region reported exposure to online abuse *via* popular apps like Instagram, TikTok and Twitter [2]–[5]. The ITU and Arab Social Media Observatory have similarly flagged cyberbullying as a major digital threat to the mental health of children and adolescents [6]-[7]. These findings point to an urgent need for scalable, data-driven solutions that go beyond manual moderation.

Despite growing efforts in English-language research, Arabic cyberbullying detection remains underexplored. The increasing use of Arabic on social media—especially Twitter—demands more targeted approaches, but the language's rich morphology, diverse dialects and limited annotated resources present ongoing challenges [8]. The situation was further exacerbated by the COVID-19 pandemic, which saw young users spending more time online and becoming more vulnerable to digital abuse [9].

To address these gaps, this study proposes a deep learning-based model for detecting cyberbullying in Arabic text. By combining three datasets representing different Arabic dialects into a single corpus and applying a Recurrent Neural Network (RNN)—a relatively underutilized method in this context—we achieve significant improvements in detection performance. Our work contributes to the development of more robust and linguistically aware systems for identifying abusive content in Arabic-language social media.

The rest of this paper is organized as follows: Section 2 reviews related work and the datasets used; Section 3 outlines the proposed methodology; Section 4 presents and analyzes the results; Section 5 offers a comprehensive discussion; and Section 6 concludes the paper.

M. Hawa, T. Kmail and A. Hasasneh (Corresponding Author) are with Department of Natural, Engineering and Technology Sciences, Faculty of Graduate Studies, Arab American University (AAUP), Ramallah, Palestine. Emails: m.hawa1@student.aaup.edu, t.kmial@student.aaup.edu and Ahmad.Hasasneh@aaup.edu

2. LITERATURE REVIEW

Natural-language processing (NLP) technologies have evolved substantially over the decades, becoming vital for enabling effective human-computer interaction [11]. Fundamentally, NLP transforms natural-language texts into machine-processable digital formats, enabling sophisticated tasks, such as machine translation and sentiment analysis [12]-[13]. The roots of NLP trace back to the 1950s with early systems, like the Georgetown-IBM translation experiment, which laid the groundwork for subsequent advances in AI-driven text understanding.

A critical initial step in NLP pipelines is text pre-processing, which ensures high-quality input data for improved model performance. Tokenization breaks text into meaningful units, such as sentences or words, facilitating downstream analysis. Techniques, like stemming, which reduce words to their root forms and stop-word removal, which excludes frequent, but semantically light words, are essential for reducing noise and dimensionality [14], [16]. Kanaan et al. [15] demonstrated that combining stemming with truncation, normalization and stop-word removal significantly boosts classification accuracy and F1-scores in document-classification tasks.

In the realm of machine learning (ML), classical algorithms such as Support Vector Machines (SVMs), Naive Bayes (NB), Logistic Regression (LR) and Extreme Gradient Boosting (XGBoost), have remained popular due to their efficiency and interpretability. These models have been applied extensively for Arabic-cyberbullying detection, yielding solid baseline results. For example, Hani et al. [23] reported over 89% accuracy using linear SVM with TF-IDF features on a small Arabic dataset, while Rashid et al. [24] and Moheb et al. [21] achieved accuracies up to 95% with NB classifiers. Logistic regression also performs competitively, with Rashid et al. [24] improving F1-scores through dataset balancing and feature engineering. XGBoost, a powerful ensemble method, showed promising results with 85% accuracy [24].

2.1 Classification Methods

Many researchers have collected data from popular social-media platforms, such as Twitter and Facebook, to study cyberbullying. For instance, Aladdin et al. [17] utilized the Twitter API to gather their dataset. Similarly, Haidar et al. [18]-[19] developed dedicated tools in Python and PHP to collect data from both Facebook and Twitter, storing it in a MongoDB database. Al-Harbi and colleagues [20] compiled a large dataset comprising 100,327 tweets and comments collected from Twitter, YouTube and Microsoft platforms. Meanwhile, Mohib et al. [21] gathered 25,000 tweets and comments from Twitter and YouTube using their respective APIs. Other studies employed tools, such as NLTK, for text analysis or platforms, like RStudio, for extracting tweets [22]-[23]. Although most of these datasets were primarily in English, some research focused on Arabic data collected from sources, including Twitter, Facebook and YouTube [22]. Most datasets were processed and manually annotated, while Arabic-cyberbullying datasets remain comparatively limited.

The literature highlights the significant role of machine-learning algorithms in addressing cyberbullying challenges by detecting harmful patterns and behaviors through classification and text analysis. Support Vector Machines (SVMs) have been widely used for text classification in Arabic-cyberbullying research. For example, Hani et al. [23] achieved over 89% accuracy using a linear SVM on a small dataset of 1.6K publications after extracting features with the term frequency-inverse document frequency (TF-IDF) method. The Naive Bayes (NB) classifier has also been extensively applied in Arabic-text analysis [12], [24]-[25]. Rashid et al. [24] employed NB with the bag of words model, reaching 87% accuracy and 35% recall, while Moheb et al. [21] reported up to 95% accuracy. Kanaan et al. [20] further demonstrated that NB attained 91% accuracy following demodulation and stop-word removal. Logistic regression (LR) is another common classification algorithm used in both binary and multi-class problems. Rashid et al. [24] used LR as a baseline model and, after balancing the dataset, improved the F1-score to 84% using TF-IDF features. Alfageh et al. [25] applied LR with TF-IDF, reporting results slightly lower by 1.8% compared to count vectorization. Lastly, the Extreme Gradient Boosting (XGBoost) algorithm has shown effectiveness in handling text data for cyberbullying detection, with Rashid et al. [24] reporting 85% accuracy using this approach.

2.2 Deep-learning Techniques

These methods have demonstrated impressive effectiveness in addressing the challenge of identifying cyberbullying in the Arabic language. For example, the researchers in [21], [25] developed a CNN-based model specifically tailored for this task. Their methodology involved four key steps: converting textual data into numerical representations, applying convolutional operations to extract significant features, reducing the convolution output to preserve only the most relevant information and finally feeding the processed data into a dense layer fully connected to all neurons in the network. This approach was tested on a dataset of 39,000 Arabic tweets collected *via* the Twitter API, achieving an impressive accuracy exceeding 95% without requiring manual intervention. Similarly, Banerjee et al. [26] extended the use of CNN to a larger dataset of 69,000 Arabic tweets, reporting an accuracy rate above 93%. In another study, Benaissa et al. [24] compared CNN with other deep-learning architectures, including Gated Recurrent Units (GRUs), Long Short-Term Memory (LSTM) and Bidirectional LSTM (BLSTM). Their analysis, conducted on a dataset of 32,000 Arabic comments from Aljazeera.net, showed that CNN outperformed the other models by a margin of one percent in the F1-score. Across the balanced dataset, these models collectively achieved an average F1-score of 84%. Further insights were provided by Srivastava et al. [27], who explored GRU, LSTM and BLSTM models for detecting objectionable content in online conversations. Their methodology incorporated rigorous data pre-processing steps, such as text cleaning, tokenization, stemming, lemmatization and stop-word removal prior to training the deep-learning algorithms. Among the models tested, BLSTM achieved the highest accuracy of 82.18%, followed closely by GRU (81.46%) and LSTM (80.86%). These results highlight the transformative potential of deep-learning techniques, particularly CNN, in enhancing the detection of cyberbullying within Arabic social media posts. Although these findings are promising, they also emphasize the need for continued research to further refine these models and effectively manage the growing volume and complexity of Arabic content on social-media platforms.

Building on the promising results of deep-learning techniques, such as CNN and RNN, in Arabic-cyberbullying detection, recent studies have explored hybrid and transformer-based approaches to further enhance performance. The study in [35] proposed a hybrid deep-learning model that combines LSTM networks with CNNs to detect cyberbullying in Arabic tweets. Their study focused on applying deep learning techniques to social-media data, specifically targeting the challenges of NLP. They demonstrated that their hybrid model outperformed several traditional ML algorithms, including SVM and NB, in terms of classification accuracy. While their contribution is significant, the study did not explicitly address dialectal variations within Arabic, nor did it elaborate on the size and linguistic diversity of the dataset used, which are important considerations in the context of Arabic social-media text. Abu Kwaik et al. [36] introduced an advanced methodology for identifying hate speech in Arabic tweets by integrating Recurrent Neural Network architectures—namely GRU and BiLSTM—with contextual word embeddings derived from AraBERT. Their experiments on dialectal Arabic-tweet datasets demonstrated strong discriminatory power, achieving an AUROC of approximately 0.84 in binary classification, 87.05% accuracy for the 2-class task, 78.99% for the 3-class task and 75.51% for the 6-class task. This study highlights the effectiveness of combining transformer-based embeddings with recurrent neural models when handling Arabic social-media content.

Building on these advances, a very recent study in [39] proposed state-of-the-art deep-learning techniques and provided comparative benchmarks closely aligned with the methodology of this research. The study applied a combination of CNN, RNN and transformer-based models to large-scale datasets of Arabic social-media content, emphasizing the importance of handling dialectal diversity and semantic nuances. Their results surpassed previous benchmarks, achieving improvements in both accuracy and F1 score metrics, demonstrating significant progress in the field between 2022 and 2024. Including such up-to-date research enhances the understanding of current capabilities and helps guide future work toward more robust cyberbullying-detection systems. Based on the previous studies referenced [13], [26], [23], [27], it has been observed that detecting bullying in the Arabic language remains a critical topic that requires significant attention in research. There is an urgent need for more studies on this topic. The existence of new technologies can help reduce the harmful impact of social media to prevent unwanted occurrences. Obeidat et al. [37] conducted a comparative study evaluating deep-learning models, such as RNN and CNN, against traditional machine-learning classifiers, like SVM and Random Forest for Arabic sentiment analysis on Twitter datasets. Their findings demonstrated that deep-learning approaches outperform traditional machine-learning methods in effectively handling the complexity and dialectal variations of Arabic social-media text. This is highly relevant to cyberbullying detection, which shares similar linguistic challenges. Our work extends these findings by applying RNN architectures on a larger, multi-dialectal dataset specifically focused on cyberbullying detection, further confirming the superior performance of deep-learning techniques over traditional models in Arabic NLP tasks. Earlier work by Al-Hassan and Al-Dossari [38] proposed one of the earliest benchmark datasets for Arabic-cyberbullying detection, compiling approximately 10,000 tweets labelled for offensive content. They evaluated both ML (Random Forest) and DL models (CNN, RNN), highlighting the promising performance of RNNs. Their dataset, however, is limited in scale and dialectal coverage. In contrast, our study utilizes a larger and more dialectally diverse dataset and focuses on standard RNN architectures, allowing for a more detailed exploration of their effectiveness in cyberbullying detection. Furthermore, other deep-learning models have also demonstrated promising results in Arabic-text classification tasks outside the cyberbullying domain. For instance, Jamaluddin et al. [43] proposed a multi-channel deep-learning model for Arabic news classification, emphasizing the importance of capturing semantic features through parallel architectures. Similarly, Al Qadi et al. [44] introduced a scalable shallow learning approach for tagging Arabic news articles, highlighting the benefits of lightweight models for Arabic NLP. These contributions further underline the growing applicability of both deep and shallow models across diverse Arabic-language NLP tasks.

While previous research demonstrates considerable progress using classical ML and deep learning for Arabic-cyberbullying detection, several gaps remain. Most studies rely on limited datasets with narrow dialectal coverage and modest sample sizes. The increasing linguistic complexity of Arabic social-media content necessitates larger, more diverse datasets and efficient deep-learning models. Our study addresses these gaps by utilizing extensive, dialect-rich datasets and focusing on RNN architectures that balance performance and complexity. This approach contributes to advancing robust cyberbullying detection in Arabic, complementing recent transformer-based innovations.

Therefore, a group of ML and deep-learning algorithms that were observed in the literature was chosen. Table 1 provides a summary of some of the literature on Arabic cyberbullying.

3. Materials and Methods

It is well known that the detection of a cyberbullying attack involves several steps, including data collection, visualization, pre-processing, feature extraction, model training and then model evaluation, as illustrated in Figure 1.

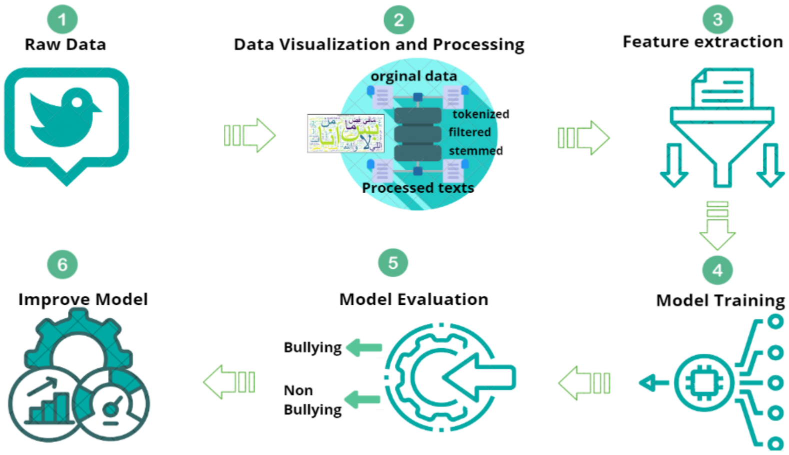


Figure 1. A general workflow of the proposed methodology.

3.1 Data Description

The data used in this research consists of public datasets published on Kaggle and divided into three separate and linguistically varied datasets, as shown in Table 2. The initial dataset, consisting of 5,846 Syrian/Lebanese political tweets, is included in the "Levantine Arab Hate Speech" dataset [42], which is divided into three groups: abusive tweets, hate-speech tweets and normal tweets. The second set, known as the "Arabic Sentiment Twitter Dataset Corpus" [43], consists of 56,795 Arabic tweets divided into two categories: positive and negative. The third group, "Arabic Dataset 1" [44], consists of a relatively small dataset of 1,100 tweets, divided into two categories using binary classification:

Table 1. Brief summary of the literature on Arabic cyberbullying.

Ref.	Classifier	Year	Dataset (Size)	Evaluation matric
[9]	XGBoost, NB.SVM, LR	2024	Twitter 9000 Tweets	Accuracy: 88%,78%, 84.4%, 83.95%
[14]	SVM	2021	Twitter API, (17.748 Tweets)	Accuracy: 85.49%
[15]	SVM, KNN, NB, RF	2020	X API, (4000 Tweets, Facebook2138Posts)	N/A
[24]	Deep Learning	2020	Aljazeera.net (32000 Comments)	Accuracy: 84%
[28]	NB	2023	YouTube Platform (4760 Comments)	Accuracy: 94%
[29]	SVM, NB	2024	Twitter and YouTube (30000 Tweets)	Accuracy: 95%, 70%
[30]	LSTM	2023	Twitter 10000 Tweets	Accuracy: 88%
[31]	MLP	2023	Twitter API 4140	Accuracy: 89%
[32]	LR, voting classifier, SVM	2024	Twitter 12000 Tweets	Accuracy: 65%, 71%, 98%
[33]	Codellama, DeepSeekCoder, Llama2	2025	10000Comments	Accuracy: 35%, 26%, 16.4%
[34]	AraBERT	2025	4240 Comments	N/A
[35]	Hybrid (CNN, LSTM)	2022	N/A	Accuracy: 97%
[36]	GRU and BiLSTM combined with contextual embeddings (AraBERT)	2023	N/A	Accuracy: 87.05% (2-class), 78.99% (3-class), 75.51% (6-class)
[39]	CNN, LSTM and BiLSTM	2025	50000 comments	Accuracy: 91%

negative speech (1) and positive or neutral speech (0). This data is characterized by the diversity of dialects used, including local dialects and classical Arabic, making it comprehensive and covering different linguistic styles in the Arab world. Tweets are categorized into two main categories: bullying, which contains offensive words or phrases and non-bullying, which does not. The final dataset comprises labels of 0 or 1 depending on whether the comment is bullying or not. Additionally, the data used is balanced, as shown in Figure 2.

Table 2. Data description.

Ref.	Group Name	No. of Tweets	Categories	Size - Notes
[42]	Arabic -Levantine Hate Speech	5846	Normal, Abusive, Hate	Syrian–Lebanese Politics
[43]	Arabic Twitter Sentiment Dataset	56795	Positive, Negative	Training 45275, Testing 11920
[44]	Arabic Dataset	1100	Negative, Positive	Relatively Small Data Size
Total		63741		

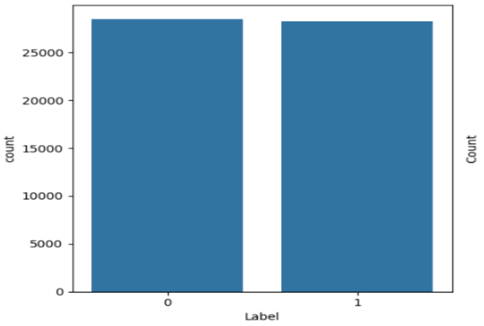
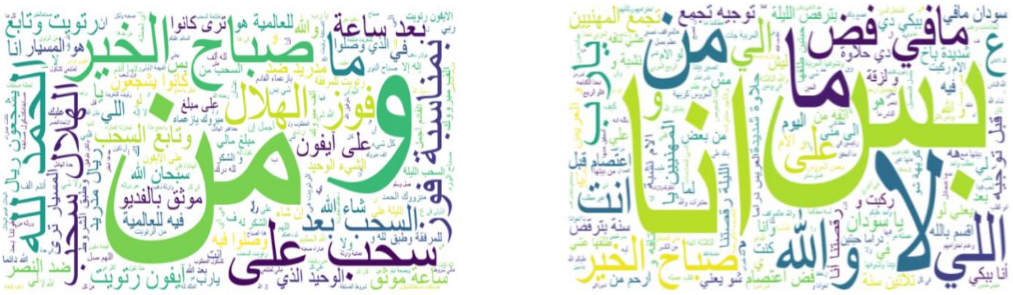


Figure 2. Distribution of positive and negative text, where (0 = Non-bullying, 1 = Bullying).

3.2 Data Visualization and Pre-processing

To know the most frequent words for bullying and non-bullying comments, this is expressed by displaying the word sizes, where large words are frequently repeated, as shown in Figure 3.



(a) (b)

Figure 3. Word cloud negative (a), Word cloud positive (b).

Figure 4 shows the representation of the most frequent words using the Count Vectorizer technique, where the frequency of words within texts is counted and converted into a numerical representation. The graph displays the twenty most frequently used words, ranked by frequency.

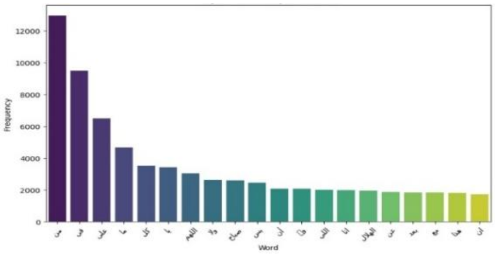


Figure 4. Count-vectorizer technique (top-20 most frequent words).

The first word is shown to have the highest frequency, occurring more than 250 times, followed by other words with decreasing frequencies. This representation is useful for understanding the distribution of words within the text data and discovering words that may be of high analytical interest in the context under study. The pre-processing stage is an important step in an ML technique, because it cleans and prepares the dataset, so that it can be used to train the model. In this study, the tweets are written in various dialects that differ from traditional Arabic. Therefore, we have used the NLP technique to address issues presented by comments on Twitter written in Arabic. This was applied in Figure 5.

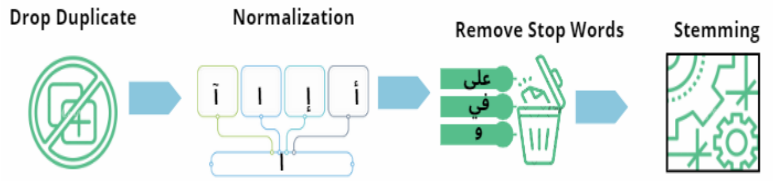
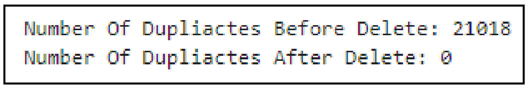


Figure 5. Data pre-processing main steps.

3.2.1 Removing Duplicates

There is a duplicate tweet; with bullying the duplicate count is 9896 and without bullying the duplicate count is 11122. So, by using the Python code, we remove these duplicates and they become zero duplicates, as shown in Figure 6.

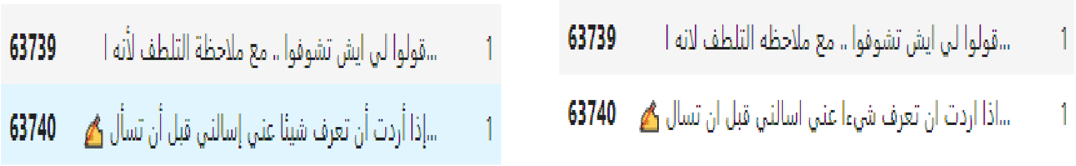


```
Number Of Dupliactes Before Delete: 21018
Number Of Dupliactes After Delete: 0
```

Figure 6. Remove duplicate.

3.2.2 Normalization

We applied the normalization to the dataset and converted it into a uniform text. The Python programming language implemented this process. It significantly contributes to improving the performance of models in ML tasks by reducing unnecessary linguistic variations. By converting texts into a standardized format, such as removing diacritics or similar characters, the model becomes better at understanding underlying patterns, which reduces noise in the data. This step leads to improved model accuracy and increased efficiency in handling unstructured and diverse texts, such as those found in cyberbullying tasks. In our study, we remove the English Letters, URLs, Hashtags, Special Characters and emojis. After applying the normalization process this led to the text being normalized and the result is shown here in Figure 7.



63739	قولوا لي ايش تشوفوا .. مع ملاحظة التلطف لأنه ا...	1
63740	إذا أردت أن تعرف شيئا عني إسالني قبل أن تسأل 👆...	1

(a)

63739	قولوا لي ايش تشوفوا .. مع ملاحظه التلطف لانه ا...	1
63740	اذا اردت ان تعرف شيءا عني اسالني قبل ان تسال 👆...	1

(b)

Figure 7. Text; before normalization (a); after normalization (b).

In Figure 7. (a), we see the tweets (“ إذا أردت أن تعرف شيئا همي إسألني قبل أن تسأل“), we see (“ي,ئ,أ”) converted into uniform text, as shown in Figure 7. (b). For example, (“أ,إ “) converted into (ا).

3.2.3 Removing Stop Words

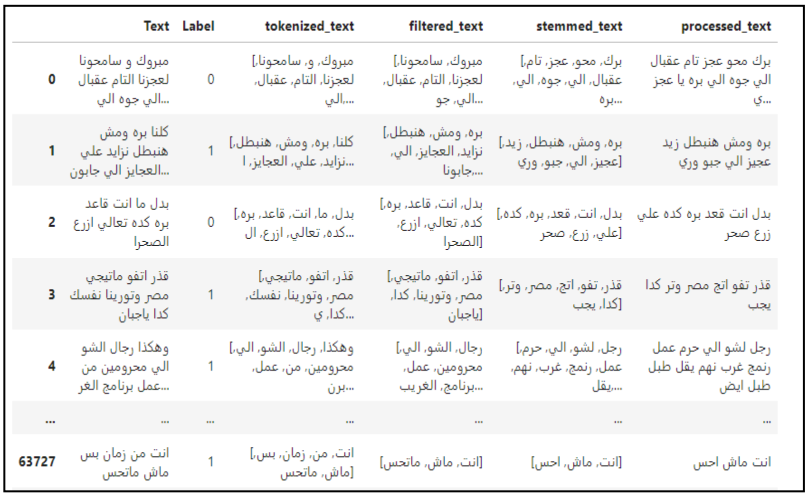
Stop words are meaningless in our study; we normalize text by removing the stop words. In applications, omitting standard words is a good way to implement and emphasize the most important words.

3.2.4 Tokenization

The texts were converted into words using natural-language units based on language rules defined by word boundaries. This step enabled the RNN model to treat each word as an independent unit within a sentence string, creating innovations in learning and processing context. When tokenization is carried out accurately, it makes it easier for the model to handle a wide variety of texts, such as those found in cyberbullying, which can include offensive words and complex phrases. Through good segmentation, the model can "understand" these offensive elements separately from other words, improving the accuracy of its predictions.

3.2.5 Stemming

In this step, words are reduced to their original roots by removing good suffixes or additions such as" ه" أو "ين" أو "ات". The goal of stemming is to reduce word diversity, helping the model understand that words derived from the same trigger have the same underlying meaning, such as "كتب" ,"كتابة" ,"كتاب" being reduced to "كتب". If stemming is applied effectively, it can improve the model's accuracy by reducing the variety of words associated with the same root. However, sometimes it can have a negative impact if it excessively reduces words, leading to weakened differentiation between important words. After we applied the pre-processing steps shown in Figure 5, Figure 8 shows a sample of the pre-processing phase.



	Text	Label	tokenized_text	filtered_text	stemmed_text	processed_text
0	مبروك و سامحونا لعجزنا التام عقبال الي جوه الي...	0	[مبروك, و, سامحونا, لعجزنا, التام, عقبال, الي,...	[مبروك, سامحونا, لعجزنا, التام, عقبال, الي, جو...	[برك, محو, عجز, تام, عقبال, الي, جوه, الي, بره...	برك محو عجز تام عقبال الي جوه الي بره يا عجز ي...
1	كلنا بره ومش هنبطل نزايد علي العجايز الي جابون...	1	[كلنا, بره, ومش, هنبطل, نزايد, علي, العجايز, ا...	[بره, ومش, هنبطل, نزايد, العجايز, الي, جابونا,...	[بره, ومش, هنبطل, زيد, عجيز, الي, جبو, وري]	بره ومش هنبطل زيد عجيز الي جبو وري
2	بدل ما انت قاعد بره كده تعالي ازرع الصحرا	0	[بدل, ما, انت, قاعد, بره, كده, تعالي, ازرع, ال...	[بدل, انت, قاعد, بره, كده, تعالي, ازرع, الصحرا]	[بدل, انت, قعد, بره, كده, علي, زرع, صحر]	بدل انت قعد بره كده علي زرع صحر
3	قذر اتفو ماتيجي مصر وتورينا نفسك كدا ياجبان	1	[قذر, اتفو, ماتيجي, مصر, وتورينا, نفسك, كدا, ي...	[قذر, اتفو, ماتيجي, مصر, وتورينا, كدا, ياجبان]	[قذر, تفو, اتج, مصر, وتر, كدا, يجب]	قذر تفو اتج مصر وتر كدا يجب
4	وهكذا رجال الشو الي محرومين من عمل برنامج الغر...	1	[وهكذا, رجال, الشو, الي, محرومين, من, عمل, برن...	[رجال, الشو, الي, محرومين, عمل, برنامج, الغريب...	[رجل, لشو, الي, حرم, عمل, رنمج, غرب, نهم, يقل,...	رجل لشو الي حرم عمل رنمج غرب نهم يقل طبل طبل ايض
...	...	...	...	...	...	...
63727	انت من زمان بس ماش ماتحس	1	[انت, من, زمان, بس, ماش, ماتحس]	[انت, ماش, ماتحس]	[انت, ماش, احس]	انت ماش احس

Figure 8. A sample of the pre-processed data.

As Figure 8 illustrates, there are six columns. The first and second columns represent the dataset before processing. Column 3 (“Tokenize Text”) shows how the text is tokenized into small words. Column 4 (“Filter Text”) displays the result after removing meaningless data using stop words. Column 5 (“Stem Text”) shows the text converted into its original form in Arabic and the final column presents the uniform data after pre-processing. The study then investigates the best model features that yielded the highest accuracy to identify the most effective ML algorithms for detecting cyberbullying in Arabic tweets using TF-IDF techniques. In this study, Twitter tweets are categorized into two groups: bullying and non-bullying. The TF-IDF feature-extraction method was employed to enhance the textual data representation by measuring the importance of terms within individual tweets relative to the entire dataset. Additionally, the study utilized n-grams to analyze the sequences of words rather than isolated terms, allowing the capture of contextual information. This significantly improved the model's understanding of language patterns associated with bullying behavior. This approach was essential for classifying tweets in a relevant and accurate manner.

3.3 Machine-learning Classification and Tuning

3.3.1 Support Vector Machine

In this study, the SVM algorithm was used as one of the basic ML techniques for tweet classification and cyberbullying analysis. SVM is an effective tool for handling high-dimensional text data and finding the best hyperplane between different categories, such as bullying-free tweets and bullying tweets. SVM has been applied to text features extracted using NLP techniques, such as converting text into numerical representation *via* TF-IDF vectorization. The algorithm has improved classification accuracy thanks to its ability to handle multi-dimensional text, especially in light of the diversity of dialects and linguistic patterns within the dataset [9].

3.3.2 Naïve Bayes

In this study, Arabic tweets were categorized into cyberbullying-related groups using the NB algorithm. As a result of its effectiveness and simplicity, NB was an appropriate option when handling huge and high-dimensional data. The algorithm's output also showed strong performance in rapidly and precisely gathering data, which aided in the efficient identification of cyberbullying in tweets [9].

3.3.3 Logistic Regression

In this study, Arabic tweets were analyzed using LR as a classification method and they were divided into two groups: cyberbullying and non-bullying. LR is a good choice for this kind of data, because it can handle binary problems well and has shown promise in identifying the correlation between textual characteristics and the degree of bullying in tweets. Obtaining precise and comprehensible classification models was also beneficial [31].

3.3.4 Extreme Gradient Boosting

XGBoost was employed in this study as a technique to classify Arabic tweets into two groups: those that involved cyberbullying and those that did not. XGBoost was selected because of its top-ranking performance and good accuracy in handling data with many different dimensions. Furthermore, the XGBoost method enhances performance by employing strategies, like regularization to lessen overfitting and enhance generalization [31].

3.3.5 Deep-learning Approach

In this research, RNNs were used to analyze Arabic tweets related to cyberbullying and categorize them into two classes: "bullying" and "non-bullying." This technique was selected due to its ability to recognize sequential patterns in text data, such as understanding context within a series of words. RNNs are particularly well-suited for tweet analysis, as they account for the chronological order of words and expressions, helping identify offensive messages influenced by contextual nuances.

The model was trained using Keras’s sequential interface, incorporating an embedding layer, followed by a simple RNN layer and ending with a dense layer. The embedding layer transformed words into numerical representations, with an input dimension of 5,000 and an output dimension of 64. The sequence length was determined based on the maximum length of tweets in the dataset.

The training process was conducted using a set of pre-defined hyperparameters, with multiple tests performed to determine the optimal configuration. The model was initialized with random weights and the number of units in the output layer was tailored for binary classification, as the task requires categorizing tweets into two classes: "bullying" and "non-bullying."

The learning rate was optimized using the Adam optimizer, chosen for its efficiency in training deep-learning models. The batch size was set to 64, enabling the model to process a sufficient number of samples per iteration. The training spanned 27 epochs. The text data was also classified using a combination of ML and DL algorithms to enhance performance and identify the optimal model. The ML algorithms included SVC, LR and NB, while RNNs represented the deep-learning component. Texts were transformed into numerical representations using the TF-IDF technique and hyperparameters were fine-tuned using GridSearch to achieve the best possible performance. Below is a summary of the parameter settings used for each algorithm [24].

3.3.6 Hyper-parameter Fine-tuning and Evaluation Measures

Several algorithms were used with parameter adjustments to enhance performance. In SVC, the parameter *C* was set to control regularization, *max_iter* for the number of iterations, *length* for defining stopping criteria and *TF-IDF max_features* to specify the number of words considered in the TF-IDF representation. In LR, *C and solver* were configured to select the solution method, along with adjustments to *TF-IDF max_features* and *TF-IDF ngram_range* to define the range of words considered. In NB, the *alpha* parameter was used to regulate the influence of rare words and *TF-IDF ngram_range* was used to define the word range considered in the model. In the RNN model, the learning rate was determined using the Adam optimizer. The parameters *input_dim*, *output_dim* and *input_length* were set to properly format the text input, while *epochs* and *batch size* were selected for the training process. All models used TF-IDF to convert textual data into a numerical format and *GridSearch* was employed to determine the optimal values for each model's parameters.

3.3.7 Model Generation and Evaluation

In this study, the Python, Sklearn and XGBoost libraries were used to develop four supervised ML models to classify the data. The results were evaluated using several performance metrics, including accuracy as given in Equation (1), precision as given in Equation (2), recall as given in Equation (3) and F1-score as given in Equation (4). These measures were calculated using the following equations [31], where TP is the true positives, TN is the true negatives, FP is the false positives and FN is the false negatives.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

$$F1 - Score = \frac{2*(precision*recall)}{(precision+recall)} \tag{4}$$

We also calculated the F1-score computed at the class level (Macro-F1) and at the sample level (Micro-F1). The Macro-F1 was calculated as the simple average of the F1 scores for each class and the Micro-F1 was calculated based on the confusion matrix, which takes into account all true positives, false positives, false negatives and true negatives across all classes, as follows:

$$Macro - F1 = \frac{F1\ class_0+F1\ class_1}{2} \tag{5}$$

$$Micro - F1 = \frac{TP+TN+FP+FN}{TP} \tag{6}$$

4. Results

Experiments were conducted to analyze the performance of models used in text classification, using ML and deep-learning algorithms with parameter adjustment to improve accuracy. The aim was to compare the effectiveness of the models and choose the most appropriate for the available data. The results are presented below.

4.1 Experiment Results Using Different Machine-learning Algorithms

The models were built and tested using a dataset collected and processed for this study. The collection contains 42,723 tweets after initial processing, obtained from Kaggle. This study used four ML algorithms: SVM, NB, LR and XGBoost. The performance of these models was evaluated using measures of accuracy, precision, recall and F1-score, as shown in Table 3.

Table 3. Experimental results of various machine learning methods.

ML	Feature Extraction	Accuracy	Precision		Recall	
SVM	TF - IDF	75%	0	76%	0	76%
			1	75%	1	75%
NB	TF - IDF	72%	0	73%	0	73%
			1	72%	1	72%
LR	TF - IDF	74%	0	76%	0	74%
			1	73%	1	75%
XGBoost	TF - IDF	74%	0	77%	0	70%
			1	71%	1	78%

Table 3 compares the performance of four models (SVM, NB, LR and XGBoost) in the cyberbullying-detection task using the TF-IDF feature-extraction method. Each model's performance was evaluated based on the mentioned metrics. The SVM model performed best, recording 75% accuracy, 76% precision, 76% recall and 76% F1-score. This makes it the most effective of all models, showing a good balance across all metrics. The NB model recorded 72% accuracy, 73% precision, 73% recall and 73% F1-score. Despite its weaker performance compared to SVM, it still offers acceptable results, particularly in recall. The LR model achieved 74% accuracy, 76% precision, 72% recall and 75% F1-score. LR performed close to SVM, but was lower in terms of recall and F1 score. The XGBoost model showed balanced performance, achieving 74% accuracy, 77% precision, 70% recall and 74% F1-score. XGBoost outperformed other models in terms of precision, with the highest score (77%), demonstrating its ability to make more accurate positive predictions. Although the initial results obtained using traditional ML algorithms were acceptable, they were not sufficient to meet the required objectives. Therefore, the accuracy and overall performance of the model were enhanced by applying deep-learning techniques using RNN.

4.2 Experiment Results Using RNN

To improve model performance and achieve better outcomes, we transitioned to using deep learning, with a focus on RNNs, to process the same large and complex dataset. During our experiments, neural networks demonstrated their ability to outperform traditional algorithms. In the first experiment, the model was trained for 20 epochs, resulting in an excellent accuracy of 96%. In the second experiment, we used 27 epochs and achieved an accuracy of 97%. These findings highlight the high proficiency of deep-learning techniques in extracting complex patterns from large datasets and underscore their significance as an effective approach to enhancing performance in this context. Table 4 presents the results of the experiment using RNNs.

Table 4. Experimental results of the deep-learning approach.

Experiments	Classifier	Accuracy	Precision		Recall		F1- Score
Experiment 1	RNN	96%	0	97%	0	94%	96%
			1	94%	1	97%	
Experiment 2	RNN	97%	0	97%	0	96%	97%
			1	96%	1	97%	

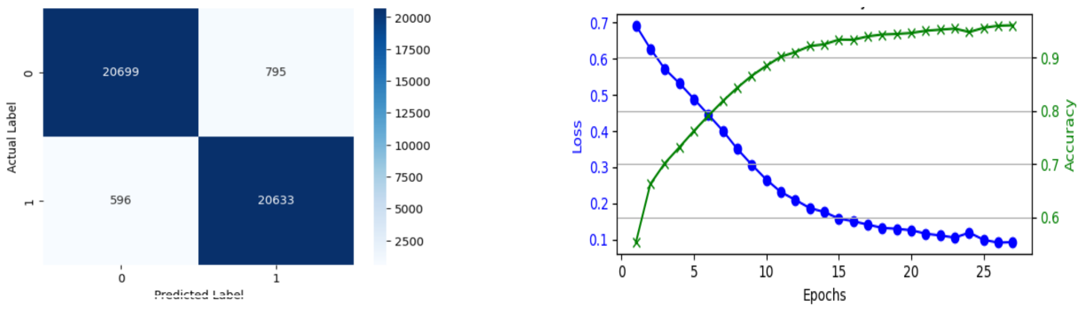
To ensure fair evaluation, we report precision, recall and F1-score separately for each class (0: non-bullying, 1: bullying). As shown in Experiment 2, Table 5, the model achieved high precision and recall for both classes (class 0: 97% recall, 96% F1-score; class 1: 96% precision, 97% F1-score), indicating balanced performance and minimal bias. In addition to per-class metrics, we computed the macro-averaged F1-score (97%) and micro-averaged F1-score (97%), confirming consistent performance across classes. We also include the confusion matrix to visualize the distribution of true positives, false positives, true negatives and false negatives, further supporting the reliability of our results.

Furthermore, the training process was stable, as evidenced by accuracy and loss curves, which show no signs of overfitting. These results highlight the model's robustness and its ability to distinguish between bullying and non-bullying instances effectively.

The attached diagrams illustrate the performance of the RNN model used in the experiment. Figure 9(a) shows the confusion matrix, which reflects the model's prediction accuracy, where the values in the cells indicate the number of correctly and incorrectly classified cases. For example, the model correctly classified 20,699 instances of the negative category (0) and 20,633 instances of the positive category (1), while misclassifications were limited to 795 and 596, respectively. These results indicate strong performance in data classification.

Figure 9(b) presents the loss and accuracy curve. This curve illustrates the relationship between the number of epochs and the corresponding values of loss and accuracy. The loss is shown to continuously decrease as the number of epochs increases, indicating the model's learning progress and improvement. Conversely, accuracy steadily increases to high levels, reflecting model stability and the ability to achieve accurate results over time.

Figure 10 displays the ROC Curve, which is used to evaluate model performance by comparing the True Positive Rate (TPR) with the False Positive Rate (FPR). The curve indicates that the model achieved an Area Under the Curve (AUC) of 97%, reflecting high effectiveness in distinguishing between categories. These results demonstrate the model's efficiency and its ability to process and classify data with high accuracy.



(a) (b)

Figure 9. (a) Confusion matrix and (b) Loss and accuracy curve for RNN.

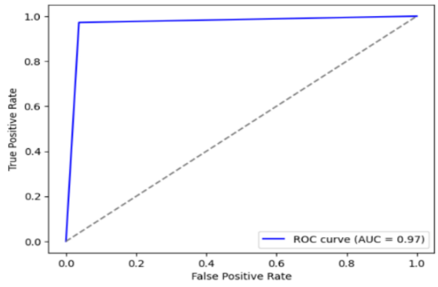


Figure 10. ROC curve for RNN results.

An important aspect of evaluating model performance involves analyzing false positives (FP) and false negatives (FN), as they directly impact the precision and recall scores, especially in sensitive tasks, like cyberbullying detection. As shown in the RNN confusion matrix (Figure 9A), the model misclassified 795 non-bullying tweets as bullying (false positives) and 596 bullying tweets as non-bullying (false negatives). While both types of errors are undesirable, false negatives are particularly critical in this context, as failing to identify a bullying instance could allow harmful content to persist unflagged. However, the low number of false negatives relative to the total sample suggests strong recall, particularly for the bullying class (97%). Similarly, the limited number of false positives supports the model's high precision (96%) in identifying actual bullying content without over-flagging benign posts. This balance between FP and FN reinforces the model's robustness and practical reliability in real-world applications.

5. DISCUSSION

To compare the proposed system with the latest available methods, we used a quantitative comparison between studies by selecting some recent studies that share three common aspects (language (Arabic), social media (Twitter) and data-collection source). However, the dataset used in those studies is different from the proposed one. In this regard, a comparison was made with three recent methods from 2023 [29]-[31], [35]-[36].

Table 5. Comparison between the proposed approach and state-of-the-art.

Approach	Feature Extraction	Classifier	Accuracy	Precision	Recall	F1-Score
[29]	TF- IDF, Wob	SVM, NB	95%,70%	92%	84%	88%
[30]	Non	LSTM	88%	88%	88%	88%
[31]	TF- IDF	MLP	89%	88%	90%	89%
[35]	Automatic feature extraction	hybrid LSTM-CNN	87.8%	N/A	83.6%	84.1%
[36]	AraBERT embeddings	GRU and BiLSTM with AraBERT embeddings	Accuracy: 87.05% (2-class), 78.99% (3-class), 75.51% (6-class)	N/A	N/A	N/A
[39]	standard text embeddings	CNN, LSTM and BiLSTM architectures.	91%	N/A	N/A	N/A
Proposed	TF- IDF	RNN	**97%**	**97%**	**97%**	**97%**

Based on the results shown in Table 5, previous studies utilizing traditional machine-learning techniques, such as SVM and NB, with feature-extraction methods, like TF-IDF and Bag of Words, have reported accuracies reaching up to 95%. However, these studies typically relied on smaller datasets, which may have contributed to inflated performance metrics due to reduced complexity. In contrast, our approach employed a standard RNN model trained on a large-scale (42,000 samples), multi-dialectal dataset, achieving an accuracy of 97%. This underscores the capacity of deep-learning models, particularly RNNs, to generalize effectively across more diverse and complex data, outperforming traditional algorithms when evaluated on a broader scale. Our findings are consistent with Obeidat et al. [37], who demonstrated that deep-learning models, such as RNNs, significantly outperformed traditional machine-learning approaches (e.g. SVM, Random Forest) in Arabic sentiment analysis on Twitter. This further supports the superiority of neural architectures in handling complex linguistic features in Arabic social-media content. The dataset referenced in [29]-[31], [35]-[36], [39] was used to evaluate the performance of our algorithms. In the broader context of Arabic-cyberbullying detection, our results extend prior literature by emphasizing the impact of both dataset size and dialectal diversity. For example, Al-Hassan and Al-Dossari [38] introduced one of the earliest benchmark datasets (~11K tweets) and reported an F1-score of 73% using CNN-LSTM models. Our study, leveraging a more comprehensive dataset, achieved significantly higher F1-scores using a simpler RNN architecture, highlighting the value of rich data over architectural complexity. Similarly, Al-Azani and El-Alfy [35] proposed a hybrid CNN-LSTM model that attained an F1-score of 84.1%. Despite their more intricate design, our RNN-based model achieved comparable or superior accuracy without relying on hybrid or ensemble methods, affirming that a well-optimized standard RNN can deliver state-of-the-art results when trained on appropriate data. Furthermore, Abu Kwaik et al. [36] combined GRU/BiLSTM models with AraBERT embeddings, reporting an AUROC of 0.84 and accuracies ranging from 75% to 87% across various classification tasks. Although their use of transformer-based contextual embeddings enhanced performance, our model demonstrated that even without such embeddings, classical RNNs can achieve competitive results, particularly when trained on diverse and large-scale datasets. In addition, the recent study by Alshahrani et al. [39] employed CNN, LSTM and BiLSTM architectures on a dataset of approximately 50,000 Arabic tweets, achieving an accuracy above 94%. However, their work did not focus on dialectal diversity or use RNNs. By contrast, our approach incorporated three distinct Arabic dialects and applied a standard RNN, achieving superior accuracy. This demonstrates that simpler architectures, when supported by carefully curated and dialect-diverse data, can outperform more complex models lacking linguistic variation. Collectively, these comparisons reinforce two critical conclusions of our study: (1) the effectiveness of deep learning in Arabic-cyberbullying detection is closely tied to dataset size and dialectal diversity and (2) standard RNN architectures remain a viable and efficient alternative to more complex hybrid or transformer-based models.

Although the size of the dataset was limited in the previous studies, applying the RNN algorithm yielded outstanding results. Regarding reference [29], we used their same dataset and applied our proposed RNN model to it. Our approach achieved an accuracy of 99.6%, compared to 95% reported in [29] using SVM. This confirms the superiority of our method, since the improvement was demonstrated on the same dataset under comparable conditions. Therefore, the performance gain is not only due to the size or structure of the dataset, but is directly related to the effectiveness of the proposed RNN-based architecture in capturing sequential patterns in Arabic text better than traditional classifiers, such as SVM. The comparison of the proposed approach with a closely related study is shown in Table 6.

Table 6. Comparison of the proposed approach with a closely related study [29].

Approach	Classifier	Accuracy	Precision	Recall	F1- Score
Proposed	RNN	**99.6%**	**99%**	**98%**	**98%**
[29]	SVM	95%	92%	84%	88%

While this study has demonstrated the effectiveness of the proposed algorithm in detecting cyberbullying in Arabic text, several limitations should be addressed in future work. Firstly, one challenge lies in the imbalance of the dataset, as the amount of cyberbullying content is often significantly lower than neutral or non-bullying content. This can affect the performance of the model and lead to a bias towards the majority class. In the future, techniques, such as data augmentation and oversampling, can be explored to balance the dataset and improve the detection accuracy. Furthermore, while our model achieved promising results, it may struggle to accurately interpret the context in longer and more complex sentences. In future studies, hybrid models combining RNNs with Transformers [45] could be explored to leverage the strengths of both approaches. Transformers, with their ability to capture long-range dependencies, could complement the sequential learning nature of RNNs, improving the overall model's understanding of the context. Moreover, challenges related to the diverse use of language and slang in cyberbullying cases, especially in Arabic, require further attention. Future research could focus on developing advanced pre-processing techniques and word embeddings to more effectively handle such linguistic variations. Finally, while this study provides valuable insights into cyberbullying detection using deep learning, future work should focus on overcoming these limitations through the integration of advanced techniques, such as hybrid models and better handling of data imbalance and contextual complexities.

6. Conclusions

Cyberbullying is becoming increasingly difficult to detect, as users can bully without being identified. Cyberbullying poses a threat to individuals and can lead to suicide or depression among victims, making its detection essential. Although there are many studies on this topic, most of them have focused on the English language, while there are only a few studies in Arabic. In the current study, we proposed and trained a different ML model to detect cyberbullying in Arabic comments of tweets from different dialects. This study achieved significant improvements in the performance of the proposed model using feature-extraction techniques. RNNs produced the best results when utilizing 27 echoes in perfect time.

References

[1] W. N. H. Wan Ali, M. Mohd and F. Fauzi, "Cyberbullying Detection: An Overview," Proc. of the 2018 Cyber Resilience Conf. (CRC), pp. 1–6, DOI: 10.1109/CR.2018.8626869, Putrajaya, Malaysia, 2018.

[2] B. Srinandhini and J. I. Sheeba, "Online Social Network Bullying Detection Using Intelligence Techniques," Procedia Computer Science, vol. 45, pp. 485–492, DOI:10.1016/j.procs.2015.03.085, 2015.

[3] TechJury, "50 Alarming Cyberbullying Statistics to Know in 2024," [Online], Available: https://techjury.net/blog/cyberbullying-statistics/, Accessed: Jan. 2, 2025.

[4] Cyberbullying Research Center, "2023 Cyberbullying Data - Cyberbullying Research Center," [Online], Available: https://cyberbullying.org/2023-cyberbullying-data, Accessed: Aug. 27, 2024.

[5] Statista, "COVID-19 Vaccine: Adverse Events by Age and Gender in Spain," [Online], Available: https://www.statista.com/statistics/1220543/covid-19-vaccine-number-of-adverse-events-reported-by-age-and-gender-spain/, Accessed: May 10, 2025.

[6] UNICEF, "Search | UNICEF," [Online], Available: https://www.unicef.org/search?query=Statistic+cybe Rbullying, Accessed: May 10, 2025.

[7] 7amleh, "7amleh - Annual Report 2023," [Online], Available: https://7amleh.org/annual23/eng/, Accessed: May 10, 2025.

[8] Ditch the Label, "Youth Charity | Mental Health, Bullying & Relationships," [Online], Available: https://www.ditchthelabel.org/cyber-bullying-statistics-what-they-tell-us, Accessed: Aug. 27, 2024.

[9] D. Musleh et al., "A Machine Learning Approach to Cyberbullying Detection in Arabic Tweets," Computers, Materials and Continua, vol. 80, no. 1, pp. 1033–1054, Jul. 2024.

[10] Statista, "Most Used Languages Online by Share of Websites 2024," [Online], Available: https://www.statista.com/statistics/262946/most-common-languages-on-the-internet/, Aug., 2024.

[11] A. Alqarni and A. Rahman, "Arabic Tweets-based Sentiment Analysis to Investigate the Impact of COVID-19 in KSA: A Deep Learning Approach," Big Data and Cognitive Computing, vol. 7, no. 1, p. 16, DOI: 10.3390/bdcc7010016, Jan. 2023.

[12] W. J. Hutchins, "The Georgetown-IBM Experiment Demonstrated in January 1954," Lecture Notes in Computer Science, vol. 3265, pp. 102–114, DOI: 10.1007/978-3-540-30194-3_12, 2004.

[13] A. Mandal, "Evolution of Machine Translation," Towards Data Science, [Online], Available: https://towardsdatascience.com/evolution-of-machine-translation-5524f1c88b25, Aug. 27, 2024.

[14] S. Almutiry, M. Abdel Fattah and S. Arabia-Almadinah Almunawarah, "Arabic CyberBullying Detection Using Arabic Sentiment Analysis," Egyptian Journal of Language Eng., vol. 8, no. 1, pp. 39–50, 2021.

[15] T. Kanan, A. Aldaaja and B. Hawashin, "Cyber-Bullying and Cyber-Harassment Detection Using Supervised Machine Learning Techniques in Arabic Social Media Contents," Journal of Internet Technology, vol. 21, no. 5, pp. 1409–1421, DOI: 10.3966/160792642020092105016, Sep. 2020.

[16] I. Abu El-Khair, "Effects of Stop Words Elimination on Arabic Information Retrieval," International Journal of Computing & Information Sciences, vol. 4, no. 3, pp. 119–133, 2006.

[17] M. A. Al-Ajlan and M. Ykhlef, "Deep Learning Algorithm for Cyberbullying Detection," Int. J. of Advanced Computer Science and Applications, vol. 9, no. 9, pp. 199-205, 2018.

[18] B. Haidar, M. Chamoun and A. Serhrouchni, "Arabic Cyberbullying Detection: Using Deep Learning," Proc. of the 2018 7th Int. Conf. on Computer and Communication Engineering (ICCCE), pp. 284–289, DOI: 10.1109/ICCCE.2018.8539303, Kuala Lumpur, Malaysia, Nov. 2018.

[19] B. Haidar, M. Chamoun and A. Serhrouchni, "A Multilingual System for Cyberbullying Detection: Arabic Content Detection Using Machine Learning," Advances in Science, Technology and Engineering Systems J., vol. 2, no. 6, pp. 275–284, DOI: 10.25046/AJ020634, 2017.

[20] B. Y. Alharbi et al., "Automatic Cyber Bullying Detection in Arabic Social Media," Int. J. of Engineering Research & Technology, vol. 12, pp. 2330–2335, 2019.

[21] D. Mouheb et al., "Detection of Arabic Cyberbullying on Social Networks Using Machine Learning," Proc. of the 2019 IEEE/ACS 16th Int. Conf. on Computer Systems and Applications (AICCSA), DOI: 10.1109/AICCSA47632.2019.9035276, Abu Dhabi, UAE, Nov. 2019.

[22] K. Reynolds et al., "Using Machine Learning to Detect Cyberbullying," Proc. of the 10th Int'l Conf. Mach. Learn. Appl. (ICMLA), vol. 2, pp. 241–244, DOI: 10.1109/ICMLA.2011.152, Honolulu, USA, 2011.

[23] J. Hani et al., "Social Media Cyberbullying Detection Using Machine Learning," Int. J. of Advanced Computer Science and Applications, vol. 10, no. 5, pp. 703–707, 2019.

[24] B. A. Rachid et al., "Classification of Cyberbullying Text in Arabic," Proc. of the IEEE Int. Joint Conf. on Neural Networks (IJCNN), DOI: 10.1109/IJCNN48605.2020.9206643, Glasgow, UK, Jul. 2020.

[25] T. D. Alsubait, "Comparison of Machine Learning Techniques for Cyberbullying Detection on YouTube Arabic Comments," Int. J. of Computer Science and Network Security, vol. 21, no. 1, pp. 1–5, 2021.

[26] V. Banerjee et al., "Detection of Cyberbullying Using Deep Neural Network," Proc. of the IEEE 2019 5th Int. Conf. on Advanced Computing & Communication Systems (ICACCS), pp. 604–607, DOI: 10.1109/ICACCS.2019.8728378, Coimbatore, India, Mar. 2019.

[27] C. Iwendi et al., "Cyberbullying Detection Solutions Based on Deep Learning Architectures," Multimedia Systems, vol. 29, no. 3, pp. 1839–1852, DOI: 10.1007/S00530-020-00701-5, Jun. 2023.

[28] D. A. Musleh et al., "Arabic Sentiment Analysis of YouTube Comments: NLP-based Machine Learning Approaches for Content Evaluation," Big Data and Cognitive Computing, vol. 7, no. 3, p. 127, Jul. 2023.

[29] K. T. Mursi et al., "ArCyb: A Robust Machine-learning Model for Arabic Cyberbullying Tweets in Saudi Arabia," Int. J. of Advanced Computer Science and Applications, vol. 14, no. 9, pp. 1059–1067, 2023.

[30] M. Alzaqebah et al., "Cyberbullying Detection Framework for Short and Imbalanced Arabic Datasets," J. of King Saud Uni. - Computer and Information Sciences, vol. 35, no. 8, p. 101652, Sep. 2023.

[31] A. M. Alduailaj and A. Belghith, "Detecting Arabic Cyberbullying Tweets Using Machine Learning," Machine Learning and Knowledge Extraction, vol. 5, no. 1, pp. 29–42, Jan. 2023.

[32] M. Khairy et al., "Comparative Performance of Ensemble Machine Learning for Arabic Cyberbullying and Offensive Language Detection," Language Resources and Evaluation, vol. 58, no. 2, pp. 695–712, DOI: 10.1007/S10579-023-09683-Y, Jun. 2024.

[33] A. H. Zahid et al., "Evaluation of Hate Speech Detection Using Large Language Models and Geographical Contextualization," arXiv, arXiv: 2502.19612, Feb. 2025.

[34] A. Charfi et al., "Hate Speech Detection with ADHAR: A Multi-dialectal Hate Speech Corpus in Arabic," Frontiers in Artificial Intelligence, vol. 7, p. 1391472, DOI: 10.3389/FRAI.2024.1391472, May 2024.

[35] A. Altayeva et al., "Hybrid Deep Learning Model for Cyberbullying Detection on Online Social Media Data," Int. J. of Computer Science, vol. 8, no. 3, Sep. 2022.

[36] A. Alhazmi et al., "Code-mixing unveiled: Enhancing the hate speech detection in Arabic dialect tweets using machine learning models," PLOS One, vol. 19, no. 7, p. e0305657, 2024.

[37] R. Obeidat et al., "Deep Learning *vs.* Traditional Machine Learning for Arabic Sentiment Analysis: A Comparative Study," Int. J. of Advanced Computer Science and Appl., vol. 12, no. 4, pp. 188–195, 2021.

[38] A. Al-Hassan and H. Al-Dossari, "A Benchmark Dataset for Arabic Cyberbullying Detection on Twitter: Design and Evaluation," Int. J. of Advanced Computer Science and Appl., vol. 11, no. 2, pp. 72–78, 2020.

[39] G. Jaradat et al., "Deep Learning Approaches for Detecting Cyberbullying on Social Media," J. of Computational and Cognitive Engineering, vol. 2025, no. 00, pp. 1–15, Mar. 2025.

[40] I. Jamaleddyn, R. El Ayachi and M. Biniz, "Novel Multi-channel Deep Learning Model for Arabic News Classification," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 10, no. 4, pp. 453–468, DOI: 10.5455/jjcit.71-1720086134, Dec. 2024.

[41] L. Al Qadi, H. El Rifai, S. Obaid and A. Elnagar, "A Scalable Shallow Learning Approach for Tagging Arabic News Articles," Jordanian Journal of Computers and Information Technology (JJCIT), vol. 6, no. 3, pp. 263–280, DOI: 10.5455/jjcit.71-1585409230, Sep. 2020.

[42] Haithem Hermessi, "Arabic Levantine Hate Speech Detection," [Online], Available: https://www.kaggle.com/datasets/haithemhermessi/arabic-levantine-hate-speech-detection, Jan. 2025.

[43] M. K. Saad, "Arabic Sentiment Twitter Corpus," [Online], Available: https://www.kaggle.com/datasets/mksaad/arabic-sentiment-twitter-corpus, Jan. 2025.

[44] A. Saleh, "Arabic Dataset1," [Online], Available: https://www.kaggle.com/datasets/ahmadsaleh2001/arabicdataset1, Jan. 2025.

[45] M. Tami et al., "Transformer-based Approach to Pathology Diagnosis Using Audio Spectrogram," Information, vol. 15, no. 5, p. 253, DOI: 10.3390/info15050253, 2024.

ملخص البحث:

لقد ظهر التّنمُّر السّيبراني كقضية مُلحّة في العصر الرّقمي، وبخاصة في المجتمعات المتحدّثة بالعربية حيث لا يزال البحث في هذا المجال محدوداً. وهذه الورقة تبحث في الكشف عن التّنمُّر السّيبراني بالعربية على وسائل التّواصل الاجتماعي باستخدام تقنيات التّعلُّم الآلي التّقليدية، وتقنيات التّعلُّم العميق. وقد جرى استخدام مجموعة بيانات لتغريدات بالعربية مُتاحة للعموم لتدريب وتقييم عدّة نماذج تعلُّم آلي، إلى جانب نموذج تعلُّم عميق قائم على شبكة عصبية (RNN).

وقد أثبتت النّتائج أنّ النّموذج القائم على الشّبكة العصبية تفوّق على النّماذج الأخرى المستندة إلى التّعلُّم الآلي، الأمر الّذي يُشير إلى فاعلية التّعلُّم العميق في التّحديد الدّقيق للمحتوى السّيّء في النّصوص المكتوبة باللُّغة العربية. وتؤكّد النّتائج ضرورة دمْج البيانات الغنية لغوياً والبِنى المتقدّمة المرتكزة إلى الشّبكات العصبية لتحسين عمل أنظمة الكشف عن التّنمُّر السّيبراني في اللُّغات المختلفة، وبخاصّةٍ العربية؛ لما تنطوي عليه من تعقيد وتنوُّع.