Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.062724

ARTICLE

A Deep Learning Framework for Arabic Cyberbullying Detection in Social Networks

Yahya Tashtoush[1,*], Areen Banysalim[1], Majdi Maabreh[2], Shorouq Al-Eidi[3], Ola Karajeh[4] and Plamen Zahariev[5]

[1]Computer Science Department, Jordan University of Science and Technology, Irbid, 22110, Jordan
[2]Department of Information Technology, Faculty of Prince Al-Hussein Bin Abdallah II For Information Technology, The Hashemite University, Zarqa, 13133, Jordan
[3]Computer Science Department, Tafila Technical University, Tafila, 66110, Jordan
[4]Department of Digital Media Software Engineering, Ferris State University, Big Rapids, MI 49503, USA
[5]Department of Telecommunications, University of Ruse "Angel Kanchev", Ruse, 7017, Bulgaria

*Corresponding Author: Yahya Tashtoush. Email: yahya-t@just.edu.jo

Received: 25 December 2024; Accepted: 26 February 2025; Published: 16 April 2025

ABSTRACT: Social media has emerged as one of the most transformative developments on the internet, revolutionizing the way people communicate and interact. However, alongside its benefits, social media has also given rise to significant challenges, one of the most pressing being cyberbullying. This issue has become a major concern in modern society, particularly due to its profound negative impacts on the mental health and well-being of its victims. In the Arab world, where social media usage is exceptionally high, cyberbullying has become increasingly prevalent, necessitating urgent attention. Early detection of harmful online behavior is critical to fostering safer digital environments and mitigating the adverse effects of cyberbullying. This underscores the importance of developing advanced tools and systems to identify and address such behavior effectively. This paper investigates the development of a robust cyberbullying detection and classification system tailored for Arabic comments on YouTube. The study explores the effectiveness of various deep learning models, including Bi-LSTM (Bidirectional Long Short-Term Memory), LSTM (Long Short-Term Memory), CNN (Convolutional Neural Networks), and a hybrid CNN-LSTM, in classifying Arabic comments into binary classes (bullying or not) and multiclass categories. A comprehensive dataset of 20,000 Arabic YouTube comments was collected, preprocessed, and labeled to support these tasks. The results revealed that the CNN and hybrid CNN-LSTM models achieved the highest accuracy in binary classification, reaching an impressive 91.9%. For multiclass classification, the LSTM and Bi-LSTM models outperformed others, achieving an accuracy of 89.5%. These findings highlight the effectiveness of deep learning approaches in the mitigation of cyberbullying within Arabic online communities.

KEYWORDS: Arabic text classification; arabic text mining; cyberbullying detection; neural networks; deep learning; CNN; LSTM; YouTube; Bi-LSTM

1 Introduction

Social media is one of the most recent developments on the internet, gaining widespread popularity and emerging as the primary tool for communication and interaction. Moreover, it serves various purposes, such as job seeking or hiring. It has become a platform for searching for jobs, and currently, organizations and companies utilize social media platforms for trading, marketing, and hiring. According to the Arab Social

Media Report, social media has rapidly spread in the Arab world. In some Arab countries, over 90% of the population actively engages on various social platforms [1].

While social media offers numerous advantages and valuable purposes, the presence of individuals with malicious intentions, such as bullying users and cyberbullying, has become a widespread phenomenon. This study focuses on the Arab users. The youth and adolescents in the Arab world suffer from cyberbullying, with high percentages of 33.6% in Lebanon, 44.2% in Jordan, 39.1% in Oman, 20.9 in UAE, and 31.9% in Morocco [2].

Cyberbullying involves the use of various technologies of information, such as mobile phones and messaging platforms to intentionally and repeatedly harm others. Bullying behaviors include scaring, provoking, defaming, or bullying others' color, gender, values, and/or religion. Compared to classical bullying, cyberbullying is more harmful and substantial. This is because cyberbullying uses the internet as a way to spread harm quickly, where the bullies can post a victim's image, video, or personal information to the social network, and it can reach many users in a short time.

Removing harmful content from the internet is challenging, often bordering on impossibility. This results in significant psychological harm to victims, including lack of concentration, loss of self-confidence, and depression [3,4]. This phenomenon has been widespread among young people in recent times due to several reasons, including the absence of parents' role, the deficiency of universities, schools, and media channels to increase awareness of cyberbullying risk and its harm to health, the absence of the concept of respect for others and acceptance of their differences [5].

In terms of detecting or diminishing the impact of this phenomenon using smart services, several challenges are encountered in building cyberbullying detection systems (CDS), such as (1) language challenges, (2) dataset challenges, and (3) labeling challenges. It is popularly known that the Arabic language is one of the challenging mining and analytics languages due to its complexity. For example, “كره بس من باب النصيحه ماتغلبي حالك وتغيري بشكلك انتي شو ماعملتي بتضلي مش حلوه وتقبلي نفسك زي ما انتي انا ما بحكي هيك“. In English, this could be translated as “Do not bother changing your appearance; no matter what you do, you will not be beautiful. Try to accept yourself as you are. I am not saying this because I hate you, but out of advice”. Using the fundamental Arabic keywords-based analysis, the previous sentence expressed mixed sentiments of bullying and non-bullying; however, in reality, it is a form of bullying based on one of the girls' appearance.

So, in many cases, a logical and deep analysis needs to be applied to classify it as a cyberbullying or non-cyberbullying comment [6]. On the other side, some sentences seem positive to express sarcasm, such as “ جد ضحكت من كل قلبي على الرغم من انكم بيض هههههه”. This can be translated as “I laughed with all my heart even though you are white”. However, because bullying is very subjective and subtle, it might be challenging to identify it. In addition, there are also rapid changes in the usage of languages, particularly the language used by teens and young adults, which change quickly and will impact the keywords used as a feature in the detection of cyberbullying [6].

The dataset is another challenge in cyberbullying detection, where dataset extraction is not a simple task since it is related to user and privacy information and social media platforms usually do not openly provide this data. So, the data extraction process requires secured and licensed tools or connections to a secure API (Application Programming Interface) such as Twitter API [7]. Collected data from social media platforms might be inconsistent, incomplete, and contain errors and extra characters such as punctuation marks, special characters, duplication, URLs (Uniform Resource Locator), and hashtags. This introduces the challenge of data preprocessing, which requires several steps to obtain meaningful social media data. Furthermore, data labeling is a labor-intensive process, critical and time-consuming, and where each data row in the corpus could need to be manually labeled [8]. Initially, defining an appropriate set of labels associated with each data sample is essential. An expert from different domains might intervene to help identify the intention behind the post, given its context. The annotation step is critical to designing effective labeling strategies to produce a high-quality labeled dataset.

The Arabic language is the official language of 22 countries in the Arab world. The Arabic language consists of 28 letters [9]. The Arabic Language can be used in different forms classified as 1) Clear standard Arabic, which is the Quran language (The holy book of Muslims) 2) Modern Standard Arabic (MSA), which is the formal language used in schools, universities, and media channels, and 3) dialects are the slang language, where each Arab country has its different dialect [9,10]. Arabic is a morphologically rich language; adding one or more alphabets to a root can generate dozens of words with different meanings [11], for example, لاعب, يلعب, لعب. Moreover, synonyms are also increasing ambiguity in the Arabic language where the same meaning can be expressed in many different words, for example, صبوحه مليحه, فاتنه, جميله, غانيه, حسناء, all these words meaning beautiful. In addition, the Arabic language contains a lot of orthography and diacritics, where all previously mentioned facts increase the complexity of the Arabic language, especially in the domain of the natural language process.

The detection of cyberbullying on social media has recently received significant attention due to its devastating impact on social interactions and mental health. However, despite the growing research in this field, there remains a noticeable gap in addressing the use of content from Arabic social media to tackle the issue of cyberbullying. Most existing works have been limited to binary classification, failing to consider that cyberbullying practices include nuanced and multifaceted dimensions such as racism, sexism, and general abuse. Further, the Arabic language presents different challenges due to morphological richness, semantic intricacies, and a wide presence of dialects that further complicate natural language processing. Besides, high-quality, annotated datasets suitable for Arabic content at a premium constrain the creation of effective detection systems. This gap shows the necessity for a framework that is inclusive of these challenges and does so to provide a multiclass classification model for detecting various cyberbullying forms in Arabic social media.

Consequently, this paper attempts to fill such gaps by presenting utilizing deep learning approaches for cyberbullying detection in Arabic social media content. The deficiencies found in the study are overcome by proposing an end-to-end methodology that applies natural language processing techniques for innovative data annotations and modeling methods.

In summary, our contributions can be summarized below:

- Develop and annotate a novel dataset of 20,000 Arabic YouTube comments specifically curated for cyberbullying detection tasks.
- Leveraging the pre-trained GloVe-Arabic embedding to effectively capture semantic and syntactic relationships in the Arabic language, addressing complexities brought about by its rich morphology and diverse dialects.
- Evaluate the performance of various deep learning models, including CNN, Bi-LSTM, LSTM, and a hybrid CNN-LSTM architecture, for binary and multiclass classification to provide more insight into the depth of cyberbullying behaviors. In contrast, the developed method will result in a scalable and practical approach to detecting cyberbullying on Arabic social media.

The rest of this paper is organized as follows: Section 2 overviews the related work. Section 3 discusses research methodology. Section 4 evaluates the model performance and results. In the last, Section 6 shows the conclusions and future works of this work.

2 Related Works

This section provides a comprehensive review of the literature on cyberbullying detection, focusing on deep learning methods and datasets. Section 2.1 examines recent advancements in deep neural networks for detecting cyberbullying, highlighting key models, their performance, and limitations. Section 2.2 discusses the datasets used in these studies, emphasizing their properties, languages, and challenges, particularly in the context of Arabic and other underrepresented languages. These subsections identify gaps and opportunities for future research in this domain.

2.1 Cyberbullying Detection Methods

Recently, several deep learning models have been proposed to detect cyberbullying on social media platforms. For example, Al-Ajlan et al. [9] proposed a model called CNN-CB based on using a CNN with word embedding to detect cyberbullying. The model was evaluated using a dataset consisting of 39,000 tweets in English text. The CNN-CB model was compared with the performance of the SVM, and the results showed that the CNN-CB outperformed the SVM (Support Vector Machine) with accuracy by 95% and 81.32%, respectively. However, it focuses on several works on English datasets and limits applicability to languages of various structural and morphological complexities, including Arabic. Moreover, the model does not address dialectical variations and rich morphology challenges that are important to establishing robust systems in the Arab world for detecting cyberbullying.

Similarly, Bu et al. [10] proposed a hybrid model combining two deep learning models (CNN-LSTM). The CNN level was used to extract syntactic knowledge of character series and LSTM was used to extract high-level semantic information from word series. The models were evaluated using 8815 comments from Kaggle, and the accuracy obtained from the hybrid model was 70.8%. The hybrid model also significantly outperformed other ML models for cyberbullying detection. The hybrid model presented here identifies a potential approach by which improved performance might be derived from a combination of CNN and LRCN (Long-term recurrent Convolutional Networks); unfortunately, low accuracy was realized along with limiting dataset size.

Moreover, for using the CNN algorithm, Zhang et al. [11] proposed a deep-learning model for pronunciation using CNN. The model was evaluated using 13,000 samples extracted from Formspring (a social networking service) and 1313 from Twitter. The results showed that the model returns inaccurate results due to the unbalanced dataset, reflected in a precision score where the precision is 56%, recall is 78%, and accuracy is 96%. Although the model achieved high accuracy, precision, and recall, it had a significant skew due to the imbalance of the dataset, making it less effective generally in binary classification tasks. This points to the fact that balanced datasets are crucial for models to perform reliably and fairly, specifically in sensitive applications like cyberbullying detection.

Ahmad et al. [12] compared several machines and deep learning models for detecting cyberbullying in Bangla and Romanized writings. The models were evaluated using 12,000 text comments from YouTube using YouTube API. The dataset was manually labeled into two classes: bullying and not bullying. The results show that the CNN model achieved the best accuracy of 84%. While this study has pointed out the flexibility of CNN across different languages, it is limited to binary classification, leaving subtler behaviors such as sexism and racism unaddressed.

Mohaouchane et al. [13] proposed a set of deep-learning models, including CNN and Bi-LSTM, and a hybrid model that combines CNN and LSTM to detect offensive language in social media. The models were evaluated using 15,050 Arabic comments. The data was collected from YouTube and manually labeled into bullying and non-bullying classes. The results showed that the hybrid model (CNN-LSTM) achieved the best recall of 83.64%, while the CNN model achieved the best accuracy, precision, and F1-score with values of 87.84%, 86.10%, and 84.05%, respectively. This study makes a valuable contribution to Arabic cyberbullying detection. However, it is limited to binary classification and does not identify specific types of bullying in Arabic social media.

Agrawal et al. [14] proposed an LSTM model for cyberbullying detection using multiple social media platforms. The model was evaluated using 16,000 tweets with a word embedding layer for feature extraction, and the result showed that the LSTM model performed well, where it achieved precision = 92%, recall = 91%, F1-score = 91%, and AUC (Area Under Curve) = 93%. This study demonstrates the strength of using LSTM models in handling sequential data. However, it focused solely on English tweets, and reliance on binary classification limits its effectiveness in multiclass scenarios.

Recently, Lanasri et al. [15] attempted to address the research gap dealing with Arabic dialects, especially the Algerian dialect, for hate speech detection on social networks. The authors developed a comprehensive study using deep learning architectures to analyze over 13,500 Algerian dialect documents from platforms like YouTube, Facebook, and Twitter. The proposed models have shown promising results in detecting hate speech. Similarly, this paper examines deep learning methodologies to detect cyberbullying on Twitter. Additionally, Seetharaman et al. [16] reviewed the application of deep learning methodologies to cyberbullying detection in Twitter. Their work focused on how to make use of models like CNN, RNN (Recurrent Neural Network), and LSTM for the detection of harmful content in tweets. The study also highlights specific challenges for the Twitter data: limited length and expressiveness, which can be effectively addressed using deep learning models and thus efficiently recognize instances of cyberbullying. The performance of the proposed models on various datasets demonstrates competitive results in detecting cyberbullying content.

Azzeh et al. [17] proposed a detection model that combines CNN with Multi-Head Attention. The system addresses challenges such as the rich morphology of Arabic and the informal language used in social media, preprocesses text through techniques such as tokenization and stemming, and utilizes embedding such as AraBERT or FastText to extract features. This architecture captures local patterns using CNN, whereas Multi-Head Attention focuses on essential contextual elements, enhancing detection accuracy. Compared to baseline models, this approach has demonstrated significant promise in online safety on Arabic platforms.

Table 1: Deep learning models used for cyberbullying detection

Ref.	Size (# of samples)	Year	Source	Features	Model	Performance evaluations
[9]	39,000	2018	Twitter	Word embedding, Text features	CNN-CB, SVM	The best model is CNN-CB. Accuracy = 95%
[10]	8815	2018	Kaggle	Word embedding, Text features	CNN, LRCN	The best model is LRCN Accuracy = 70.8%
[11]	14,313	2016	Twitter, Formspring	Word embedding, Text features	PCNN (Pulse Coupled Neural Network)	Precision = 56%, Recall 78%, Accuracy 96%
[12]	12,000	2021	YouTube	Word embedding, Text features	CNN	Accuracy = 84%

(Continued)

Table 1 (continued)

Ref.	Size (# of samples)	Year	Source	Features	Model	Performance evaluations
[13]	15,050	2019	YouTube	Word embedding, Text features	CNN, BI-LSTM, CNN-LSTM	The best model is CNN Accuracy = 87.84% Precision = 86.10% F1 = 84.05%
[14]	16,000	2018	Twitter	Word embedding, Text features	LSTM	Precision = 92%, Recall = 91%, F1-score = 91%, AUC = 93%
[15]	12,000	2023	YouTube Comments	TF-IDF Word embedding	CNN	Accuracy = 84%
[16]	13,016 tweets	2023	Twitter	Word embedding, Text features	LSTM RNN DEA-RNN SVM, RF (Random Forest), MNB(Multinomial Naive Bayes)	Accuracy = 90%
[17]	4505 textual	2024	Twitter	Word embedding, Text features	ResNet (Residual Network) CNN	Accuracy = 81%
[18]	Formspring = 40,952 Twitter = 7321 MySpace = 381,557	2019	Formspring, Twitter, and MySpace	TF-IDF BoW	CNN	F1-Score = 85%
[19]	Formspring = 12 K Twitter = 16 K Wikipedia = 100 K	2022	Formspring, Twitter, and Wikipedia	TF-IDF, Word embedding	CNN	F1-Score = 91%
[20]	1.6 Million	2021	Kaggle	TF-IDF, Word2vec	CNN-LSTM	Precision = 76%, Recall = 31%, F1-score = 44%

(Continued)

Table 1 (continued)

Ref.	Size (# of samples)	Year	Source	Features	Model	Performance evaluations
[21]	100 K	2023	Kaggle	Word embedding	LSTM	Accuracy = 80.86 Accuracy = 82%
[22]	Instagram = 2188 Twitter = 7321	2020	Instagram Twitter	TF-IDF Word2vec	LSTM	Accuracy = 84% Precision = 85%, Recall = 81%, F1-score = 83%,
[23]	18 K	2020	Twitter	GloVe Wors2vec	Bi-LSTM	precision = 86%, Recall = 86.5%, F1-score = 86.2%,

2.2 Related Cyberbullying Detection Dataset

We have examined the most recently used datasets for cyberbullying detection tasks with a focus on the main properties of each dataset, such as size, source, language, and classes, as summarized in Tables 1 and 2. We noticed that there is a lack of (1) a few of the cyberbullying datasets collected in the Arabic language, (2) most of them are used for binary classification tasks, mainly categorized as (bullying or non-bullying), and (3) most of the datasets are manually labeled. Bozyiğit et al. [24] presented a balanced novel dataset for cyberbullying detection tasks collected from Twitter containing 5000 labeled Turkish contents. The dataset is publicly available in comma-separated format. Alsubait et al. [25] proposed a novel dataset for cyberbullying detection tasks collected from Twitter and YouTube, comprising 15,050 labeled Arabic content. The dataset is unbalanced because it consists of 5000 bullying samples and 10,050 non-bullying ones.

Table 2: Example of datasets used for cyberbullying detection

Ref.	Year	Dataset size	Language	Content type	Classes	Source	ML algorithms
[24]	2021	5000	Turkish	Text	Binary	Twitter	SVM, LR, KNN (K Nearest Neighbor), NBM (Naive Bayes Multinomial) AdaBoost, RF
[25]	2021	15,000	Arabic	Text	Binary	YouTube	MNB, CNB, LR (Logistic Regression)
[26]	2021	8154	Arabic	Text	Binary	Twitter	SVM
[27]	2021	17,748	Arabic	Text	Binary	Twitter	SVM
[28]	2019	100,327	Arabic	Text	Binary	Twitter, Microsoft Flow, YouTube	PMI, Chi-square, Entropy

(Continued)

Table 2 (continued)

Ref.	Year	Dataset size	Language	Content type	Classes	Source	ML algorithms
[29]	2019	9854 tweets	Indonesian	Text	Binary	Twitter	LSTM, Bi-LSTM, CNN
[30]	2020	60,000 comments	English	Text	Binary	Weibo	LR, SVM, CNN, CNN
[31]	2019	–	English	Text and image	Binary	Instagram	CNN
[32]	2018	Fromspring 12 K, Twitter 16 K, Wikipedia 13,590 YouTube 54 K	English	Text	Form spring (is not especially about any single topic). Twitter (racism and sexism). Wikipedia (personalattack) YouTube (Binary)	Form spring Twitter Wikipedia YouTube	CNN, LSTM, Bi-LSTM
[33]	2020	6138	Arabic	Text	Binary	Facebook and Twitter	KNN, SVM, NB, RF, J48
[34]	2021	Instagram (2218) Vine (970)	English	Session (overall post image and comment)	Binary	Instagram Vine	LR, SVM, XGBoost CNN, LSTM HAN, HAND, SEND

Almutairi et al. [26] proposed a novel dataset for cyberbullying detection tasks. The dataset was collected in Arabic from Twitter, especially for the Saudi Arabia region, and was based on bullying keywords. The data collection took one year and seven months and was collected in separate time intervals. Finally, the total dataset is 8154, and dataset samples were manually labeled as 1 for bullying and 0 for non-bullying.

Almutiry et al. [27] proposed a dataset called "AraBully-Tweets" for detection the of cyberbullying. The dataset was collected in Arabic language using ArabiTools and Twitter API via two different methods: query-oriented and random selection. The size of the dataset was 17,748, where 4178, was the number of bullying tweets, while 3570 was the number of non-bullying tweets. The authors create a lexicon called "Ara-Bully-Words" containing all bullying words to perform the annotation process automatically using Python code that compares the content of each tweet with the "Ara-Bully-Words" where if the tweet includes one or more bully words considered as bullying otherwise is non-bullying. The authors applied manual annotation to check out the efficiency and performance of the Python-based Automatic Annotation model.

AlHarbi et al. [28] proposed a novel dataset to generate a lexicon-based cyberbullying dataset. The data set was collected in Arabic from three resources, Microsoft Flow, YouTube, and Twitter, and then all data were compiled into one file with 100,327 tweets and comments. Three annotators manually labeled the data set as 1 for bullying and 0 for non-bullying. After preprocessing, the dataset was transformed into a lexicon using three different approaches: PMI (Pointwise Mutual Information), Chi-square, and Entropy, and the result showed that the PMI outperformed others with 81% recall.

Anindyati et al. [29] proposed a novel dataset for detection tasks. The data was collected in the Indonesian language from Twitter via Twitter API by searching for harsh words commonly used for cyberbullying. The total data set is 9854 tweets manually labeled into two labels, "neutral" and" bully," by a group of annotators. Table 2 shows the summary of the datasets currently available in the literature.

3 Methodology

This section presents the underlying rationale for developing the deep learning-based cyberbullying detection framework, tailored explicitly for robust and privacy-aware cyberbullying detection on Arabic social media. The model framework includes four main components, as shown in Fig. 1: the collection and annotation of the dataset, data preprocessing to address the challenges of the Arabic language, and the last steps are the selection and evaluation of deep learning classifier includes: CNN, LSTM, Bi-LSTM, and hybrid CNN-LSTM. The performance of deep learning models is evaluated based on using several performance metrics such as accuracy, precision, F1-score, and recall to ensure its effectiveness in detecting cyberbullying. Each of these steps will be discussed in detail.

3.1 Dataset Collection and Annotation Stages

A new Arabic dataset creation is necessary to grow cyberbullying detection, especially with multiclass classification. Unlike binary datasets, multiclass datasets provide more depth in Arabic's linguistic and semantic variations. So, a cyberbullying dataset was created based on dataset collection and annotation. The dataset was crawled from YouTube comments for videos that addressed the most prominent events and issues of public opinion that occurred in Jordan between 2019 and 2022 to study the percentage of the cyberbullying phenomenon in Jordan. The collected comments were retrieved from videos related to rising prices, cases of violence against women, honor crimes, and parliament resolutions.

Our system retrieved 20,000 Arabic text comments and no criteria to include or exclude any comments to ensure the fairness of the study results. A scraper tool was used to collect the data, including fields such as Comment_ID, Author, Date, and the Arabic text content. We added three additional columns: Class (1 for bullying and 0 for non-bullying comments) and Type, which specifies the subclass of bullying. The annotations were done by experts who marked these labels based on the context of the content. A sample of the dataset is shown in Table 3.

The annotation process involved revisiting the comments by a group of three native Arabic speakers to determine if they contained any form of cyberbullying, such as general bullying, racism, sexism, or bullying of officials, based on the context as shown in Table 4. A senior researcher validated the annotated data to ensure consistency in labeling, using majority voting to assign a final label to each comment. The resultant dataset consisted of 20,000 comments, each labeled as one of the four classes of cyberbullying. The annotation was done based on the context of each comment, as detailed below.

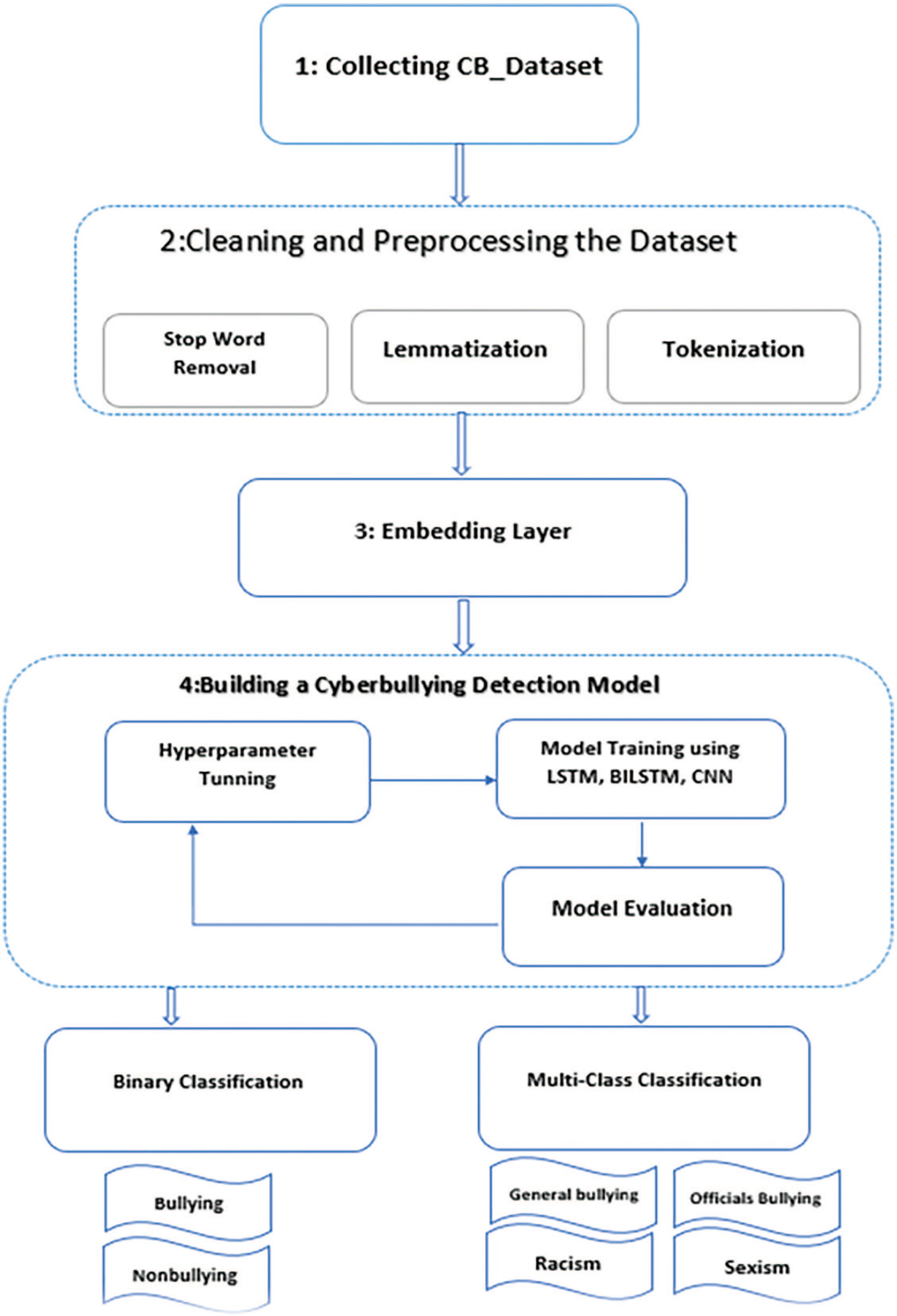

Figure 1: The cyberbullying detection framework

Table 3: Sample of the cyberbullying data

ID	Author	Date	Text (Arabic)	Text (English translation)	Class	Type
UgzZE6n0b-l6FK soUjf4AaABAg	"_"	2021-08-01T235559	اتفه يا ناديا عمتلي بالعالم المقابلة مع وحده	Oh Nadia, you did the interview with the most ridiculous person in the world.	1	1
Ugz-Rq1pCzffIc X0dUh4AaABAg	"_"	2020-07-29T215706	الله يبارك فيك ويزيدك	May God bless you and grant you more.	0	0
UgxYv8ENa-iQxE FxPbp4AaABAg	"_"	2020-07-29T142541	مهما كان فقد سخر أمواله فيواستثمرها الاردن والاردنيونالله بكثرمن امثاله	Regardless, he dedicated his wealth and invested it in Jordan, and Jordanians pray for more people like him.	0	0
UgwIN_y_9hMwS-W2XwV4AaABAg	"_"	2021-02-13T115951	تفاهات تفاهات تفاهات ما في مضمون	Nonsense, nonsense, nonsense—there's no substance.	1	1

Table 4: Classes and subclasses labels and descriptions

Main class	Type (Subclass)	Class name	Interpretation
0	0	No bullying	No bullying.
1	1	General bullying	Any general type of bullying, whether it contains obscene or insulting words, is not mentioned in the below classes.
1	2	Racism	The comment includes bullying of religion, race, color, and form or bullying of a particular tribe or state.
1	3	Sexism	Sexism Any post that bullies gender using any form of bullying or devaluation based on a person's gender. In addition, any form of bullying against women.
1	4	Bullying of officials	It is any bullying or abuse of officials.

3.2 Dataset Preprocessing Stage

Preprocessing steps are one of the most essential processes in data analysis, and acquiring a clean version of the dataset is required. It is a process that involves transforming raw data into an understandable format where actual data written in social media data has usually been inconsistent and incomplete in terms of linguistic errors as well as extraneous noise such as punctuation marks, special characters, duplication, missing values, URLs, and hashtags [35]. Non-Arabic characters and Arabizi [36] were also recorded as noise in the collected dataset. Preprocessing is, therefore, essential to refine the outcomes of analysis and decrease the amount of time and memory to be used for the processing process. Preprocessing was done in several ways using Python's regular expressions and the Natural Language Toolkit [37] to filter and clean the dataset, remove stop words, and normalize the text. This includes:

- Duplication Removal: The duplicate data instances were identified and removed, retaining unique entries.
- Dataset Cleaning: We used regular Python expressions to remove punctuation marks, special characters, URLs, hashtags, and repeating text, symbols, and non-Arabic characters from each data instance.
- Stop Words Removal: We removed frequently occurring words that provide little analytical value to improve the dataset's relevance. We created a custom stop word list based on our own experience and insights from related works or repositories, including common Arabic stop words like ‘في, ’ ‘عن, ’ ‘و, ’’لا, ’ and ‘من. ’
- Tokenization: The text has been divided into smaller fragments, called tokens, which can be analyzed.
- Normalization: The lemmatization process was applied to normalize the variation in word form, keeping them consistent across the dataset.

3.3 Cyberbullying Detection and Classification

We have conducted experiments employing various deep-learning techniques to develop and evaluate a system for efficiently identifying cyberbullying. The pre-trained word embedding, GloVe-Arabic, was used to extract features, which were fed into diverse deep-learning classifiers of instances in binary and multiclass classification. Pre-trained embedding models significantly improved the accuracy of neural network-based natural language processing, as words were represented as vectors in a space showing semantic and syntactic relationships. Out of the pre-trained Arabic embedding, the GloVe-Arabic dictionary was used in this work, with a vocabulary size of 1.5 million words (1.75 billion tokens) and 256 dimensional [38]. The preprocessing text was done by tokenization to convert words into integer sequences, padding to make sequences of the same length, and training sequences with the GloVe-Arabic embedding matrix as input for the deep learning models.

This study involves two different classification tasks to determine cyberbullying content effectively. Binary classification for comments is classified into two classes: cyberbullying and non-bullying. Multiclass classification to classify comments into five distinct classes: general bullying, racism, sexism, bullying of officials, and non-bullying. The classification focuses on the most prevalent forms of bullying in Jordan, with a special focus on bullying of officials due to its high prevalence in the region.

3.3.1 Deep Learning Models

This section explores several deep learning models, including LSTM, bi-directional LSTM, CNN, and the hybrid model, for cyberbullying detection.

- **Long Short-Term Memory Networks (LSTM)**

Long Short-Term Memory Networks model sequential dependencies by maintaining what is coming in as input within the "cell state" information store [39,40]. This LSTM comprises one memory cell and three gates: input gate, output gate, and forget gate- a trio controlling the information of training time. There is one input gate for adding new information, and one forgets the gate to eliminate useless information, as shown in Fig. 2. The input words are translated into embedding vectors using GloVe-Arabic pre-trained embedding in our implementation. These are fed into LSTM units whose output is piped into a sigmoid neuron for binary classification, for example, cyberbullying vs. non-bullying or a SoftMax neuron for multiclass classification.

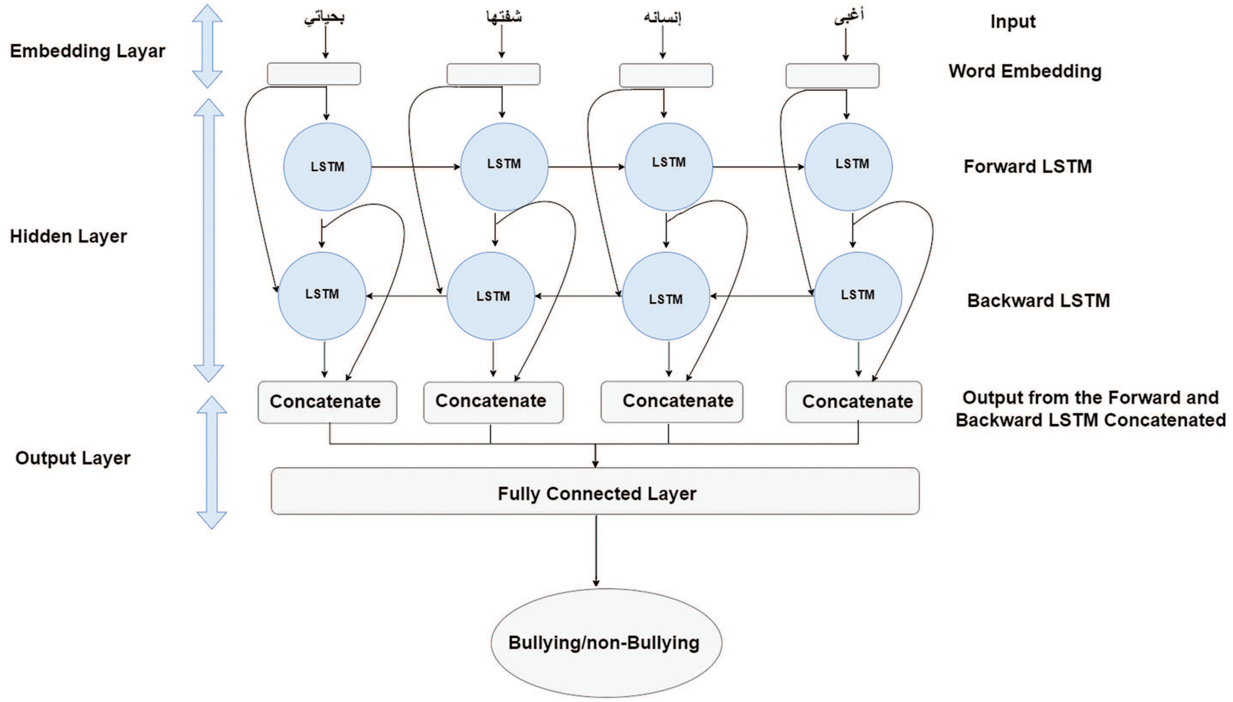

Figure 2: The architecture of the Bidirectional LSTM model

- **Bi-Directional Long Short-Term Memory Networks (Bi-LSTM)**

Bi-LSTM is an extension of the traditional LSTM, where input sequences are scanned forward and backward to capture contextual dependencies from the preceding and following words. This concept finds its application in various aspects of natural language processing. In our implementation, the input sequence of words W (1) to W(n) (n: number of input words) are converted into embedding vectors using GloVe-Arabic pre-trained embeddings. Then, these vectors are fed into two LSTM units-one for front-to-back processing and one for back-to-front processing. The output is concatenated and fed into a sigmoid neuron for binary classification, such as cyberbullying vs. non-cyberbullying, and a SoftMax neuron is used for multiclass classification.

- **Convolutional Neural Networks (CNN)**

The CNNs traditionally used for image classification have lately been adapted for text classification by adding an embedding layer and a one-dimensional convolution layer [41]. CNNs are a deep learning model comprising input, hidden, and output layers. This architecture has achieved outstanding performance in various tasks, such as spam filtering, cyberbullying detection, and sentiment analysis. In this work, the input sequence of words W(1) to W(n) will be embedded using pre-trained GloVe-Arabic embedding. These would undergo a convolution block constituted with a convolution layer to elicit a representation for local features, supported with a MaxPooling to shrink these representations, as shown in Fig. 3. The output is fed into a fully connected hidden layer, where a dropout rate of 0.5 is applied to prevent overfitting.

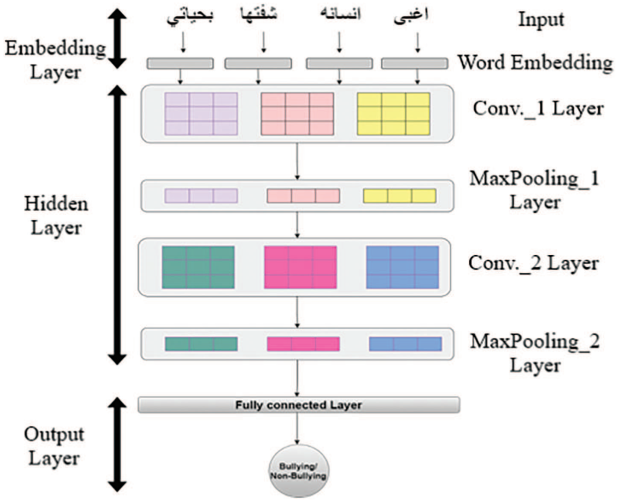

Figure 3: The implemented architecture of the CNN model

- **Hybrid Model (CNN-LSTM)**

The hybrid model we implemented combines (CNN-LSTM); the CNN is for extracting local features while the LSTM learns long-term dependencies. The input sequence of words W(1) to W(n) passes through an embedding layer to change each word into its corresponding embedding vector using GloVe-Arabic pre-trained embedding. This embedding matrix is further passed to the Conv1D layer to capture the local features. A Max-Pooling layer on these features reduces dimensionality and the number of parameters. The result now acts as the input to the LSTM layer, which then captures the long-term dependencies from feature maps, as shown in Fig. 4. The output of LSTM finally goes through a sigmoid or SoftMax neuron that classifies a comment as either cyberbullying/non-bullying or with a specific type.

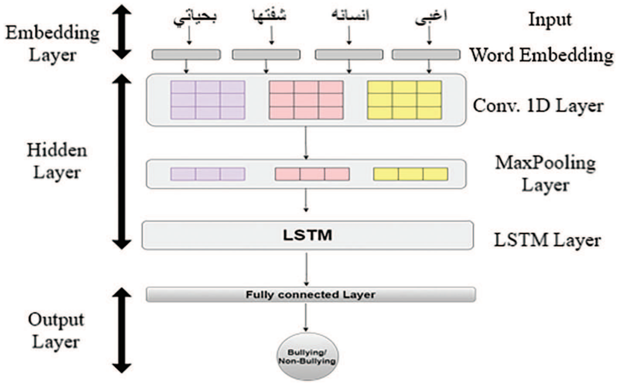

Figure 4: The implemented architecture of the hybrid model

3.3.2 Hyperparameter Tuning

The prefix 'hyper' implies the 'top-level' parameters whose values control and guide the learning and model parameters resulting from it, such as optimizers, activation functions, dropout rates, learning rates, and so it is recommended to be wisely tuned to obtain the optimal performance for a model. Hyperparameter tuning is challenging, and a set of combination optimal parameters for the learning algorithms should be chosen to maximize the models' performance. This step is usually performed manually based on the rule of thumb, copying them from another experiment, or based on trial and error. So, we use the grid search algorithm to obtain the optimal value for the essential hyperparameters such as learning rate, optimizers, and batch size. Tables 5 and 6 show the best hyperparameter values of the deep learning models that achieved the best performance of classification process.

Table 5: Hyperparameters for LSTM and bidirectional LSTM models

Hyperparameter	Value (LSTM)	Value (Bi-LSTM)	Values examined by grid search
Learning rate	0.001	0.001	0.1,0.01,0.001,0.0001,0.00001, 0.2,0.02,0.002,0.0002,0.00002
Number of epochs	15	15	–
Dropout rates	0.3	0.3	–
Batch size	128	128	32,64,128,256
Loss function	Binary cross-entropy, categorical_ cross-entropy	Binary cross-entropy, categorical_ cross-entropy	–
Activation function	Sigmoid, softmax	Sigmoid, softmax	–
Optimizer	Adam	Adam	'Adam', 'Adamax', 'Nadam'

4 Evaluation and Result Analysis

In this paper, the performance of cyberbullying classification models was evaluated across various Arabic dialects using several performance metrics such as accuracy, precision, recall, and F1-score. These metrics are defined as:

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

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

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

$$\text{F1} - \text{score} = 2 \times \frac{(\text{Precision} \times \text{Recall})}{(\text{Precision} + \text{Recall})} \tag{4}$$

TP: True Positive, FP: False Positive, TN: True Negative, FN: False Negative

Table 6: Hyperparameters for CNN and hybrid models

Hyperparameter	Value (CNN)	Value (hybrid model)	Values examined by grid search
Number of filters	64	64	
Kernel size	3	3	
Learning rate	0.001	0.001	0.1,0.01,0.001,0.0001,0.00001, 0.2,0.02,0.002,0.0002,0.00002
Number of epochs	20	15	–
Dropout rates	0.3	0.4	–
Batch size	128	128	32,64,128,256
Hyperparameter	Value (CNN)	Value (hybrid model)	Values Examined by Grid Search
Loss function	Binary cross-entropy, categorical_ cross-entropy	Binary cross-entropy, categorical_ cross-entropy	–
Activation function	Sigmoid, softmax	Sigmoid, softmax	–
Optimizer	Adam	Adam	'Adam', 'Adamax', 'Nadam'

In the binary classification results, the CNN and hybrid model (CNN-LSTM) achieved the highest performance with an accuracy of 91.9%, as shown in Table 7. The comparisons of our deep learning models' performance with other studies in Table 1 make it clear that our models show strong performance for binary and multiclass classification tasks. Our CNN, CNN-LSTM, LSTM, and Bidirectional LSTM models perform well in binary classification tasks, with accuracies ranging from 91.1% to 91.9%. The CNN and CNN-LSTM models achieve the highest accuracy at 91.9%. These results are competitive with other models developed in the studies in Table 1. In addition, the F1-scores for our models range from 0.910 to 0.919, comparable to or better than models like CNN (F1 = 0.840) and LRCN (F1 = 0.96).

Table 7: Results of deep learning models for the binary classification task

Deep learning models	Accuracy	Precision	Recall	F1-score
CNN	0.919	0.917	0.921	0.919
CNN-LSTM	0.919	0.917	0.921	0.919
LSTM	0.915	0.935	0.893	0.914
Bidirectional LSTM	0.911	0.933	0.888	0.910

In the multiclass classification, the LSTM and Bi-LSTM models showed the best performance with an accuracy of 89.5% using the cyberbullying dataset, with the hyperparameter for the LSTM and Bi-LSTM models as shown in Table 8. In the multiclass classification task, our LSTM and Bidirectional LSTM models achieve an accuracy of 89.5%, which is quite competitive with others in Table 1. While other works reported varying accuracies, 80.86%, 84%, and 90%, our models consistently demonstrated high precision and recall, with F1-scores ranging from 0.836 to 0.896 across the multiclass models. This indicates that our models perform well in terms of balance and reliability.

Table 8: Results of deep learning models for multi-class classification task

Deep learning models	Accuracy	Precision	Recall	F1-score
LSTM	0.895	0.896	0.896	0.896
Bidirectional LSTM	0.895	0.898	0.895	0.896
CNN	0.881	0.883	0.882	0.882
CNN-LSTM	0.835	0.837	0.837	0.836

In artificial intelligence, a key consideration when training deep learning models is determining the optimal point to stop the training process. The challenge lies in finding the right balance—stopping when the model hasn't learned enough (underfitting) or when it has learned too much (overfitting). To address this, we employ Early Stopping, a useful technique integrated into Keras for training deep learning models. Early Stopping intervenes by halting the training process as soon as a decline in performance is detected. Figs. 5–8 show the training and validation loss values (development) for each model. These figures demonstrate that all models underwent training without encountering underfitting or overfitting issues.

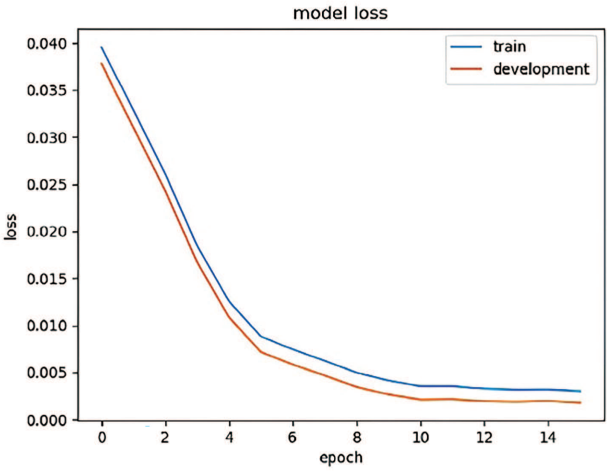

Figure 5: LSTM model loss of training and validation data

Musleh et al. [42] highlight the significance of detecting cyberbullying; it is essential to recognize its impact on individuals and society. The findings in [42] underscore the harmful effects of cyberbullying on individual well-being and the broader social environment, making the development of effective detection models even more crucial. Given these severe impacts, the models developed in this study play a significant role in addressing this issue. We can detect cyberbullying with high accuracy and precision using advanced approaches like CNN, LSTM, a hybrid model of CNN and LSTM, and Bidirectional LSTM. This allows for early identification and intervention, which, in turn, mitigates the harmful effects of cyberbullying and supports a safer online environment for all users.

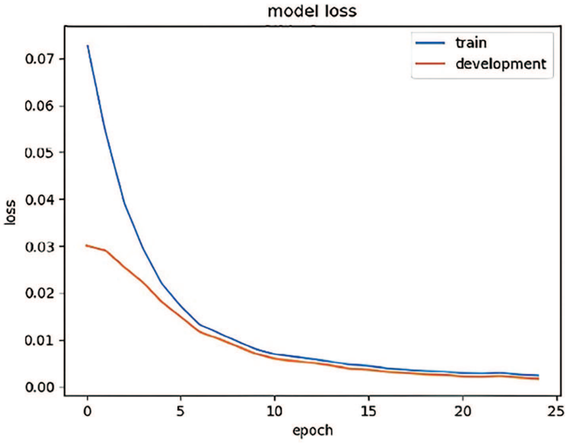

Figure 6: Bidirectional LSTM model loss of training and validation data

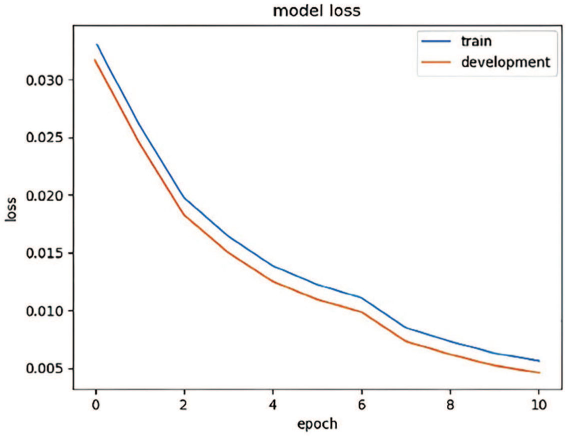

Figure 7: CNN model loss of training and validation data

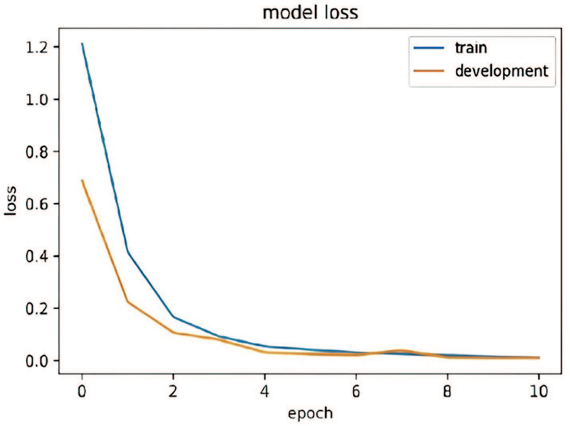

Figure 8: Hybrid model loss of training and validation data

5 Discussion and Limitations

While our approach is highly competitive in handling Arabic dialects, it has several limitations. One major limitation is the limited coverage of dialects. Although we focus on major Arabic dialects, some dialects are underrepresented in the training data. This can lead to degraded performance when the model encounters these underrepresented dialects. While our approach is effective for dialectal variations, it may struggle with subtle or indirect forms of cyberbullying, where context plays a critical role. To address this limitation, we plan to expand our dataset to include a broader range of Arabic dialects in future work. Additionally, we will explore more advanced models, such as transformer-based architectures, which are better suited to capturing sophisticated contextual relationships often present in subtle forms of cyberbullying. Another important consideration is scalability. While our approach performs efficiently on smaller datasets, it may face challenges when applied to real-world, large-scale applications. Therefore, further research will focus on optimizing the model for faster processing and improved scalability, ensuring its effectiveness in real-world scenarios.

6 Conclusion and Future Work

The extensive use of social media has led to an increase in bullying and other undesirable behaviors, resulting in the widespread issue known as cyberbullying. In this study, we address this issue by developing models for detecting cyberbullying on social media platforms using deep learning. These models were trained on a cyberbullying dataset consisting of 20,000 comments collected from YouTube's most popular videos in Jordan between 2019 and 2022. The evaluation results demonstrate the quality of the new dataset and the performance of the proposed models in detecting cyberbullying. The highest accuracy of 91.9% was achieved for binary classification using the CNN and hybrid models, while the highest accuracy of 89.5% for multiclass classification was achieved using the LSTM and Bi-LSTM models. All deep learning experiments were conducted using cross-validation, with the data split into 80% for training and 20% for testing.

One of the interesting findings reveals that the dominant bullying content in Jordan is related to sexist and racist themes. In the future, to enhance the availability of Arabic resources on cyberbullying, we plan to expand the Arabic Cyberbullying dataset and support additional NLP (Natural Language Processing) tasks. We also aim to evaluate additional annotation strategies and explore other transformer models like BERT (Bidirectional Encoder Representations from Transformers) and XLM (Extensible Markup Language). Furthermore, we will investigate and analyze more data on bullies, including their age, gender, and nationality.

Acknowledgement: The authors thank all the editors and anonymous reviewers for their valuable comments and suggestions.

Funding Statement: This study is financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, Project No. BG-RRP-2.013-0001-C01.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Yahya Tashtoush; draft manuscript preparation: Majdi Maabreh, Shorouq Al-Eidi, Areen Banysalim, Ola Karajeh; funding acquisition and supervision: Plamen Zahariev; review: Yahya Tashtoush, Shorouq Al-Eidi, Majdi Maabreh, Plamen Zahariev. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Salem F. Arab social media report 2017: social media and the internet of things: towards data-driven policymaking in the Arab world: potential, limits and concerns; 2017. [Internet]. [cited 2025 Feb 25]. Available from: https://mbrsg.ae/en/research/innovation-and-future-governments/arab-social-media-report-2017?id=70493&rel=4228.
2. AL Nuaimi A. Effectiveness of cyberbullying prevention strategies in the UAE. In: ICT Analysis and Applications: Proceedings of ICT4SD 2020; 2021; Singapore: Springer. p. 731–9.
3. Salawu S, He Y, Lumsden J. Approaches to automated detection of cyberbullying: a survey. IEEE Trans Affective Comput. 2020;11(1):3–24. doi:10.1109/TAFFC.2017.2761757.
4. Raisi E, Huang B. Cyberbullying identification using participant-vocabulary consistency. arXiv:1606.08084, 2016.
5. Cheng L, Silva YN, Hall D, Liu H. Session-based cyberbullying detection: problems and challenges. IEEE Internet Comput. 2021;25(2):66–72. doi:10.1109/MIC.2020.3032930.
6. Haidar B, Chamoun M, Serhrouchni A. A multilingual system for cyberbullying detection: Arabic content detection using machine learning. Adv Sci Technol Eng Syst J. 2017;2(6):275–84. doi:10.25046/aj020634.
7. Habash NY. Introduction to Arabic natural language processing. Synth Lect Hum Lang Technol. 2010;3(1):1–185. doi:10.1007/978-3-031-02139-8.
8. Tashtoush Y, Magableh A, Darwish O, Smadi L, Alomari O, ALghazoo A. Detecting Arabic YouTube Spam using data mining techniques. In: 2022 10th International Symposium on Digital Forensics and Security (ISDFS); 2022 Jun 6–7; Istanbul, Turkey: IEEE; 2022. p. 1–5. doi:10.1109/ISDFS55398.2022.9800840.
9. Al-Ajlan MA, Ykhlef M. Deep learning algorithm for cyberbullying detection. Int J Adv Comp Sci Appl . 2018;9(9). doi:10.14569/IJACSA.2018.090927.
10. Bu SJ, Cho SB. A hybrid deep learning system of CNN and LRCN to detect cyberbullying from SNS comments. In: Hybrid artificial intelligent systems. Cham: Springer International Publishing; 2018. p. 561–72. doi: 10.1007/978-3-319-92639-1_47.
11. Zhang X, Tong J, Vishwamitra N, Whittaker E, Mazer JP, Kowalski R, et al. Cyberbullying detection with a pronunciation based convolutional neural network. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA); 2016 Dec 18–20; Anaheim, CA, USA: IEEE; 2016. p. 740–5. doi:10.1109/ICMLA.2016.0132.
12. Ahmed MT, Rahman M, Nur S, Islam A, Das D. Deployment of machine learning and deep learning algorithms in detecting cyberbullying in bangla and romanized bangla text: a comparative study. In: 2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT); 2021 Feb 19–20; Bhilai, India: IEEE; 2021. p. 1–10. doi:10.1109/ICAECT49130.2021.9392608.
13. Mohaouchane H, Mourhir A, Nikolov NS. Detecting offensive language on Arabic social media using deep learning. In: 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS); 2019 Oct 22–25; Granada, Spain: IEEE; 2019. p.466–71. doi:10.1109/snams.2019.8931839.
14. Agrawal S, Awekar A. Deep learning for detecting cyberbullying across multiple social media platforms; 2018. p. 141–53.
15. Lanasri D, Olano J, Klioui S, Lee SL, Sekkai L. Hate speech detection in Algerian dialect using deep learning. arXiv:2309.11611. 2023.
16. Seetharaman AR, Jahankhani H. Cyberbullying detection in twitter using deep learning model techniques. In: International Conference on Global Security, Safety, and Sustainability 2023; 2023; Cham, Switzerland: Springer Nature. p. 147–67.
17. Azzeh M, Alhijawi B, Tabbaza A, Alabboshi O, Hamdan N, Jaser D. Arabic cyberbullying detection system using convolutional neural network and multi-head attention. Int J Speech Technol. 2024;27(3):521–37. doi:10.1007/s10772-024-10118-4.
18. Zhang A, Li B, Wan S, Wang K. Cyberbullying detection with birn and attention mechanism. Vol. 294. In: Zhai X, Chen B, Zhu K, editors. Machine learning and intelligent communications. MLICOM 2019. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering. Cham: Springer; 2019. doi:10.1007/978-3-030-32388-2_52.
19. Paul S, Saha S. CyberBERT: bert for cyberbullying identification. Multimed Syst. 2022;28(6):1897–904. doi:10.1007/s00530-020-00710-4.

20. Gada M, Damania K, Sankhe S. Cyberbullying detection using LSTM-CNN architecture and its applications. In: 2021 International Conference on Computer Communication and Informatics (ICCCI); 2021 Jan 27–29; Coimbatore, India: IEEE; 2021. p. 1–6. doi:10.1109/ICCCI50826.2021.9402412.
21. Iwendi C, Srivastava G, Khan S, Maddikunta PKR. Cyberbullying detection solutions based on deep learning architectures. Multimed Syst. 2023;29(3):1839–52. doi:10.1007/s00530-020-00701-5.
22. Rezvani N, Beheshti A, Tabebordbar A. Linking textual and contextual features for intelligent cyberbullying detection in social media. In: Proceedings of the 18th International Conference on Advances in Mobile Computing & Multimedia; 2020; Chiang Mai Thailand: ACM. p. 3–10. doi:10.1145/3428690.3429171.
23. Zhao Z, Gao M, Luo F, Zhang Y, Xiong Q. LSHWE: improving similarity-based word embedding with locality sensitive hashing for cyberbullying detection. In: 2020 International Joint Conference on Neural Networks (IJCNN); 2020 July 19–24; Glasgow, UK: IEEE; 2020. p. 1–8. doi:10.1109/ijcnn48605.2020.9207640.
24. Bozyiğit A, Utku S, Nasibov E. Cyberbullying detection: utilizing social media features. Expert Syst Appl. 2021;179:115001. doi:10.1016/j.eswa.2021.115001.
25. Alsubait T, Alfageh D. Comparison of machine learning techniques for cyberbullying detection on youtube Arabic comments. Int J Comput Sci Netw Secur. 2021;21(1):1–5.
26. Almutairi AR, Al-Hagery MA. Cyberbullying detection by sentiment analysis of tweets' contents written in Arabic in Saudi Arabia society. Int J Comput Sci Netw Secur. 2021;21(3):112–9.
27. Almutiry S, Abdel Fattah M. Arabic cyberbullying detection using Arabic sentiment analysis. Egypt J Lang Eng. 2021;8(1):39–50. doi:10.21608/ejle.2021.50240.1017.
28. AlHarbi BY, AlHarbi MS, AlZahrani NJ, Alsheail MM, Alshobaili JF, Ibrahim DM. Automatic cyber bullying detection in Arabic social media. Int J Eng Res Technol. 2019;12(12):2330–5.
29. Anindyati L, Purwarianti A, Nursanti A. Optimizing deep learning for detection cyberbullying text in Indonesian language. In: 2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA); 2019 Sep 20–21; Yogyakarta, Indonesia: IEEE; 2019. p. 1–5. doi:10.1109/icaicta.2019.8904108.
30. Wu J, Wen M, Lu R, Li B, Li J. Toward efficient and effective bullying detection in online social network. Peer Peer Netw Appl. 2020;13(5):1567–76. doi:10.1007/s12083-019-00832-1.
31. Drishya SV, Saranya S, Sheeba JI, Devaneyan S. Cyberbully image and text detection using convolutional neural networks. CiiT Int J Fuzzy Syst. 2019;11(2):25–30.
32. Dadvar M, Eckert K. Cyberbullying detection in social networks using deep learning based models; a reproducibility study. arXiv:1812.08046. 2018.
33. Kanan T, Aldaaja A, Hawashin B. Cyber-bullying and cyber-harassment detection using supervised machine learning techniques in Arabic social media contents. J Internet Technol. 2020;21(5):1409–21.
34. Ge S, Cheng L, Liu H. Improving cyberbullying detection with user interaction. In: Proceedings of the Web Conference 2021 (WWW '21); 2021; Association for Computing Machinery, New York, NY, USA. p. 496–506. doi:10.1145/3442381.3449828.
35. Haddi E, Liu X, Shi Y. The role of text pre-processing in sentiment analysis. Procedia Comput Sci. 2013;17:26–32. doi:10.1016/j.procs.2013.05.005.
36. Talafha B, Abuammar A, Al-Ayyoub M. Atar: attention-based LSTM for Arabizi transliteration. Int J Electr Comput Eng. 2021;11(3):2327–34. doi:10.11591/ijece.v11i3.
37. Hardeniya N, Perkins J, Chopra D, Joshi N, Mathur I. Natural language processing: python and NLTK. Mumbai, India: Packt Publishing Ltd.; 2016.
38. Fawzy M, Fakhr MW, Rizka MA. Word embeddings and neural network architectures for Arabic sentiment analysis. In: 2020 16th International Computer Engineering Conference (ICENCO); 2020 Dec 29–30; Cairo, Egypt: IEEE; 2020. p. 92–96. doi:10.1109/icenco49778.2020.9357377.
39. Zhang J, Li Y, Tian J, Li T. LSTM-CNN hybrid model for text classification. In: 2018 IEEE 3rd Advanced Information Technology, Electronic and Automation Control Conference (IAEAC); 2018 Oct 12–14; Chongqing: IEEE; 2018. p. 1675–80. doi:10.1109/IAEAC.2018.8577620.
40. Borna K, Ghanbari R. Hierarchical LSTM network for text classification. SN Appl Sci. 2019;1(9):1124. doi:10.1007/s42452-019-1165-1.

41. Yu S, Liu D, Zhu W, Zhang Y, Zhao S. Attention-based LSTM, GRU and CNN for short text classification. J Intell Fuzzy Syst. 2020;39(1):333–40. doi:10.3233/JIFS-191171.

42. Musleh D, Rahman A, Alkherallah MA, Al-Bohassan MK, Alawami MM, Ali Alsebaa H, et al. A machine learning approach to cyberbullying detection in Arabic tweets. Comput Mater Contin. 2024;80(1):1033–54. doi:10.32604/cmc.2024.048003.