Engineered Science
DOI: https://dx.doi.org/10.30919/es1272

Enhancing Cyberbullying Detection in Arabic Text Through Ensemble Stacking Models

Reem Albayari,[1, 2] Arwa A. Al Shamsi[3] and Muath Alrammal[4,*]

Abstract

Cyberbullying has become a major social concern in this modern era of digital communications. Cyberbullying can have detrimental effects on the individuals involved ranging from psychological to pathological. Hence, detecting any act of cyberbullying in an automated manner will help to prevent any unfortunate results. In this regard, data-driven approaches, such as Machine Learning (ML), particularly Deep Learning (DL), have shown promising results. DL approaches provide highly accurate predictive models for text classification. However, literature shows that ML approaches, particularly DL, have not been extensively studied for Arabic text classification of cyberbullying. The prevalence of non-diacritical writing, dialectal variability, and morphological complexity presents challenges in developing high-accuracy text classification systems for Arabic text. Subsequently, the application of DL to cyberbullying detection problems within Arabic text classification can be considered a novel approach due to the complexity of the problem and the tedious process involved, besides the scarcity of relevant research studies. Thus, this research aims to develop a highly advanced DL model that can automatically detect cyberbullying. We evaluated seven deep learning (DL) models for Arabic cyberbullying classification: Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional Gated Recurrent Unit (Bi-GRU), hybrid Bi-LSTM-LSTM, CNN-Bi-GRU, CNN-Bi-LSTM, and Bi-LSTM-Bi-GRU. Subsequently, an ensemble stacking model was implemented, integrating the top-performing DL models in terms of accuracy. The stacking models were designed to optimize predictive accuracy by synergistically combining the strengths of the individual models. The ensemble stacking model consisted of DL models with a meta-learner layer of classifiers. In this research, the first model combined the two best-performing DL models: Bi-LSTM and Bi-LSTM-Bi-GRU. The second model combined the four best-performing models: CNN, Bi-LSTM, Bi-GRU, and Bi-LSTM-Bi-GRU. The final model combined all seven trained DL models. Our results indicate that the stacking DL model with the meta-learner layer of the Random Forest (RF) classifier achieved the highest accuracy of 94.73%, outperforming other models.

Keywords: Deep learning; Ensemble stacking; Arabic text; Machine learning; Cyberbullying.
Received: 08 April 2024; Revised: 13 September 2024; Accepted: 08 October 2024.
Article type: Research article.

1. Introduction

Unstructured Arabic text used in blogs, news' sites, social networking sites, and online forums has increased at a never-before-seen rate due to the growing use of smart mobile devices and digital platforms.[1-3] This increase in data has promoted the issue of cyberbullying. This widespread problem has a serious negative influence on people's mental health and presents a major challenge in the digital age.[3-5]

Manual classification of these bullying words requires a huge time-consuming effort compared to the automated method,[6,7] Cyberbullying autodetection is a critical area of research that has the potential to make online spaces safer and more inclusive for everyone.[8,9] Recently, data-driven approaches, particularly Machine Learning (ML) and Deep Learning (DL) models, have shown promising results in text classification, however, their application to Arabic text classification remains underexplored.[9-11] However, many studies have shown that DL techniques outperform standard ML techniques in text classification,[12-14] and others. Despite

[1] *Business Analytics Program, Abu Dhabi School of Management, Abu Dhabi 6844, United Arab Emirates.*
[2] *Department of Software Engineering and Computer Science, Al Ain University, Abu Dhabi 64141, United Arab Emirates.*
[3] *Faculty of Engineering and IT, The British University in Dubai, Dubai 345015, United Arab Emirates.*
[4] *Faculty of Computer Information System, Higher Colleges of Technology, Abu Dhabi 41012, United Arab Emirates.*
**Email:* malrammal@hct.ac.ae; muath.alrammal@gmail.com (M. Alrammal)

the prevalence of Arabic content across social media, e-commerce, and mobile apps, DL algorithms have not yet been as extensively studied as they have been in English text.[8,15-17]

Arabic is one of the world's six most widely spoken languages. Nearly 300 million people speak Arabic as their first language.[9,18,19] The Arabic language contains linguistic features that make it one of the most difficult languages to speak compared to other languages.[9,19,20] Furthermore, Arabic words do not have sequential forms because their word structures vary depending on where they appear in the phrase and what they imply.[19,21,22] The Arabic language encompasses three main types: Classical Arabic, which is the language of the Holy Qur'an; Modern Standard Arabic; and Dialectal Arabic, which varies significantly.[2-23] This is unlike English, which has only one standardized form. However, most Natural Language Processing (NLP) tools and applications are designed for the English language and Modern Standard Arabic form of the Arabic language,[23] thus, the need for an automatic approach to detect and classify offensive Arabic text is critical. We limited our study to the classification of Arabic text.

This research study aims to fill this highlighted gap by conducting empirical analysis of the performance of seven DL models (Convolutional Neural Network (CNN), Bidirectional Long Short-Term Memory (Bi-LSTM), Bidirectional Gated Recurrent Unit (Bi-GRU), hybrid Bi-LSTM-LSTM, CNN-Bi-GRU, CNN-Bi-LSTM, Bi-LSTM-Bi-GRU) for Arabic cyberbullying and developing an innovative ensemble stacking approach using multiple DL models.

The ML literature has recently shown an increasing interest in ensemble learning techniques. However, their application to cyberbullying autodetection is still limited, particularly given that most current ensemble learning research focuses on homogenous ensembles.[24-28] Ensemble learning, the method of combining several DL or ML models into a single prediction model, is regarded as one of the best ways to improve performance.[24-25] Heterogeneous ensembles combine various model types, whereas homogeneous ensembles use many instances of the same model. Ensemble learning is preferable over the use of a single model. For two main reasons: (a) Ensemble tactics consistently result in higher performance when compared to baseline single learner processes;[24,26,28] (b) The ensemble lowers the spread or dispersion of the predictions which increases the robustness, stability, and dependability of a model's average performance.[24,26] Hence, this study advances the field of the research community by presenting an ensemble model of DL models that include a meta learner layer of classifiers to improve the accuracy of Arabic dialect cyberbullying autodetection. The application of DL to cyberbullying detection problems within Arabic text classification can be considered a novel approach due to the complexity of the problem and the tedious process involved, besides the scarcity of relevant research studies.

The issue being considered is a classification problem in the data science field.[19,29,30] The input is the labelled dataset, benchmark dataset, which was collected from Instagram by Albayari and Abdallah[7] and evaluated for detecting cyberbullying.[5] This labelled dataset was used to develop a predictive model from the input dataset using classification algorithms such as DNN.[31] The predictive model is then used to detect new text data automatically. The logical representation of multi-class classification can be formulated as follows: Let x be the input feature vector (n-dimensional array of numerical features representing the set of keywords corresponding to Arabic text), and y be the corresponding class label (Positive/Neutral/Bullying). The DNN-based learning algorithm consists of multiple hidden layers, each with a set of artificial neurons that perform a non-linear transformation of the input features. The output of the final hidden layer is passed through a final meta meta-learner layer (classical ML classifier) to produce the predicted class label. The specific representation can vary depending on the architecture of the DNN or the learning algorithm and the problem. The model's performance will then be evaluated on test datasets using a variety of performance metrics. More information about the evaluated models and their implementation is provided in Section 3.

The study is structured as follows: Section 2 presents an overview of related work; Section 3 describes the methodology of this research study; Section 4 analyses and discusses the results, and the last section concludes the findings and discusses future work.

Social media platforms such as Facebook, Twitter, and YouTube are rapidly expanding globally. The ability for anyone to express and share their opinions on a wide range of topics was made possible by these social networks, and as a result, billions of comments and reviews are posted online every day, making it necessary to identify bullying automatically. Filtering and categorizing this generated data by hand would be a challenging task.[6,7] The development of cyberbullying autodetection systems is a result of the increased interest in automating the text classification process. Recently, data-driven approaches, particularly DL models, have shown promising results in text classification although their application to Arabic text classification remains underexplored.[9-11]

The DL approach attempts to emulate the functioning of the human brain through algorithms, comprehensive network architecture, and many neurons and layers so that data can be elaborated and defined with abstractions.[32,24] The architecture is composed of layers hidden between the input and output layers performing highly complex calculations for obtaining features from raw data. The goal is to develop a model that describes high-level abstractions in a dataset, implying learning layers of representations.[24,32]

Researchers have found that DL models such as the CNN, LSTM, and Bi-LSTM are effective at analyzing Arabic content.[33,34,35] In 2020, Husain *et al*.,[36] investigated multiple

DL models: LSTM, Bi-LSTM, RNN, GRU, and Bi-GRU on an unbalanced dataset with instances of offensive text: 1,900, and not offensive: 8,100. This imbalance in the dataset could result in biased results. Albadi (2018)[37] developed various classification models using lexicon-based, n-gram-based, and DL approaches for identifying religious hate speech on Arabic social platforms. Currently, ensemble learning is acknowledged as one of the superior methods to improve performance compared to individual DL models. It entails the fusion of several ML or DL models into one unified predictive mode.[38] Despite gaining substantial attention in ML literature, ensemble learning is not extensively applied in cyberbullying autodetection particularly in Arabic text.[39-41]

Stacking aims to integrate various ML models (base learners) through an additional data mining technique.[42] Following the training of base learners, a combiner, also referred to as a meta-classifier, is trained to generate a final prediction based on those made by the base learners.[42] Research has shown that these stacked ensembles serve as an asymptotically optimal system for learning, often surpassing the performance of individual base learners.[42]

Ensemble methods offer diverse approaches to improving model performance by working with multiple models.[42,38] For instance, Bagging, or Bootstrap Aggregating, implements this task by repeatedly training the same algorithm on different subsets of the training data and averaging their predictions to minimize variance. Sequential enhancement trains several models where each one learns from the errors of the past model by increasing the strength of misclassified data points leading to a competent, changeable model.[42]

2. Methodology

This section provides a detailed overview of the research methodology. It begins with an outline of the dataset selection criteria and its characteristics. This is followed by an in-depth examination of the data preprocessing techniques employed to prepare the dataset for analysis. Next, the section introduces the DL models that were selected based on their potential applicability to the research question. The final part of this section is devoted to elucidating the three ensemble models, detailing their integration and function within the study's analytical framework. The research methodology for this study takes the sentiment analysis techniques from the literature already in existence and creatively applies them to the little-studied field of cyberbullying detection in Arabic dialects' text.

2.1 Dataset

To assess the performance of the DL models in this study, we employed a dataset previously established.[7]

2.1.1 Dataset description

This dataset comprises 46,898 Instagram comments, making it one of the most comprehensive public resources for detecting cyberbullying. Notably, it encompasses a diverse range of languages, enhancing its inclusivity across various linguistic variations. The dataset, titled 'AA-MCU' (authors' initials-multi class unbalanced dataset) includes 17,376 positive entries, 18,193 neutral entries, and 11,329 unbalanced entries, all sourced from [7].

2.1.2 Dataset annotation

The dataset was annotated manually. Three annotators (one from Jordan, one from Egypt, and one from Iraq) manually annotated the dataset. The annotators' ages ranged from 23 to 27 and each had a bachelor's degree. As far as we are aware, this dataset is the first to use multi-labeling for the detection of cyberbullying in Arabic text. The comments were categorized as positive, negative, or neutral. The negative comments were further classified into two categories (toxic and bullying) based on their negativity. If the dialect of the comment was not immediately apparent, the annotators wrote not available (NA); if the comments only contained emojis, the dialect was also classified as NA.[43]

2.1.3 Data preprocessing

The pre-processing phase is important for reducing noise and improving text classification performance. Text pre-processing is a crucial stage in developing any word embedding model as it can have a significant impact on the results.[44] In this research, we have used the NLTK library in Python to implement the following pre-processing steps:

• Tokenization: The first stage of pre-processing in which the text is divided into words (tokens) and punctuated or divided by white space.[45]

• Filtering Non-Arabic Content: this step involved removing non-Arabic content from the gathered dataset. This is crucial particularly when working with web data such as the Instagram data we used for our experiment.

• Data Cleaning: The experiment involves the removal of various unknown characters including punctuation, numbers, symbols, and the tatweel (kashida) character; hashtags and URLs have also been eliminated. It is important to note that the hashtag text will stay in the text and will not be removed because most of the time it expresses an idea that should be considered.

• Stop Words Removal: According to [46], stop words are used to structure language, but do not contribute to its content in any way.

• Normalization: Aligning Arabic characters (آ،إ،أ) with (ا), (ة) with (ه), (ى،ي) with (ي), and (ؤ) with (و). Also, words that are elongated are reverted to their original form.

2.1.4 Splitting data

The dataset was split, allocating 80% for training, 10% for testing, and another 10% for validation. The performance of

each model on the test set was documented.

2.1.5 Ethical considerations

The collected comments in the dataset are in the form of brief sentences that do not comply with copyright requirements. To ensure adherence to privacy regulations, we concealed the identities of the owners of the accounts from which the texts were obtained. The collected comments are from public accounts and this action is permitted and legal under the websites' terms of use policies.

2.2 Deep learning model

In our research, we adopted the model architecture previously developed by Al Shamsi and Abdallah for Arabic dialect sentiment analysis.[47] Our study involved evaluating seven distinct DL models designed for auto-detecting cyberbullying in Arabic text. The seven DL models developed include deep CNN, Bidirectional-LSTM, Bidirectional-GRU, hybrid Bidirectional-LSTM-LSTM, CNN-Bidirectional-GRU, CNN-Bidirectional-LSTM, Bidirectional-LSTM-Bidirectional-GRU, and ensemble stacked models. We provide below a detailed description of these DL models within our architecture.

CNN Model: The models employed in our research were configured with specific parameter values to optimize their performance. The Deep CNN model, for instance, incorporated a sequence of filter numbers, namely 16, 32, 64, and 128, coupled with a kernel size of 3 and a pool size of 2. Its dense layer consisted of 600 units, utilizing the Adam optimizer, a batch size of 512, and 100 epochs with a learning rate set at 0.001. These parameter values were meticulously selected to enhance the model's efficacy and overall performance in the sentiment analysis tasks conducted during our research.

Bi-LSTM Model: Incorporated a dropout rate of 0.25 to prevent overfitting, followed by a dense layer comprising 60 units. For optimization, we utilized the Adam optimizer with a batch size of 512 and trained the model for 100 epochs, maintaining a learning rate of 0.001. These parameter settings were selected to enhance the model's performance and prevent potential overfitting during the training process.

Bi-GRU Model: We incorporated a dropout rate of 0.25 to mitigate overfitting concerns, followed by a dense layer with decreasing units (128, 64, 32). We employed the Adam optimizer with a batch size of 512 and conducted training for 100 epochs, maintaining a learning rate of 0.001. These parameter specifications were deliberately chosen to optimize the model's performance and minimize potential overfitting during the training phase.

Bi-LSTM Model: We used a dropout rate of 0.5 for regularization, paired with a densely connected layer featuring decreasing units (128, 64, 32). Employing the Adam optimizer with a batch size of 512, we trained the model for 100 epochs while maintaining a learning rate of 0.001. The choice of these parameters was designed to enhance the model's effectiveness and minimize the likelihood of overfitting throughout the training phase.

CNN-Bi-LSTM Model: We defined this with a kernel size of 3 and a pool size of 2 for the convolutional layers, accompanied by a dropout rate of 0.5. It included a densely connected layer with units decreasing in size (128, 64, 32). For optimization, we employed the Adam optimizer with a batch size of 512 and trained the model for 100 epochs, maintaining a learning rate of 0.001. These parameter choices were carefully made to enhance the model's performance and mitigate overfitting during the training phase.

CNN-Bi-GRU Model: We used a kernel size of 3 and a pool size of 2 for the convolutional layers, incorporating a dropout rate of 0.5 for regularization purposes. It featured a densely connected layer with units decreasing in size (128, 64, 32). To optimize the model, we employed the Adam optimizer with a batch size of 512 and conducted training for 100 epochs, maintaining a learning rate of 0.001. These parameter selections were carefully chosen to maximize the model's effectiveness and prevent potential overfitting during the training phase.

Bi-LSTM-Bi-GRU Model: We integrated a dropout rate of 0.25 for regularization purposes and featured a densely connected layer with units decreasing in size (128, 64, 32). To optimize performance, we employed the Adam optimizer, setting a batch size of 512 for training over 100 epochs while maintaining a consistent learning rate of 0.001. These parameter choices were thoughtfully selected to enhance the model's efficacy and mitigate potential overfitting during the training process.

Please note that we used the same DL models as the authors in [47]. Our framework enhanced the model used in [47] by using additional basic ML algorithms in the ensemble stacking DL model as a meta-learner layer of classifiers. Table 1 below illustrates the parameters used in DL models.

To maximize performance for cyberbullying detection in texts written in Arabic dialects, in our study we carefully adjusted each DL model's hyperparameters. The activation function (ReLU) kernel size (set to 3) and the number of filters (16, 32, 64, and 128) were critical hyperparameters for the CNN model. The model's ability to capture subtle features was generally enhanced by adding more filters. The number of Neuron units (32, 64, 128) and dropout rate (0.25) were the focus points for the Bi-GRU, Bi-LSTM, and the hybrid of the DL models. The model's capacity to learn long-term dependencies, crucial for deciphering context in text data, was enhanced by higher units. Because of its straightforward architecture and similar use of hyperparameters, we discovered that the Bi-GRU model provided competitive performance with less computational overhead. We used a grid search for hyperparameter tuning and evaluated combinations based on validation set performance. Our tests revealed that the model's recall, accuracy, precision, and F1-score were significantly increased by the optimal hyperparameters. A list of the DL models and their hyperparameters is illustrated in

Table 1. DL models' parameters.

Models	Parameters	Values
Deep CNN	Number of filters	[16, 32, 64, 128]
	Kernel size	3
	Pool size	2
	Dense (hidden layers)	2
	Neurons	600
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001
	Dropout	0.25
Bi-LSTM	Dense (hidden layers)	2
	Neurons	60
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001
	Dropout	0.25
Bi-GRU	Dense (hidden layers)	3
	Neurons	[128, 64, 32]
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001
	Dropout	0.5
Bi-LSTM-LSTM	Dense (hidden layers)	3
	Neurons	[128, 64, 32]
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001
	Kernel size	3
	Pool size	2
	Dropout	0.5
CNN-Bi-LSTM	Dense (hidden layers)	3
	Neurons	[128, 64, 32]
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001
	Kernel size	3
	Pool size	2
	Dropout	0.5
CNN-Bi-GRU	Dense (hidden layers)	3
	Neurons	[128, 64, 32]
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001
	Dropout	0.25
Bi-LSTM-Bi-GRU	Dense (hidden layers)	3
	Neurons	[128, 64, 32]
	Optimizer	Adam
	Batch size	512
	Epoch	100
	Learning rate	0.001

Table 1 below.

These models were selected based on their track record of handling text data, especially for challenging tasks such as cyberbullying detection, sentiment analysis, and text classification.[47]

Though they employ different strategies, LSTM Bi-LSTM and Bi-GRU are RNN models that share the goal of extracting features from text data. CNNs excel in capturing local spatial hierarchies through convolutional operations, while RNNs, particularly LSTM and GRU variants, are designed to handle sequential data and capture temporal dependencies. Bi-LSTM and Bi-GRU perform better than their unidirectional counterparts and enhance context comprehension by processing the input in two directions. By fusing CNN's local pattern recognition skills with RNN's sequence learning capabilities the hybrid models offer a more comprehensive approach to text analysis.[47]

2.3 Ensemble Stacking Model

In Ensemble models, several ML or DL models are combined into a single prediction model positively impacting the performance.[48] Hybrid types of CNN, LSTM, and DL models have been increasingly used recently and significantly improved the performance of text classification.[49]

Bagging, boosting, and stacking are the ensemble learning techniques. Bagging is a technique where the same algorithm is trained multiple times using different subsets sampled from the training data. The average of all the predictions made by each sub-model is then computed for the final output. Bagging generally lowers the variance of classification errors improving classification accuracy.[50]

In boosting, several models are sequentially trained and each model gains knowledge from the errors of its predecessors. Adaptive resampling is used to obtain learning by selecting a misclassified data point from an earlier classifier more often than a correctly classified one. A weight divided by the number of cycles in the training iteration is assigned to each training data point. In the next iteration, the classifier needs to concentrate on reweighting data points that were incorrectly classified in the previous iteration. The final classifier is the weighted average of the ensemble predictions.[50]

Combining different ML models (base learners) with another data mining technique is the aim of stacking. After the base learners, a combiner also referred to as a meta-classifier is trained to produce a final prediction that is based on the base learners' predictions. Such stacked ensembles have been shown to outperform any individual base learners and to constitute an asymptotically ideal system for learning.[50]

In our experiment, we used ensemble stacking techniques combining different ML models with DL models; these

models illustrated outstanding performance in the classification of Arabic dialects' text. We implemented three ensemble stacking models as follows: the first combined the two DL models that exhibited the highest performance based on validation accuracy; the second merged four DL models that demonstrated the best performance in terms of validation accuracy, while the third encompassed all seven DL models. To evaluate the performance of these ensemble stacking DL models, we focused on accuracy as the primary metric. Accuracy serves as a widely used evaluation measure for assessing text classification models, representing the ratio of correctly classified instances to the total instances. Renowned for its simplicity and clarity, accuracy offers a comprehensive understanding of the model's precision in making predictions. In amalgamating these high-performing DL models into a stacking ensemble, our objective was to harness their unique strengths while capitalizing on the possible synergies and complementary aspects among these models.

Incorporating a stacking technique enables the consolidation of forecasts generated by multiple models, offering the potential to enhance the overall effectiveness and resilience of sentiment analysis.

Three experiments involved the use of varying numbers of pre-trained DL models: two in the first, four in the second, and seven in the third. A stacked input dataset was created by consolidating predictions from the base models for each data model. The meta-learner was trained using this dataset and actual labels, predicting test data using the base models. These predictions were then merged to form a new input for the meta-learner. Finally, the ensemble stacking model's performance on test data was assessed, visually represented in Figs. 1, 2, and 3, combining the 2, 4, and 7 DL models, respectively.

2.4 Framework for cyberbullying detection in Arabic dialects' texts

Firstly, the comments are gathered from the Instagram platform and then annotated. This is the general workflow with our framework. Secondly, pre-processing is performed on the labeled comments. Thirdly, the characteristics are removed. Fourthly, a variety of ensemble stacking, and DL models are used to classify the comments. Ultimately, measurements are made of F-score, accuracy, recall, and precision. A framework for detecting cyberbullying in Arabic dialects is shown in Fig. 4.

3. Results and discussion

To train and evaluate the DL models we used the TensorFlow and Keras libraries. The datasets were sectioned into three splits for these experiments: Training, Testing, and Validating (80/10/10), and the best outcomes were noted. We used Adam optimizers with epoch 100 and an early stopping parameter for all the models so that the iteration would end if no gains were achieved. For every experiment, we used a state-of-the-art PC running Windows 11 with a Core i7 processor and 32GB of RAM.

Table 2 presents the results of an experimental evaluation of various neural network models conducted for a classification task. The standout performer was the Bi-LSTM, achieving the highest accuracy at 87.4%. This model showcased excellence across multiple metrics, including recall (86.8%), precision (86.2%), and F-measure (86.5%). Following closely, the Bi-LSTM-Bi-GRU secured the second-highest accuracy at 87.1%, demonstrating strong recall (86.9%), precision (85.9%), and an impressive F-measure of 86.3%. The Deep CNN also delivered notable performance, securing the third highest accuracy at 86.7%, with competitive scores in recall (86.6%), precision (85.3%), and F-measure (85.9%). This finding aligns which the Bi-LSTM performance on the text classification task of sentiment analysis of the Tunisian dialect was impressive achieving an accuracy of 87%.[51]

Additionally, these results align with those from other investigations. Elfaik and Nfaoui (2021)[52] analyzed the sentiment in an ArTwitter dataset using several DL models. In terms of F-measure, accuracy, recall, and precision, Bi-LSTM performed the best. In their explanation of their findings, the authors stated that Bi-LSTM's ability to learn the context of each word in the text more effectively is caused by its ability to access contextual data that comes before and after it by combining a forward hidden layer with a backward hidden layer. The authors found that Bi-LSTM outperforms LSTM in finding richer semantic information and making better use of contextual data.[52]

Table 2. Results for the DL models.

Model	Accuracy	Recall	Precision	F-Measure
Deep CNN	86.7	86.6	85.3	85.9
Bi-LSTM	87.4	86.8	86.2	86.5
Bi-GRU	86.8	86.4	85.5	85.9
Bi-LSTMLSTM	86.4	86.6	85.1	85.7
CNN-Bi-LSTM	85.7	84.9	84.4	84.6
CNN-Bi-GRU	86.5	85.8	85.3	85.5
Bi-LSTM-BiGRU	87.1	86.9	85.9	86.3

The results underscore the efficacy of LSTM-based architectures, particularly the Bi-LSTM model, for the given classification task, highlighting the importance of selecting appropriate neural network structures tailored to specific tasks. Based on the highest accuracies achieved by the DL models evaluated in this study, we proceeded to construct stacked models for enhanced predictive performance. We employed a two-layer stacking approach, wherein the first layer consisted of two models: the Bi-LSTM with an accuracy of 87.4% and the Bi-LSTM-Bi-GRU with an accuracy of 87.1%. These two models were chosen based on their superior individual accuracies.

Additionally, a second layer of stacked models was formulated using the four models with the best accuracy. The

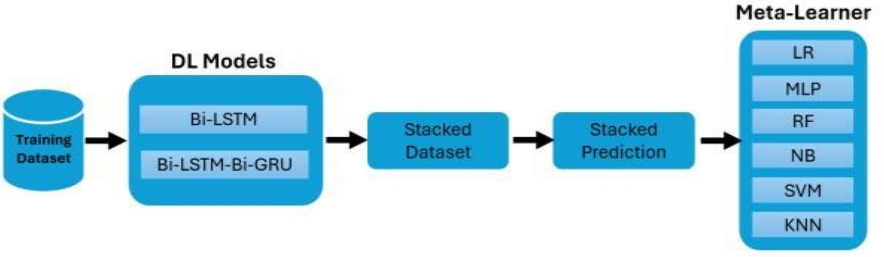

Fig. 1 Ensemble stacking model combining 2 DL models.

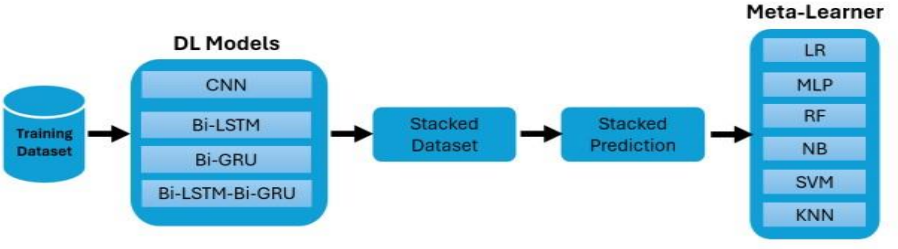

Fig. 2 Ensemble stacking model combining 4 DL models.

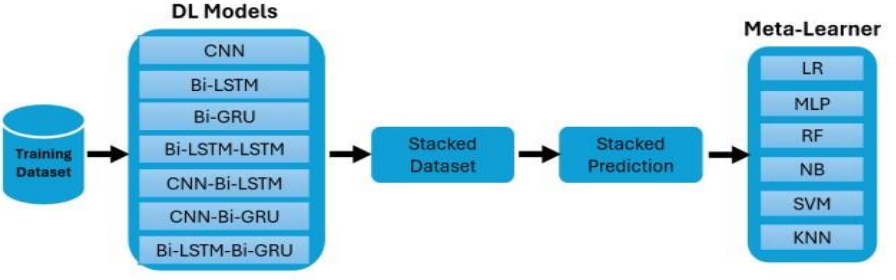

Fig. 3 Ensemble stacking model combining 7 DL models.

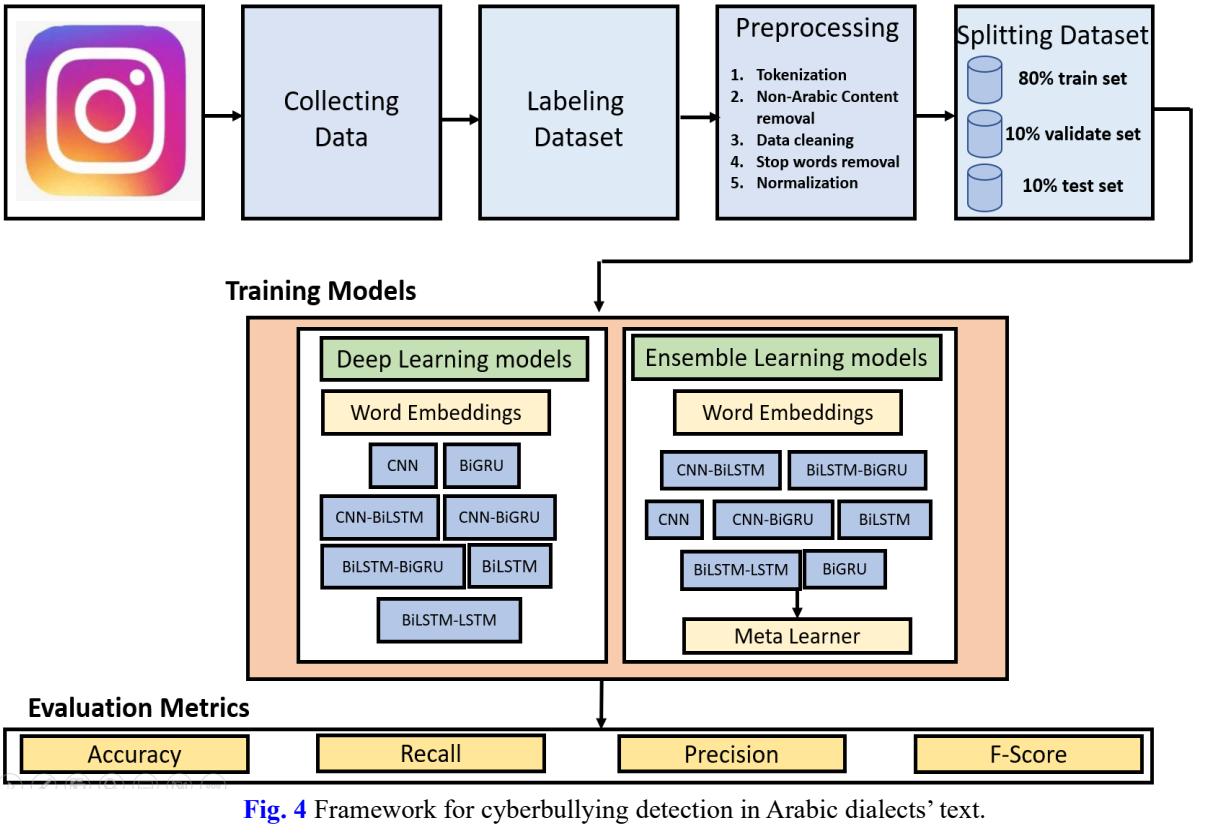

Fig. 4 Framework for cyberbullying detection in Arabic dialects' text.

Table 3. Results of the ensemble stacking models' accuracy.

Stacked Model	Stacked 2 models	Stacked 4 models	Stacked 7 models
Ensemble Stacking Model with LR	87.35	87.42	87.61
Ensemble Stacking Model with MLP Meta-Learner Layer	87.33	87.67	87.87
Ensemble Stacking Model with RF Meta-Learner Layer	94.60	94.70	94.73
Ensemble Stacking Model with NB Meta-Learner Layer	87.25	87.44	87.50
Ensemble Stacking Model with SVM Meta-Learner Layer	87.26	87.42	87.61
Ensemble Stacking Model with KNN Meta-Learner Layer	8.58	88.75	88.80

selected models for this layer were the Bi-LSTM (87.4%), Bi-LSTM-Bi-GRU (87.1%), Bi-GRU (86.8%), and Deep CNN (86.6%).

A comprehensive stacking model involving all seven DL models was created, using Bi-LSTM, Bi-LSTM-Bi-GRU, Bi-GRU, Deep CNN, CNN-Bi-GRU, Bi-LSTM-LSTM, and CNN-Bi-LSTM. This ensemble aimed to leverage the diverse strengths of each model to enhance overall predictive capabilities.

The stacking models were designed with the intent of optimizing predictive accuracy by combining the strengths of individual models synergistically. The accuracies provided for each model in the dataset (Bi-LSTM, Bi-LSTM-Bi-GRU, Bi-GRU, Deep CNN, CNN-Bi-GRU, Bi-LSTM-LSTM, CNN-Bi-LSTM) served as the foundation for the selection and construction of the stacked models, as illustrated in Figs. 1, 2, and 3. These stacking strategies represent an approach to harness the collective power of multiple DL models for improved accuracy in the context of the given dataset.

As illustrated in Table 3, the ensemble stacking models, which integrate a range of base classifiers, demonstrated remarkable outcomes in predictive accuracy. Specifically, the ensemble stacking model with a MLP meta-learner classifier layer, comprising seven base DL models, attained the highest accuracy of all stacked models. This observation aligns with the conclusions drawn by several researchers, supporting the notion that combining multiple DL models frequently leads to improved classification performance. The ensemble stacking model with the RF meta-learner layer of the classifier consistently displayed impressive accuracy across different ensemble models, reaching its peak at 94.73% and the best performance when combining 4 DL models. This consistent and robust performance of the ensemble stacking model with the RF meta-learner layer of the classifier highlights the efficacy of incorporating diverse models in ensemble learning strategies.

These results align with the results achieved in using the ensemble stacking approach for text classification and sentiment analysis experiments. Saleh *et al*. (2022)[53] presented an ensemble stacking model in which the authors integrated three DL models and employed three classifiers as a meta-learner layer to improve the accuracy of their model. Using an LR classifier as the meta-learner layer in their ensemble-stacked model allowed them to attain the highest accuracy performance of 98.08%. Furthermore,[49] employed an SVM and LR classifier meta-learner layer to improve the model's performance and the authors were able to attain a 95.81% accuracy. Habbat *et al*. (2022) suggested an ensemble stacking model that incorporated a meta-learner layer made up of LR and MLP with four DL models.[48] When the authors used MLP as a meta-learner classifier the best accuracy achieved was 97.5%. The model was tested using datasets in Arabic and French.

4. Conclusion

In conclusion, this research has significantly expanded our insight into cyberbullying detection within the Arabic language, showcasing the efficacy of employing ensemble stacking models to identify offensive language precisely. The study's thorough evaluation of seven DL models, with a special focus on ensemble stacking, demonstrates the potential of advanced computational techniques in addressing the intricate challenges of cyberbullying. Particularly, the ensemble stacking model with an RF meta-learner classifier layer consistently displayed impressive accuracy across different ensemble models with other meta-learner layers of classifiers, reaching its peak at 94.73% with a stacking model that combined seven DL models.

The reliable and strong performance of the combination stacking model with the RF meta-learning layer of the classifier highlights the success of including a variety of models, in learning approaches. In the future research, in this field could investigate incorporating these cutting-edge language detection models into platforms. By integrating models there is potential to create safer digital environments by quickly recognizing and addressing cases of cyberbullying. This proactive approach aligns with ongoing initiatives to cultivate a more inclusive and supportive online environment, fostering engagement without the fear of harassment. Ongoing research efforts could focus on refining the ensemble stacking methodology and assessing its adaptability to the evolving landscape of online communication. Sustained collaboration among researchers, industry stakeholders, and policymakers is essential for effectively implementing these advancements and ensuring a positive digital experience for users globally.

The findings of this study have significant implications for both academic research and practical applications. The generated models for detecting cyberbullying can be

practically integrated into social media sites educational monitoring systems and law enforcement tools to enhance the identification and prevention of cyberbullying in Arabic-speaking communities, The results of this study have important ramifications for both academia and applications. In order to enhance the identification and prevention of cyberbullying, within Arabic-speaking societies the created models, for detecting cyberbullying could be feasibly incorporated into social networking sites, educational supervision systems, and law enforcement tools. Academically, this study adds to the small corpus of research on the identification of cyberbullying in Arabic dialects demonstrating the suitability of sentiment analysis methods for this specialized field.

However, the study met several challenges including the dearth of datasets for cyberbullying detection with annotations for Arabic dialects. Overfitting mitigation and identifying the best model architectures and hyperparameters required a great deal of trial and error. Additionally, the management of lengthy training periods and massive data processing needs high-performance computing resources.

Acknowledgments
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Conflict of Interest
There is no conflict of interest.

Supporting Information
Not applicable.

References
[1] A. Alshamsi, R. Bayari, S. Salloum, Sentiment analysis in English texts, *Advances in Science, Technology and Engineering Systems Journal*, 2020, **5**, 1683-1689, doi: 10.25046/aj0506200.
[2] A. A. Al Shamsi, S. Abdallah, Text mining techniques for sentiment analysis of Arabic dialects: literature review, *Advances in Science, Technology and Engineering Systems Journal*, 2021, **6**, 1012-1023, doi: 10.25046/aj0601112.
[3] J. W. Patchin, S. Hinduja, Bullies move beyond the schoolyard, *Youth Violence and Juvenile Justice*, 2006, **4**, 148-169, doi: 10.1177/1541204006286288.
[4] P. W. Agatston, R. Kowalski, S. Limber, Students' perspectives on cyber bullying, *Journal of Adolescent Health*, 2007, **41**, S59-S60, doi: 10.1016/j.jadohealth.2007.09.003.
[5] R. Albayari, S. Abdallah, K. Shaalan, Cyberbullying detection model for Arabic text using deep learning, *Journal of Information & Knowledge Management*, 2024, 2450016, doi: 10.1142/s0219649224500163.
[6] A. Elnagar, L. Lulu, O. Einea, An annotated huge dataset for standard and colloquial Arabic reviews for subjective sentiment analysis, *Procedia Computer Science*, 2018, **142**, 182-189, doi: 10.1016/j.procs.2018.10.474.
[7] R. ALBayari, S. Abdallah, Instagram-based benchmark dataset for cyberbullying detection in Arabic text, *Data*, 2022, **7**, 83, doi: 10.3390/data7070083.
[8] R. ALBayari, S. Abdullah, S. A. Salloum, Cyberbullying classification methods for Arabic: a systematic review. The International Conference on Artificial Intelligence and Computer Vision, 2021, 375-385, doi: 10.1007/978-3-030-76346-6_35.
[9] J. Bale, Arabic as a heritage language in the United States. *International Multilingual Research Journal*, 2010, **4**, 125-151, doi: 10.1080/19313152.2010.499041
[10] E. A. Abozinadah, A. V. Mbaziira, J. H. Jones, Detection of abusive accounts with Arabic tweets, *International Journal of Knowledge Engineering-IACSIT*, 2015, **1**, 113-119, doi: 10.7763/ijke.2015.v1.19.
[11] P. Baby, B. Krishnapriya, Sentimental analysis and deep learning: a survey, *International Journal of Scientific Research in Science, Engineering and Technology*, 2020, 212-220, doi: 10.32628/ijsrset207135.
[12] M. A. Al-Ajlan, M. Ykhlef, Deep learning algorithm for cyberbullying detection, *International Journal of Advanced Computer Science and Applications*, 2018, **9**, 199–205, doi: 10.14569/ijacsa.2018.090927.
[13] H. Rosa, D. Matos, R. Ribeiro, L. Coheur, J. P. Carvalho, A "deeper" look at detecting cyberbullying in social networks, 2018 International Joint Conference on Neural Networks (IJCNN). Rio de Janeiro, Brazil. IEEE, 2018.
[14] B. A. H. Murshed, J. Abawajy, S. Mallappa, M. A. N. Saif, H. D. E. Al-Ariki, DEA-RNN: a hybrid deep learning approach for cyberbullying detection in twitter social media platform, *IEEE Access*, 2022, **10**, 25857-25871, doi: 10.1109/ACCESS.2022.3153675.
[15] R. Bayari, A. Bensefia, Text mining techniques for cyberbullying detection: state of the art, *Advances in Science, Technology and Engineering Systems Journal*, 2021, **6**, 783-790, doi: 10.25046/aj060187.
[16] A. Wahdan, S. AL Hantoobi, S. A. Salloum, K. Shaalan, A systematic review of text classification research based on deep learning models in Arabic language, *International Journal of Electrical and Computer Engineering (IJECE)*, 2020, **10**, 6629, doi: 10.11591/ijece.v10i6.pp6629-6643.
[17] H. Rosa, N. Pereira, R. Ribeiro, P. C. Ferreira, J. P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A. M. Veiga Simão, I. Trancoso, Automatic cyberbullying detection: a systematic review, *Computers in Human Behavior*, 2019, **93**, 333-345, doi: 10.1016/j.chb.2018.12.021.
[18] H. Mubarak, K. Darwish, W. Magdy, Abusive Language Detection on Arabic Social MediaProceedings of the First Workshop on Abusive Language Online. Vancouver, BC, Canada. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017.
[19] A. A. Al Shamsi, S. Abdallah, Sentiment analysis of emirati dialect, *Big Data and Cognitive Computing*, 2022, **6**, 57, doi: 10.3390/bdcc6020057.

[20] H. Fouadi, H. El Moubtahij, H. Lamtougui, A. Yahyaouy, Sentiment analysis of Arabic comments using machine learning and deep learning models, *Indian Journal of Computer Science and Engineering*, 2022, **13**, 598-606, doi: 10.21817/indjcse/2022/v13i3/221303003.
[21] S. M. Abdou, A. M. Moussa, Arabic speech recognition: challenges and state of the art, *Computational Linguistics, Speech and Image Processing for Arabic Language*, 2019, 1-27, doi: 10.1142/9789813229396_0001.
[22] A. A. Al Shamsi, S. Abdallah, A systematic review for sentiment analysis of Arabic dialect texts researches, International Conference on Emerging Technologies and Intelligent Systems. Cham: Springer, 2022.
[23] A. Alshutayri, E. Atwell, A social media corpus of Arabic dialect text, Computer-mediated communication and social media corpora, clermont-ferrand: Presses Universitaires Blaise Pascal, 2018.
[24] A. A. Al Shamsi, S. Abdallah, Ensemble stacking model for sentiment analysis of emirati and Arabic dialects, *Journal of King Saud University - Computer and Information Sciences*, 2023, **35**, 101691, doi: 10.1016/j.jksuci.2023.101691.
[25] N. Hicham, S. Karim, N. Habbat, Customer sentiment analysis for Arabic social media using a novel ensemble machine learning approach, *International Journal of Electrical and Computer Engineering (IJECE)*, 2023, **13**, 4504, doi: 10.11591/ijece.v13i4.pp4504-4515.
[26] H. Saleh, S. Mostafa, A. Alharbi, S. El-Sappagh, T. Alkhalifah, Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis, *Sensors*, 2022, **22**, 3707, doi: 10.3390/s22103707.
[27] S. A. Kokatnoor, B. Krishnan, Twitter Hate Speech Detection using Stacked Weighted Ensemble (SWE) Model, 2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN). Bangalore, India. IEEE, 2020.
[28] B. Haidar, M. Chamoun, A. Serhrouchni, Arabic cyberbullying detection: enhancing performance by using ensemble machine learning, 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData). Atlanta, GA, USA. IEEE, 2019.
[29] M. K. Dalal, M. A. Zaveri, Automatic text classification: a technical review, *International Journal of Computer Applications*, 2011, **28**, 37-40, doi: 10.5120/3358-4633.
[30] K. Siddhartha, K. R. Kumar, K. J. Varma, M. Amogh, M. Samson, Cyber bullying detection using machine learning, 2022 2nd Asian Conference on Innovation in Technology (ASIANCON). Ravet, India. IEEE.
[31] A. Faraz, An elaboration of text categorization and automatic text classification through mathematical and graphical modelling, *Computer Science & Engineering*, 2015, **5**, 1-11, doi: 10.5121/cseij.2015.5301.
[32] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, S. Belfkih, ASA: A framework for Arabic sentiment analysis, *Journal of Information Science*, 2020, **46**, 544-559, doi: 10.1177/0165551519849516.
[33] B. A. Rachid, H. Azza, H. H. Ben Ghezala, Classification of cyberbullying text in Arabic, 2020 International Joint Conference on Neural Networks (IJCNN). Glasgow, UK. IEEE.
[34] M. Djandji, F. Baly, W. Antoun, H. Hajj, Multi-task learning using AraBert for offensive language detection, Proceedings of the 4th Workshop on Open-source arabic corpora and processing tools, with a shared task on offensive language detection, 2020.
[35] H. Mohaouchane, A. Mourhir, N. S. Nikolov, Detecting offensive language on Arabic social media using deep learning, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). Granada, Spain. IEEE, 2019.
[36] F. Husain, J. Lee, S. Henry, A survey of offensive language detection for the Arabic language, *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 2021, **20**, 1-44, doi: 10.1145/3421504.
[37] N. Albadi, M. Kurdi, S. Mishra, Are they our brothers? analysis and detection of religious hate speech in the Arabic twittersphere, 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). Barcelona, Spain. IEEE, 2018.
[38] A. A. Al Shamsi, S. Abdallah, Ensemble stacking model for sentiment analysis of emirati and Arabic dialects, *Journal of King Saud University - Computer and Information Sciences*, 2023, **35**, 101691, doi: 10.1016/j.jksuci.2023.101691.
[39] F. Husain, Arabic offensive language detection using machine learning and ensemble machine learning approaches, 2020.
[40] H. H. Saeed, T. Calders, F. Kamiran, OSACT4 shared tasks: Ensembled stacked classification for offensive and hate speech in Arabic Tweets, Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020.
[41] M. K. A. Aljero, N. Dimililer, A novel stacked ensemble for hate speech recognition, *Applied Sciences*, 2021, **11**, 11684, doi: 10.3390/app112411684.
[42] L. Wen, M. Hughes, Coastal wetland mapping using ensemble learning algorithms: a comparative study of bagging, boosting and stacking techniques, *Remote Sensing*, 2020, **12**, 1683, doi: 10.3390/rs12101683.
[43] R. ALBayari, S. Abdallah, Instagram-based benchmark dataset for cyberbullying detection in Arabic text, *Data*, 2022, **7**, 83, doi: 10.3390/data7070083.
[44] A. B. Soliman, K. Eissa, S. R. El-Beltagy, AraVec: A set of Arabic Word Embedding Models for use in Arabic NLP, *Procedia Computer Science*, 2017, **117**, 256-265, doi: 10.1016/j.procs.2017.10.117.
[45] A. Oussous, F.-Z. Benjelloun, A. A. Lahcen, S. Belfkih, ASA: A framework for Arabic sentiment analysis, *Journal of Information Science*, 2020, **46**, 544-559, doi:

10.1177/0165551519849516.
[46] Z. Nassr, N. Sael, F. Benabbou, Preprocessing Arabic dialect for sentiment mining: state of art, *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2020, **44**, 323-330, doi: 10.5194/isprs-archives-xliv-4-w3-2020-323-2020.
[47] A. A. Al Shamsi, S. Abdallah, Ensemble stacking model for sentiment analysis of emirati and Arabic dialects, *Journal of King Saud University - Computer and Information Sciences*, 2023, **35**, 101691, doi: 10.1016/j.jksuci.2023.101691.
[48] N. Habbat, H. Anoun, L. Hassouni, H. Nouri, Analyzing Booking's comments using stacking ensemble deep learning model and neural topic model, *SSRN Electronic Journal*, 2022, 4181357, doi:10.2139/ssrn.4181357.
[49] H. Saleh, S. Mostafa, L. A. Gabralla, A. O. Aseeri, S. El-Sappagh, Enhanced Arabic sentiment analysis using a novel stacking ensemble of hybrid and deep learning models, *Applied Sciences*, 2022, **12**, 8967, doi: 10.3390/app12188967.
[50] L. Wen, M. Hughes, Coastal wetland mapping using ensemble learning algorithms: a comparative study of bagging, boosting and stacking techniques, *Remote Sensing*, 2020, **12**, 1683, doi: 10.3390/rs12101683.
[51] A. Masmoudi, J. Hamdi, L. Hadrich Belguith, Deep learning for sentiment analysis of Tunisian dialect, *Computación y Sistemas*, 2021, **25**, 129–148, doi: 10.13053/cys-25-1-3472.
[52] H. Elfaik, E. H. Nfaoui, Deep bidirectional LSTM network learning-based sentiment analysis for Arabic text, *Journal of Intelligent Systems*, 2021, **30**, 395-412, doi: 10.1515/jisys-2020-0021.
[53] H. Saleh, S. Mostafa, A. Alharbi, S. El-Sappagh, T. Alkhalifah, Heterogeneous ensemble deep learning model for enhanced Arabic sentiment analysis, *Sensors*, 2022, **22**, 3707, doi: 10.3390/s22103707.