IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 3, June 2026, pp. 2849~2862
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i3.pp2849-2862

Hybrid convolutional neural network-bidirectional long short-term memory model for Arabic sentence readability assessment

Mohamed Amine Ouassil[1], Mohammed Jebbari[1], Rabia Rachidi[2], Mouaad Errami[1], Soufiane Hamida[3,5], Bouchaib Cherradi[1,2,3,4], Abdelhadi Raihani[1,3]
[1]EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco
[2]LaROSERI Laboratory, Faculty of Science, Chouaib Doukkali University, El Jadida, Morocco
[3]2IACS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca, Mohammedia, Morocco
[4]STIE Team, CRMEF Casablanca, Provincial Section of El Jadida, El Jadida, Morocco
[5]GENIUS Laboratory, SupMTI of Rabat, Rabat, Morocco

Article Info

Article history:

Received Sep 25, 2024
Revised Mar 6, 2026
Accepted Apr 21, 2026

Keywords:

Arabic natural language processing
Bidirectional long short-term memory
Convolutional neural network
Readability evaluation
Text classification

ABSTRACT

In the current educational landscape, a large number of educators prefer using generative artificial intelligence techniques to produce textual content to be presented for learning. However, these generated texts may not meet the specific needs of learners or align with their abilities. Many traditional methods and techniques can be employed to assess the complexity of a text, such as traditional readability formulas, but these techniques are time-consuming and labor-intensive. In this paper, we introduce a deep learning approach for automatically evaluating the readability of Arabic texts by analyzing and classifying sentences into different difficulty levels within educational content. The initial stage of the proposed approach is preprocessing textual content and leveraging natural language processing (NLP) methodologies for feature extraction, such as Word2Vec. The approach then concentrates on refining and evaluating a deep learning model to classify text into different readability levels. This paper introduces a hybrid classification model that combines convolutional neural networks (CNNs) and bidirectional long short-term memory (BiLSTM) layers, attaining an accuracy of 96.68%. This model demonstrates the significance of applying hybrid deep learning models in analyzing educational materials and establishes a foundation for subsequent progress in the field of automated Arabic readability assessment.

Corresponding Author:

Bouchaib Cherradi
EEIS Laboratory, ENSET of Mohammedia, Hassan II University of Casablanca
Mohammedia 28830, Morocco
Email: bouchaib.cherradi@gmail.com

1. INTRODUCTION

In modern educational environments, it is widely recognized that students display diverse levels of learning progress. Within the same classroom or study group, the knowledge and understanding of one student can significantly vary from that of their peers. To optimize learning outcomes and ensure effective teaching, adopting innovative learning strategies is crucial for teachers [1]. One solution to addressing this challenge is the teaching at the right level (TaRL) approach, a personalized learning methodology that helps learners develop essential reading and math skills [2]. This strategy recognizes the individual learning differences and pace of each student [3]. It is pivotal in addressing the varying academic abilities within a classroom, ensuring that all students, regardless of their starting point, and have the opportunity to advance in their learning journey [4]. It calls for educators to tailor the educational content to match each student's

unique needs [5], [6]. Educators play an important role in this adaptive learning model [7]. By measuring the readability of a text, teachers can ensure the material meets the specific needs of each learner [8]. This methodology allows learners to advance according to their personal learning rhythm, avoiding any sense of being swamped with information or disinterest due to lack of challenge. Readability is influenced by a range of factors, such as vocabulary complexity, sentence and paragraph structure, and various textual features [9]. These elements are critical for educators to consider when selecting or creating educational materials, allowing them to adjust the difficulty level to match their students' abilities.

However, the manual evaluation of text readability is both time-consuming and costly, limiting its practical application in educational settings [10]. For the Arabic language, one of the main challenges that face readability assessment is the language's inherent complexity [11]. The assessment of readability in the Arabic language poses additional challenges due to the language's complex morphology, extensive lexicon, and diverse sentence construction [12]. Arabic's high level of inflection, which allows single words to convey various grammatical nuances, further complicates this task. These difficulties have spurred interest in seeking technological innovations for automating the process of readability assessment. Among these solutions, deep learning and natural language processing (NLP) have emerged as particularly promising approaches. Both NLP and deep learning are interdisciplinary fields that merge computer science with artificial intelligence [13], [14], equipping machines with the ability to process, understand, and interpret human language in a meaningful and contextually appropriate manner. NLP, in the context of readability measurement, utilizes text classification techniques to assess the readability of texts by categorizing them into various readability levels or categories [15], [16]. This approach involves training machine learning models to analyze textual features and make predictions about the readability of a given text [17]. The integration of deep learning techniques within this domain offers the potential for even more refined analysis. Deep learning's ability to process and learn from large datasets can uncover subtle patterns in language use that are indicative of text complexity, thereby enhancing the precision of readability predictions [18].

The field of text readability assessment is an area of increasing interest [19]. Reviewing previous research efforts offers important background for the current study, emphasizing their primary contributions, techniques, and results, with a particular focus on the methodologies employed to address the issues of text readability and their effectiveness in attaining high prediction accuracy. For instance, Nassiri *et al.* [20] employed machine learning techniques to evaluate the readability of modern standard Arabic (MSA) texts. They mapped the resulting readability levels to the interagency language roundtable (ILR) scale for assessing second language proficiency. From an initial set of 170 readability features, the proposed approach achieved an accuracy of 90.43% in a three-class classification task using the instance-based k-nearest neighbors (IBK) and random forest algorithms. However, this study was based on traditional machine learning techniques, which have limited capacity to attain high accuracy. Nassiri *et al.* [21] explore the use of statistical machine learning to predict the readability of Arabic texts for second-language learners. Using various feature sets and algorithms, the study aims to maintain or increase prediction accuracy while simplifying the feature set. Experiments showed maximum accuracies of up to 86.15%, confirming that a reduced feature set could achieve comparable results. Saddiki *et al.* [22] present a comprehensive computational study on Arabic readability, a challenging task given Arabic's morphological complexity and low-resource status. The study evaluates both first-language (L1) and second-language (L2) Arabic readability, uniquely combining them in the same experimental setting. By incorporating advanced features like language modeling, the study achieves a 94.8% accuracy for L1 and a 74.1% accuracy for L2, marking significant error reductions compared to traditional methods. The paper suggests that using L1-generated models can enhance L2 readability assessment.

Berrichi *et al.* [23] used various methods for assessing the complexity of text for native (L1) speakers. Two primary approaches were evaluated. The first approach involved analyzing specific, handcrafted features that help determine a text's difficulty level. In the second approach, authors explored word embeddings (including continuous bag-of-words model (CBOW), skip-gram, and Arabic bidirectional encoder representations from transformers (AraBERT)) as an alternative to the handcrafted features. The AraBERT model performed best among the techniques, achieving a prediction accuracy of approximately 76.93%. Bessou and Chenni [24] employ machine learning to predict text complexity, aiming to make Arabic text more accessible to learners. Various classifiers and features like count and term frequency-inverse document frequency (TF-IDF) were tested, revealing that a combination of n-gram features and specific data representations significantly boosted performance. The most accurate results were achieved using TF-IDF vectors and support vector machine (SVM), reaching an overall accuracy of 87.14% across four complexity classes.

Saddiki *et al.* [25] evaluate simple readability features using publicly accessible and open-source tools, achieving performance comparable to existing Arabic readability studies. The study found that cost-effective features like morpheme counts, type and token counts, and part-of-speech tags had a good

cost/benefit ratio. The best results in terms of accuracy were 73.74% achieved by random forest classifier. Baazeem *et al.* [26] explored using eye-tracking to enhance Arabic text readability prediction, revealing that eye-tracking data, when combined with linguistic features, can more accurately assess readability. The effectiveness varies by algorithm, suggesting further investigation is needed. Limitations include insufficient data and the use of basic machine learning models. Ouassil *et al.* [27] present a novel approach for Arabic readability assessment using a hybrid BERT-bidirectional long short-term memory (BiLSTM) model. This method marks a significant advancement in text difficulty classification. The proposed model achieved an accuracy rate of 89.53%.

The preceding review reveals a clear opportunity to improve upon the performance and accuracy of classical models. Therefore, in this study a novel approach based on NLP and hybrid deep learning model is presented to assess Arabic text readability by classifying sentences into different levels of difficulty. The major contributions of this study include:

– Developing an accurate system to automatically classify the difficulty of Arabic sentences using a combination of two deep learning techniques: convolutional neural network (CNN) and BiLSTM.
– Analyzing the effect of convolutional depth on the BiLSTM layer by studying the impact of adding a varying number of convolutional layers.

The remainder of this paper is structured as follows. Section 2 describes the approach, materials, and algorithms utilized in developing the proposed system. Results are presented in section 3. The final section 4 concludes and summarizes the paper.

2. METHOD

This work proposes a computational model for automatically analyzing and grading the linguistic complexity levels of Arabic text, classifying it as easy, moderately difficult, or complex to read and understand. A deep learning model was employed to achieve this, and it was trained to accurately classify sentences based on their linguistic and semantic characteristics. Figure 1 illustrates the proposed approach that consists of four main stages: data representation, preprocessing, word representation, and classification. In the data representation stage, the dataset used in the experiments was described, including its textual content, the distribution of sentences by complexity, key vocabulary statistics, and a visualization of the dataset's content using a word cloud technique. In the preprocessing stage, the text was clean and standardize to prepare it for analysis. This consists of several key operations: text cleaning removes irrelevant noise like punctuation or HTML tags, normalization ensures uniformity by standardizing word forms, and tokenization breaks each sentence down into its individual words called tokens. In the word representation stage, the word embedding technique Word2Vec was employed to convert the clean text into a numerical format the model can process. Word2Vec generates a unique vector for each word, mathematically capturing its semantic and syntactic context. Lastly, these contextual word vectors as input were used for the proposed deep learning model. The model was finetuned, trained, and tested to classify each sentence by its complexity.

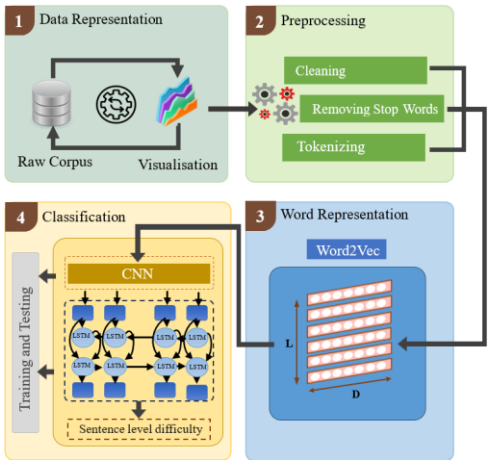

Figure 1. The flowchart of the proposed system

2.1. Dataset representation

The dataset used in this research was collected from one of the largest Arabic language learning platforms (Al-Jazeera Learning Arabic), which is an open-source tool designed to simplify Arabic language

acquisition through interactive and easy-to-navigate content [27]. For this end, the authors used a scraping tool to collect texts from this platform that were already labeled and classified by the platform's editors into three different learning categories: beginner, intermediate, and advanced. This dataset comprises two columns: the 'Text' column contains sentences as textual content, and the 'Label' column contains numbers from 1 to 3, representing the level of sentence difficulty. Each sentence was categorized into one of three proficiency levels: beginning, intermediate, or advanced. Beginning sentences are simple in structure and vocabulary for beginners. Intermediate sentences are moderately complex, suitable for learners with basic language skills. Advanced sentences contain complex grammar and extensive vocabulary for learners with high proficiency. Table 1 showcases examples from the dataset.

Table 2 provides an overview of sentence distribution across three language proficiency levels within the dataset, containing 4,519 sentences, 81,743 tokens, and 23,188 unique words. It illustrates a progression from the beginner class, with 1,418 sentences and 8,444 tokens, to the advanced class, with 1,356 sentences and 41,904 tokens, indicating increasing sentence complexity and vocabulary size. Sentence length varies from 2 to 40 tokens, and the number of letters per sentence ranges from 4 to 265 across the classes.

Table 1. Labelled samples

Sentences and its translations	Classes
يبدأ الحصاد في نوفمبر من كل سنة ويستمر لعشرة أيام	Beginning
The harvest begins in November of each year and lasts for ten days.	
وتختلف الآراء بشأن هذا النوع من العلاجات فبعض الناس لا يؤمن به وبعضهم يعتمدون عليه اعتمادا كليا ويعتبره آخرون مساعدا ومكملا للطب الحديث	Intermediate
Opinions differ about this type of treatment, some people do not believe in it, some rely on it completely, and others consider it an assistant and complement to modern medicine.	
فاجأ طفل من جنوب إفريقيا ولد بفيروس أتش أي في الخبراء بتعافيه تماما من الفيروس بعد عام واحد فقط من العلاج تلته ثمانية أعوام ونصف لم يتناول خلالها أي عقاقير	Advanced
A South African child born with HIV surprised experts by completely recovering from the virus after just one year of treatment, followed by eight and a half years without taking any medication.	

Table 2. Number of sentences per class

Classes	Number of sentences	Total tokens	Vocabulary size	Min tokens per sentence	Max tokens per sentence	Max letters per sentence	Min letters per record
Beginner	1,418	8,444	2,747	2	10	192	4
Intermediate	1,745	31,395	12,496	12	24	166	41
Advanced	1,356	41,904	15,510	25	40	265	94
Overall	4,519	81,743	23,188	2	40	265	4

To gain an initial understanding of the textual content within a dataset, the word-cloud technique was utilized. This method serves as a tool for the visual representation of word frequency, thereby providing an immediate overview of the most prominent terms and key terminologies of each class within the dataset. Figure 2 illustrate word cloud presentations of the three classes within the dataset. Figure 2(a) is the word cloud presentation of beginner class, Figure 2(b) is the word cloud presentation of intermediate class, and Figure 2(c) is the word cloud presentation of advanced class.

(a) (b) (c)

Figure 2. Word cloud presentations for the three classes within the dataset of (a) beginner, (b) intermediate, and (c) advanced

2.2. Preprocessing

Preprocessing is an essential step in NLP, aimed at ensuring data quality and consistency [28]. In this research, three steps are used. The first step is identifying and removing redundant data. Special

characters, HTML tags, punctuation marks, and unnecessary whitespace are removed. The second step is removing stop words such as common articles or prepositions; these words contribute little to the overall meaning. The final step is breaking down the text into tokens, making it easier to analyze and extract meaningful information.

2.3. Word representation

In order to transform textual preprocessed tokens into a numerical representation that can be used as input for deep learning algorithms, the CBOW-Word2Vec technique is proposed for representing tokens. Word2Vec is a word embeddings approach that encodes linguistic meaning by projecting words into a high-dimensional space where their proximity reflects semantic similarity [29], [30]. CBOW is a variant of Word2Vec that predicts a word given its context [31]; CBOW models are trained by maximizing the average log probability [32].

$$\frac{1}{L}\sum_{i=1}^{L} log\ p(t_i|t_{i-l},\dots,t_{i-1},t_{i+1},\dots,t_{i+l}) \quad (1)$$

Where $\{t_{i-l},\dots,t_{i-1},t_{i+1},\dots,t_{i+l}\}$ is a set of words that represents the context of the term t_i.

In this study, the word embedding layer using AraVec was implemented, specifically the version trained on the Arabic Wikipedia corpus using the CBOW architecture. This pre-trained model provides 300-dimensional dense vectors. It was configured with a window size of 5 and a minimum term frequency of 20 during its training phase [30].

2.4. Classification task

In this step, a hybrid CNN-BiLSTM model was developed to address the Arabic readability assessment. The CNN-BiLSTM model represents a fused neural architecture that integrates the feature extraction capabilities of CNNs with the temporal modeling of BiLSTM networks. CNNs are used for local feature extraction, often capturing spatial hierarchies in the data [33]. BiLSTMs are employed for sequence modeling, taking into account both past and future context [34]. The architecture is often represented by a sequence of equations, but a simplified form for the BiLSTM cell could be $h_t = BiLSTM(h_{t-1}, x_t)$, and for the CNN layer $\acute{x} = Conv(x; W, b)$, where W and b are the filter weights and bias.

Figure 3 illustrates the architecture of the hybrid CNN-BiLSTM model, which begins with an embedding layer that transforms input data into dense vectors of fixed size. This is followed by a dropout layer to prevent overfitting by randomly setting input units to 0 during training. The model then employs two Conv1D layers, each followed by a MaxPooling layer, to extract and downsample features from the input data, capturing local dependencies. Afterward, a BiLSTM layer is used to learn dependencies from both past and future context by processing data in both directions. To enhance training stability and speed, batch normalization is subsequently employed to normalize the input distributions across each layer. Finally, a dense layer followed by an activation layer is used to make predictions based on the learned features, with the activation layer determining the output format: beginner, intermediate, and advanced.

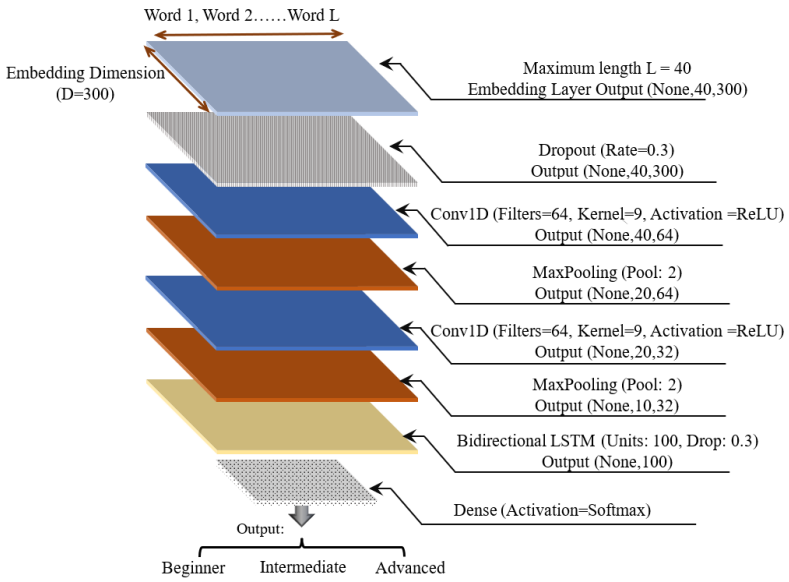

Figure 3. The layers of the hybrid CNN-BiLSTM classifier with dimensions and hyperparameters

3. RESULTS AND DISCUSSION

In this section, a detailed overview of the experimental settings and the corresponding results is provided, including descriptions of the methods used and a summary of the key findings. To assess the performance of the proposed model, standard performance metrics was employed including classification accuracy, precision, recall, F1-score, and area under the curve (AUC). The confusion also provided. These metrics are widely adopted in the field of text classification. Furthermore, to ensure a robust evaluation, a five-fold cross-validation approach was employed for all experimental procedures. Table 3 demonstrates the elements of the confusion matrix.

Table 3. Elements of the confusion matrix

	Predicted beginner	Predicted intermediate	Predicted advanced
Actual beginner	True beginner (TB)	Misclassified as intermediate	Misclassified as advanced
Actual intermediate	Misclassified as beginner	True intermediate (TI)	Misclassified as advanced
Actual advanced	Misclassified as beginner	Misclassified as intermediate	True advanced (TA)

The presented experiments were conducted on Google Collaboratory platform, on a virtual machine with T4 GPU and 54 GB of RAM. To identify optimal configurations for the proposed model, a grid search approach was employed. This research approach allowed us to systematically explore a range of parameter values and select the set that resulted in the best performance. Table 4 provides a detailed overview of the primary parameters and their respective value ranges used in this grid search.

Table 4. Set of hyperparameters used by the search grid

Hyperparameters	Values
Word embedded dimension	[100, 300]
Number of 1D convolutional layers	[1,2,3,4,5,6,7,8,9,10]
Number of epochs:	[10, 20,30]
Optimizer:	[Adam, SGD, Nadam]
Dropout rate	[0.1, 0.2, 0.3]
Learning rate	[10E-3]

Based on the grid search results, the optimal hyperparameter configuration for the model was determined. This setup employs a word embedding dimension of 300 and an architecture that includes two 1D-convolutional layers. The model is trained using the Adam optimizer for 20 epochs with a learning rate of 0.2. We reserved 20% of the dataset as an independent held-out test set. The remaining 80% was used for model training and validation using 5-fold cross-validation.

3.1. Training and validation results

In this subsection, the performance of the proposed hybrid model was examined by analyzing the progression of loss metrics over the course of the training and validation phases. Figure 4 shows the evolution of validation and training losses over a series of 20 epochs. Initially, the validation loss starts at a value of 0.686, indicating a significant error between the model's predictions and the actual targets. However, it decreases over time, reaching a low of approximately 0.158 at epoch 8, before experiencing a consistent increase, rising to 0.281 by the final epoch. The training loss follows a more consistent downward trajectory, starting from 0.429 and significantly reducing to 0.019, showcasing the model's learning and adaptation to the training data. These patterns indicate successful model training regarding the training set, highlighted by the steady decrease in loss values. However, the subsequent rise in validation loss suggests distinct overfitting. To overcome this overfitting problem, dropout techniques can be introduced, effectively improving the model's ability to generalize by preventing complex co-adaptations on training data. This method randomly disables a fraction of the neurons during the training process, which helps in making the model enhancing generalization and minimizing the risk of overfitting on specific details of the training data.

Figure 5 illustrates the progress of training and validation accuracies throughout a 20-epoch period for the hybrid deep learning model. The training accuracy commenced at 84.95%, which consistently improved, culminating in 99.29% by the end of the training. This steady increase demonstrates the model's effective learning from the training data. On the validation side, accuracy began at 66.45%, a figure that highlights the initial challenge the model faced in generalizing to unseen data. Despite some fluctuations, validation accuracy saw a notable rise, achieving highs such as 96.60% and concluding at 96.15% in the final epoch. Both training and validation accuracies improved progressively during the 20-epoch training period

highlights the model's successful adaptation and optimization, showcasing its ability to learn from the training data effectively while maintaining robust performance on unseen data.

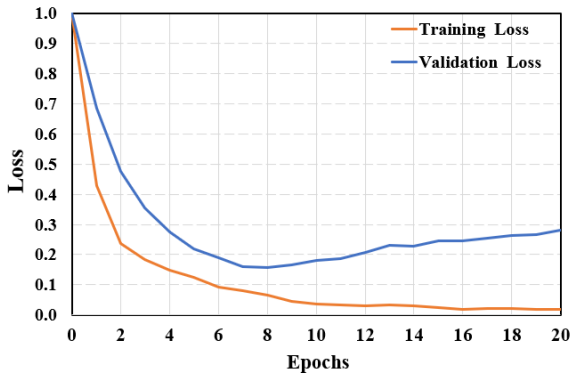

Figure 4. Validation and training loss of the proposed model

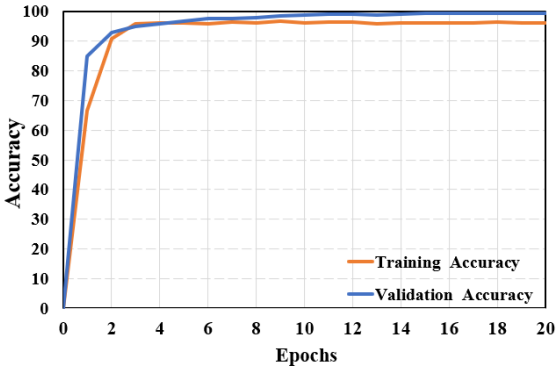

Figure 5. Training and validation accuracy of the proposed CNN-BiLSTM model

3.2. Evaluating the impact of integrating CNN layers into a BiLSTM

The impact of integrating CNN layers into a BiLSTM was examined by studying the effect of incrementally adding various numbers of Conv1D layers. This examination aimed to determine the optimal convolutional depth that enhances feature extraction while maintaining the temporal learning strengths of BiLSTM. The results illustrate the impact of Conv1D layer depth evaluated across multiple standard metrics, specifically accuracy, precision, recall, F1-score, and AUC, alongside testing execution time.

Table 5 and Figure 6 demonstrate that augmenting the number of Conv1D layers to 2 enhances performance across all evaluated metrics. Specifically, the model achieves peak values with an accuracy of 96.68%, precision of 96.70%, recall of 96.68%, and an F1-score of 96.68%. Additionally, the AUC reaches a high of 99.04%, indicating strong discriminative ability. This configuration also achieves an efficient testing execution time of 0.863 seconds, as shown in Figure 7, suggesting an improvement in feature extraction without imposing excessive computational overhead.

Table 5. Accuracy, precision, recall, F1-score, and AUC metrics across different numbers of Conv1D layers

Number of added Conv1D layers	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
1	95.58 ± 0.85	95.66 ± 0.78	95.58 ± 0.85	95.58 ± 0.86	99.09 ± 0.14
2	96.68 ± 0.25	96.70 ± 0.26	96.68 ± 0.25	96.68 ± 0.25	99.04 ± 0.15
3	96.13 ± 0.49	96.19 ± 0.44	96.13 ± 0.49	96.12 ± 0.49	98.82 ± 0.22
4	96.02 ± 0.44	96.05 ± 0.45	96.02 ± 0.44	96.02 ± 0.44	98.76 ± 0.11
5	96.28 ± 0.17	96.32 ± 0.18	96.28 ± 0.17	96.28 ± 0.17	98.91 ± 0.18
6	96.02 ± 0.52	96.07 ± 0.50	96.02 ± 0.52	96.02 ± 0.51	98.99 ± 0.11
7	95.66 ± 0.35	95.70 ± 0.34	95.66 ± 0.35	95.66 ± 0.34	99.10 ± 0.12
8	95.75 ± 0.78	95.78 ± 0.76	95.75 ± 0.78	95.75 ± 0.78	99.06 ± 0.17
9	96.13 ± 0.20	96.16 ± 0.21	96.13 ± 0.20	96.13 ± 0.20	99.22 ± 0.08
10	94.98 ± 0.55	95.06 ± 0.52	94.98 ± 0.55	94.98 ± 0.56	98.96 ± 0.26

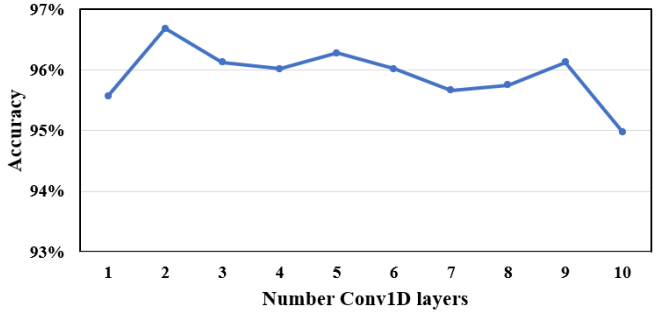

Figure 6. Accuracy variation across different numbers of Conv1D layers

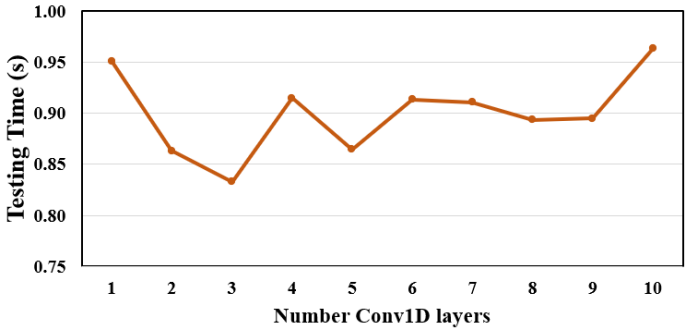

Figure 7. Testing time in seconds across different numbers of Conv1D layers

However, beyond this optimal point, further addition of layers leads to a degradation in these comprehensive metrics, suggesting overfitting. The fluctuating values with layer depths beyond 3 indicate diminishing returns from increased model complexity. For instance, at 10 layers, the accuracy drops to 94.98%, with corresponding declines in precision 95.06%, recall 94.98%, and F1-score 94.98%. Simultaneously, the computational cost rises, indicated by the testing execution time peaking at 0.963 seconds, which further suggests that deeper models may suffer from overparameterization and a failure to generalize effectively.

The findings revealed that the integration of two Conv1D layers represents the optimal configuration. This particular setup achieves a superior balance between representational capacity and generalization. This was empirically validated by the model achieving peak scores across precision, recall, and F1-score metrics while maintaining computational efficient.

3.3. Evaluating the impact of combining CNN and BiLSTM

To analyze the effect of the combination of CNN and BiLSTM, we compare the hybrid model's performance metrics and execution time with its parts: traditional CNN and BiLSTM models. Figure 8 demonstrates the superiority of the CNN-BiLSTM hybrid model, which outperforms the individual CNN and BiLSTM models across all metrics. It leads by approximately 1.75% in accuracy and 1.65% in precision compared to BiLSTM, and nearly 2.06% and 1.99% respectively over CNN, signifying a significant improvement. This advantage extends to recall and F1-score, confirming that the hybrid model excels in both precision and comprehensive data analysis, making it a more effective tool for complex classification tasks.

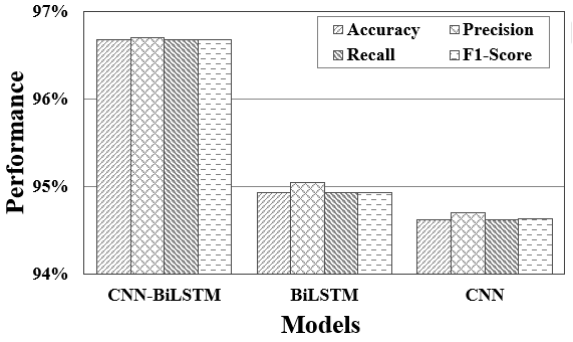

Figure 8. Comparative metrics of BiLSTM model vs. traditional CNN and BiLSTM models

Figure 9 illustrates that the CNN-BiLSTM model marks a balance between efficiency and computational time. Figure 9(a) shows that the training time for CNN-BiLSTM is 162.98 s, which is notably less than BiLSTM's 244.95 s, suggesting a more efficient training process while maintaining performance. In contrast, CNN has the shortest training time at 101.9 s but it does not provide the same level of accuracy as the hybrid model. Figure 9(b) shows that during testing CNN-BiLSTM and CNN have similar times 0.86 s and 0.9 s, but BiLSTM is significantly slower at 1.44 s. Thus, the CNN-BiLSTM hybrid offers an optimal blend of performance and training efficiency, ideal for scenarios where both factors are critical.

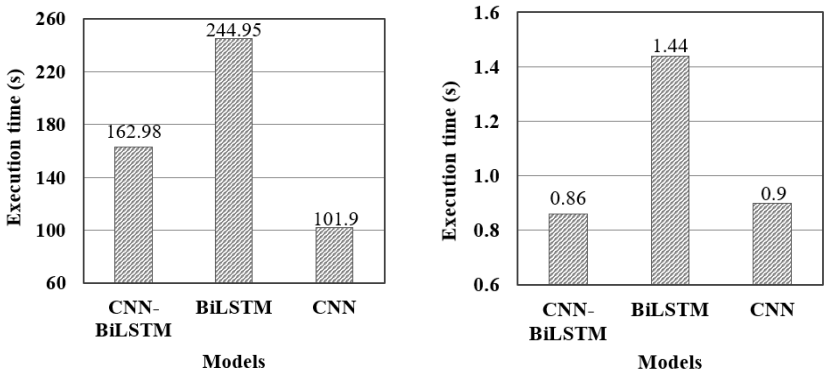

Figure 9. Execution time of CNN-BiLSTM, CNN, and BiLSTM for (a) training and time (b) testing time

Figure 10 illustrates the confusion matrices for three models, hybrid CNN-BiLSTM, CNN, and BiLSTM in classifying sentences across three difficulty levels: beginner, intermediate, and advanced. In Figure 10(a), the hybrid CNN-BiLSTM model demonstrated notable precision, correctly identifying 250 beginner sentences, 358 intermediate sentences, and 265 advanced sentences. It shows minimal misclassifications, with 5 intermediate sentences and 3 advanced sentences incorrectly classified as beginner; 1 intermediate sentence misclassified as advanced; and 9 advanced sentences wrongly identified as intermediate. In Figure 10(b), the CNN model accurately classified 250 beginner sentences but faced challenges with higher complexity levels, misclassifying 9 advanced sentences as intermediate and 16 intermediate sentences as advanced, alongside smaller errors in other categories. In Figure 10(c), the BiLSTM model showed consistent accuracy, correctly classifying 244 beginner, 350 intermediate, and 263 advanced sentences. These outcomes underscore the hybrid CNN-BiLSTM model's superior performance in accurately discerning a wide range of sentence complexities with high precision and minimal errors, marking it as particularly effective for detailed and precise classification tasks.

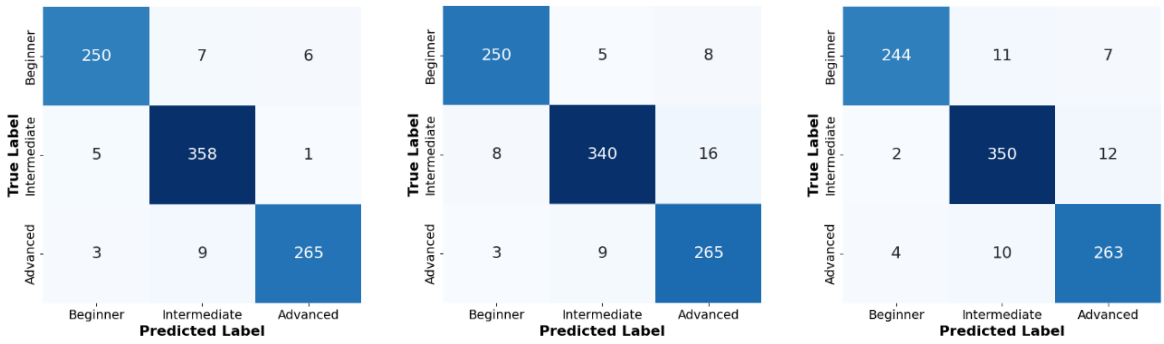

Figure 10. Confusion matrices of the three models: (a) hybrid CNN-BiLSTM, (b) CNN, and (c) BiLSTM

3.4. Discussion

In this study, a novel system for assessing the difficulty levels of Arabic sentences was introduced, employing a hybrid deep learning model. The method involves a four-step process, including data presentation, preprocessing, semantic numerical conversion using the Word2Vec technique, and sentence classification using this deep learning model. The study into the hybrid model architecture highlighted the benefits of combining CNN layers with a BiLSTM framework, focusing on finding the optimal combination for effective feature extraction. The integration of two Conv1D layers was found to be most beneficial, achieving an impressive accuracy of 96.68%, which showcases the potential of CNN and BiLSTM layers in processing complex data. The hybrid CNN-BiLSTM model's performance was compared with traditional CNN and BiLSTM models and nine traditional machine learning classifiers. Figure 11 illustrates the ROC curves and AUC scores showing the performance of the model across all target classes.

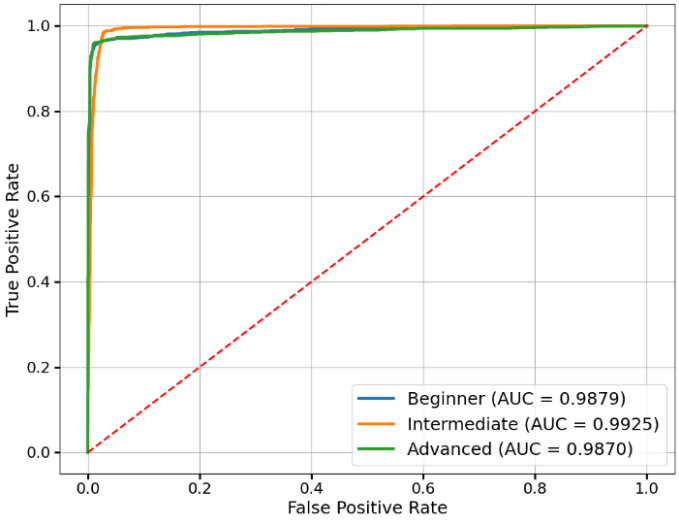

Figure 11. ROC curves and AUC scores showing the performance of the AraVec-based model across all target classes

This research underscores the advantages of hybrid deep learning models in linguistic analysis, particularly for languages with complex structures like Arabic. The outstanding performance of the CNN-BiLSTM model sets a new benchmark for sentence difficulty classification, suggesting a promising direction for future advancements in computational linguistics. Further exploration could extend the model's application to other linguistic tasks, enhancing its utility and impact in the field.

For a closer look at the error classification, Table 6 shows examples of misclassified text. In the first example, the model misclassified an intermediate sentence as advanced. Several reasons could be behind this like vocabulary and the frequent use of nouns instead of verbs (nominalization), such as 'الاقتصاد' (moderation) and 'الاكتفاء' (contentment). The sentence also includes medical term (calories, diabetes) and formal words like 'يحتم' (necessitates). The model likely considers the difficult vocabulary as a sign of the advanced class. In the second example, the level of an advanced sentence was considered as intermediate, because the of the simplicity of the sentence structure and the use of the active present-tense verbs like 'تعتبر' and 'تسهم' and common connectors like 'لكن', rather than the complex noun phrases usually found in advanced Arabic. Although the ideas are complex (cognitive faculties and deep thinking), the specific words are common ('القراءة' for reading, 'سريع' for fast). The model associates these common words with intermediate topics and misses the deeper academic context. This shows that the model relies largely on surface-level aspects; errors occur when simple concepts utilize formal language, or when complex thoughts are written explicitly.

Table 6. Examples of misclassified text

Text	Actual label	Predicted label
Example 1: ويجب الانتباه إلى أن محتواها العالي من السكر والسعرات الحرارية يحتم على مريض السكري ومن يعاني من السمنة الاقتصاد في تناولها، والاكتفاء بقطعة واحدة. Translation: Attention must be paid to its high sugar and calorie content, which requires patients with diabetes and those suffering from obesity to consume it in moderation, limiting themselves to just one piece.	Intermediate	Advanced
Example 2: لكن ماريان وولف تعتبر أن هذا النوع من التصفح السريع يمكن أن يحد تطور عمليات الفهم الأبطأ والإدراك التي تسهم في تكوين ملكات القراءة والتفكير العميق. Translation: However, Maryanne Wolf argues that this type of rapid skimming can limit the development of slower comprehension and cognitive processes, which contribute to forming the faculties of reading and deep thinking.	Advanced	Intermediate

The model relies on formal structures, largely because the word embeddings were trained on Wikipedia data. This restricts the model's ability to generalize, as the formal register used in Wikipedia does not reflect the dialects and non-standard syntax often found on social platforms. AraVec-Twitter could be

used to fix this. Since these embeddings come from tweets, they provide the dialect vocabulary the model needs to handle text outside of formal contexts.

In summary, the results presented in this paper demonstrate that the proposed model achieved remarkable accuracy. Additionally, experimental results show that the integration of the combined model based on word embedding techniques yielded a positive change in prediction performance when compared with other models presented in the literature. In Table 7, the performance of the proposed model is compared to that of similar studies examined in this paper.

Table 7. Comparison of the proposed model with recent related works

Reference	Dataset	Feature representation	Classifiers	Accuracy (%)
[23]	MoSAR corpus	BERT	CNN	77.5
[21]	Combination of global language online support system (GLOSS)-reading, GLOSS-listening, and Aljazeera-learning (AL) datasets.	Linguistics features (PoS-based frequency)	Traditional machine learning algorithms (KNN, SVM)	89.15
[27]	Aljazeera learning platform dataset	BERT	BiLSTM	89.55
Proposed method	Aljazeera learning platform dataset	Word2vec	Hybrid CNN-BiLSTM	96.68

4. CONCLUSION

The evaluation of readability is fundamentally important in education, acting as a key tool to adapt educational materials to suit the varied needs of learners. Addressing this necessity, a novel system was introduced that aims to assist educators in delivering customized and appropriate textual materials based on students' reading levels. This system measures the readability of Arabic text through the classification of sentence difficulty. The process began by collecting a dataset of Arabic sentences that have been labeled according to three levels of difficulty. Subsequently, NLP techniques are employed to extract meaningful features from the sentences. Following that a hybrid deep learning classifier is trained and compared. Experimental results demonstrate that the hybrid deep learning model achieved the highest accuracy rate of 96.68%. This research contributes to the field of Arabic text analysis by introducing an effective tool for measuring sentence readability and highlighting the importance of accurate readability assessment in Arabic language processing. The upcoming research proposes enhancing the system's performance by enlarging the dataset and implementing other word embedding techniques, along with other deep learning and transformer models.

FUNDING INFORMATION

This study was conducted without any financial support from funding agencies or organizations.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Mohamed Amine Ouassil	✓	✓	✓		✓	✓		✓	✓	✓				
Mohammed Jebbari	✓	✓		✓	✓	✓			✓	✓	✓			
Rabia Rachidi	✓	✓	✓		✓	✓		✓	✓	✓	✓			
Mouaad Errami	✓	✓	✓	✓	✓	✓			✓	✓	✓			
Soufiane Hamida	✓	✓		✓	✓					✓	✓		✓	
Bouchaib Cherradi	✓	✓		✓	✓					✓		✓	✓	
Abdelhadi Raihani	✓	✓		✓	✓					✓		✓	✓	

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Derived data supporting the findings of this study are available from the corresponding author, [BC], upon reasonable request.

REFERENCES

[1] A. Potot, L. N. Kyamko, R. R. R. -Sereño, and H. Bustrillo, "Differentiated instruction as strategy in improving reading comprehension," *Journal of English Language Teaching and Applied Linguistics*, vol. 5, no. 4, pp. 113–128, Nov. 2023, doi: 10.32996/jeltal.2023.5.4.12.

[2] M. Muammar, S. Ruqoiyyah, and N. S. Ningsih, "Implementing the teaching at the right level (TaRL) approach to improve elementary students' initial reading skills," *Journal of Languages and Language Teaching*, vol. 11, no. 4, Oct. 2023, doi: 10.33394/jollt.v11i4.8989.

[3] N. Asiza, A. Rahman, and M. Irwan, "TaRL: the potential and the challenges in learning process at the elementary school parepare," *Sang Pencerah: Jurnal Ilmiah Universitas Muhammadiyah Buton*, vol. 9, no. 2, pp. 492–500, May 2023, doi: 10.35326/pencerah.v9i2.3236.

[4] M. A. Pratama, "Improving student learning outcomes through the TaRL learning model on discussion," *Ideguru: Jurnal Karya Ilmiah Guru*, vol. 9, no. 1, pp. 53–59, Nov. 2023, doi: 10.51169/ideguru.v9i1.644.

[5] A. Binaoui, M. Moubtassime, and L. Belfakir, "The effectiveness of the TaRL approach on Moroccan pupils' mathematics, Arabic, and French reading competencies," *International Journal of Education and Management Engineering*, vol. 13, no. 3, pp. 1–10, Jun. 2023, doi: 10.5815/ijeme.2023.03.01.

[6] J. T. Guthrie and M. H. Davis, "Motivating struggling readers in middle school through an engagement model of classroom practice," *Reading & Writing Quarterly*, vol. 19, no. 1, pp. 59–85, Jan. 2003, doi: 10.1080/10573560308203.

[7] S. Amalia, S. Safrida, and S. M. Ulva, "Application of teaching at the right level (TaRL) and culturally responsive teaching (CRT) approach to increase the motivation and learning outcomes of students on the material of transport through membranes," *Jurnal Penelitian Pendidikan IPA*, vol. 10, no. 1, pp. 270–274, Jan. 2024, doi: 10.29303/jppipa.v10i1.5355.

[8] R. G. Benjamin, "Reconstructing readability: recent developments and recommendations in the analysis of text difficulty," *Educational Psychology Review*, vol. 24, no. 1, pp. 63–88, Mar. 2012, doi: 10.1007/s10648-011-9181-8.

[9] A. S. M. Selim, "Readability of reading texts for EFL students at Al-Baha University," *English Language Teaching*, vol. 16, no. 12, pp. 1–14, Nov. 2023, doi: 10.5539/elt.v16n12p1.

[10] M. Arshad, M. M. Yousaf, and S. M. Sarwar, "Comprehensive readability assessment of scientific learning resources," *IEEE Access*, vol. 11, pp. 53978–53994, 2023, doi: 10.1109/ACCESS.2023.3279360.

[11] A. Koubaa, A. Ammar, L. Ghouti, O. Necar, and S. Sibaee, "ArabianGPT: native Arabic GPT-based large language model," *Preprints*, Feb. 2024, doi: 10.20944/preprints202402.1409.v1.

[12] S. AL-Sarayreh, A. Mohamed, and K. Shaalan, "Challenges and solutions for Arabic natural language processing in social media," in *Business Intelligence and Information Technology*, Singapore: Springer, 2023, pp. 293–302, doi: 10.1007/978-981-99-3416-4_24.

[13] H. Gharaibeh *et al.*, "Arabic sentiment analysis of Monkeypox using deep neural network and optimized hyperparameters of machine learning algorithms," *Social Network Analysis and Mining*, vol. 14, no. 1, Jan. 2024, doi: 10.1007/s13278-023-01188-4.

[14] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "LLMs and NLP models in cryptocurrency sentiment analysis: a comparative classification study," *Big Data and Cognitive Computing*, vol. 8, no. 6, Jun. 2024, doi: 10.3390/bdcc8060063.

[15] S. Maqsood *et al.*, "Assessing English language sentences readability using machine learning models," *PeerJ Computer Science*, vol. 7, Jan. 2022, doi: 10.7717/peerj-cs.818.

[16] M. S. Ahmed, S. M. Maher, and M. E. Khudhur, "Arabic cyberbullying detecting using ensemble deep learning technique," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 32, no. 2, pp. 1031–1041, Nov. 2023, doi: 10.11591/ijeecs.v32.i2.pp1031-1041.

[17] M. Zulqarnain *et al.*, "Text classification using deep learning models: a comparative review," *Cloud Computing and Data Science*, vol. 5, no. 1, pp. 80–96, Oct. 2023, doi: 10.37256/ccds.5120243528.

[18] W. She, "A review of deep learning-based text sentiment analysis research," *Applied and Computational Engineering*, vol. 32, no. 1, pp. 157–164, Jan. 2024, doi: 10.54254/2755-2721/32/20230204.

[19] H. Feng, S. Hou, L.-Y. Wei, and D.-X. Zhou, "CNN models for readability of Chinese texts," *Mathematical Foundations of Computing*, vol. 5, no. 4, 2022, doi: 10.3934/mfc.2022021.

[20] N. Nassiri, A. Lakhouaja, and V. C. -Sforza, "Modern standard Arabic readability prediction," in *Arabic Language Processing: From Theory to Practice*, Cham, Switzerland: Springer, 2018, pp. 120–133, doi: 10.1007/978-3-319-73500-9_9.

[21] N. Nassiri, A. Lakhouaja, and V. C. -Sforza, "Arabic L2 readability assessment: dimensionality reduction study," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 3789–3799, Jun. 2022, doi: 10.1016/j.jksuci.2020.12.021.

[22] H. Saddiki, N. Habash, V. C. -Sforza, and M. Al Khalil, "Feature optimization for predicting readability of Arabic L1 and L2," in *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, 2018, pp. 20–29, doi: 10.18653/v1/W18-3703.

[23] S. Berrichi, N. Nassiri, A. Mazroui, and A. Lakhouaja, "Exploring the impact of deep learning techniques on evaluating Arabic L1 readability," in *Artificial Intelligence, Data Science and Applications*, Cham, Switzerland: Springer, 2024, pp. 1–7, doi: 10.1007/978-3-031-48573-2_1.

[24] S. Bessou and G. Chenni, "Efficient measuring of readability to improve documents accessibility for Arabic language learners," *Journal of Digital Information Management*, vol. 21, no. 3, pp. 75–82, Sep. 2021, doi: 10.6025/jdim/2021/19/3/75-82.

[25] H. Saddiki, K. Bouzoubaa, and V. C. -Sforza, "Text readability for Arabic as a foreign language," in *2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA)*, Nov. 2015, pp. 1–8, doi: 10.1109/AICCSA.2015.7507232.

[26] I. Baazeem, H. Al-Khalifa, and A. Al-Salman, "Cognitively driven Arabic text readability assessment using eye-tracking," *Applied Sciences*, vol. 11, no. 18, Sep. 2021, doi: 10.3390/app11188607.

[27] M. A. Ouassil, M. Jebbari, R. Rachidi, M. Errami, B. Cherradi, and A. Raihani, "Enhancing Arabic text readability assessment: a combined BERT and BiLSTM approach," in *2024 International Conference on Circuit, Systems and Communication (ICCSC)*, Jun. 2024, pp. 1–7, doi: 10.1109/ICCSC62074.2024.10616953.

[28] S. Lin, "Text emotional analysis in natural language processing," *Applied and Computational Engineering*, vol. 36, no. 1, pp. 163–172, Feb. 2024, doi: 10.54254/2755-2721/36/20230440.
[29] S. J. Johnson, M. R. Murty, and I. Navakanth, "A detailed review on word embedding techniques with emphasis on word2vec," *Multimedia Tools and Applications*, vol. 83, no. 13, pp. 37979–38007, Oct. 2023, doi: 10.1007/s11042-023-17007-z.
[30] A. B. Soliman, K. Eissa, and S. R. El-Beltagy, "AraVec: a set of Arabic word embedding models for use in Arabic NLP," *Procedia Computer Science*, vol. 117, pp. 256–265, 2017, doi: 10.1016/j.procs.2017.10.117.
[31] Q. Luo, W. Xu, and J. Guo, "A study on the CBOW model's overfitting and stability," in *Proceedings of the 5th International Workshop on Web-scale Knowledge Representation Retrieval & Reasoning*, Nov. 2014, pp. 9–12, doi: 10.1145/2663792.2663793.
[32] M. Al-Sarem, A. Alsaeedi, F. Saeed, W. Boulila, and O. AmeerBakhsh, "A novel hybrid deep learning model for detecting COVID-19-related rumors on social media based on LSTM and concatenated parallel CNNs," *Applied Sciences*, vol. 11, no. 17, Aug. 2021, doi: 10.3390/app11177940.
[33] M. Lotfy and G. Soliman, "CNN-optimized text recognition with binary embeddings for Arabic expiry date recognition," *Journal of Electrical Systems and Information Technology*, vol. 11, no. 1, Feb. 2024, doi: 10.1186/s43067-024-00136-2.
[34] M. M. A. Busst, K. S. M. Anbananthen, S. Kannan, J. Krishnan, and S. Subbiah, "Ensemble BiLSTM: a novel approach for aspect extraction from online text," *IEEE Access*, vol. 12, pp. 3528–3539, 2024, doi: 10.1109/ACCESS.2023.3349203.

BIOGRAPHIES OF AUTHORS

Mohamed Amine Ouassil received the B.Sc. in Mathematics and Computer Science Engineering in 2009 from the Faculty of Science and Technology of Beni Mellal, Morocco, and the M.Sc. degree in Data Science and Big Data in 2021 from High National School for Computer Science and Systems Analysis (ENSIAS) of Rabat, Morocco. He is currently working as guidance counsellor at National Ministry of Education Morocco. He is also a Ph.D. candidate at Electrical Engineering and Intelligent Systems (EEIS) Laboratory in ENSET Mohammedia, Hassan II University of Casablanca (UH2C). His research interests reside in the fields of machine learning, artificial intelligence, and natural language processing. He can be contacted at email: ouassil.amine@gmail.com.

Mohammed Jebbari is Ph.D. candidate enrolled at the Electrical Engineering and Intelligent Systems (EEIS) Laboratory in ENSET of Mohammedia, Hassan II University of Casablanca. He received the master's degree in Multimedia Pedagogical Engineering from the higher normal school of Tetouan, Abdelmalek Essaadi University in 2016. Currently serving as a computer science professor at the Ministry of National Education of Morocco. His research focuses on the development and enhancement of machine learning models for intelligent systems, with the aim of detecting learning difficulties in learners and integrating various aspects of the learner's model, such as predominant learning styles, predicting performance and emotions. He has published several research articles in these areas and continues to actively contribute to the development of new knowledge in these fields. He can be contacted at email: mohamed.jb@outlook.sa.

Rabia Rachidi obtained her master's degree in Business Intelligence and Big Data Analytics from the Faculty of Science at Chouaib Doukkali University, El Jadida, Morocco, in 2022. She is currently pursuing a Ph.D. at the same university. Her research interests include artificial intelligence, machine learning, deep learning, and natural language processing. She is also a Ph.D. candidate in the Optimization Research Emerging Systems Networks and Imaging Laboratory (LAROSERI) at the Faculty of Science, Chouaib Doukkali University. She can be contacted at email: rachidirabia99@gmail.com.

Mouaad Errami received his engineering degree in Data Engineering and Data Science from the National Institute of Statistics and Applied Economics of Rabat in 2020. Currently, he is working as a systems engineer at Rabat. He is also a Ph.D. candidate at Electrical Engineering and Intelligent Systems (EEIS) Laboratory in ENSET Mohammedia, Hassan II University of Casablanca (UH2C). Highly interested in the world of machine learning and artificial intelligence, his articles delve heavily into the realm of NLP such as detecting fake news and sentiment analysis; is also interested in developing this domain when it comes to Arabic language. He can be contacted at email: mouaad.errami@gmail.com.

Soufiane Hamida is a 29-year-old researcher from Rabat, Morocco, is highly knowledgeable in the field of machine learning methodologies for pattern recognition, as evidenced by his Ph.D. degree. He further honed his skills and expertise in the field of educational technology by obtaining his master's degree from the higher normal School of Tetouan, Abdelmalek Essaadi University in 2017. Currently, he is actively contributing to the advancement of research at the Electrical Engineering and Intelligent Systems Research Laboratory at Hassan II University of Casablanca, Morocco. Furthermore, he is making significant efforts towards furthering research at the GENIUS Laboratory at SupMTI in Rabat, Morocco. He can be contacted at email: hamida.93s@gmail.com.

Bouchaib Cherradi received the B.S. degree in Electronics in 1990 and the M.S. degree in Applied Electronics in 1994 from the ENSET Institute of Mohammedia, Morocco. He received the DESA diploma in Instrumentation of Measure and Control (IMC) from Chouaib Doukkali University at El Jadida in 2004. He received his Ph.D. in Electronics and Image Processing from the Faculty of Science and Technology, Mohammedia. His works as an associate professor in CRMEF-El Jadida. In addition, he is associate researcher member of Electrical Engineering and Intelligent Systems (EEIS) Laboratory in ENSET of Mohammedia, Hassan II University of Casablanca (UH2C), and LaROSERI Laboratory on leave from the Faculty of Science, Chouaib Doukali University, El Jadida, Morocco. He is a supervisor of several Ph.D. students. He can be contacted at email: bouchaib.cherradi@gmail.com.

Abdelhadi Raihani was appointed as a professor in Electronics Engineering at Hassan II University of Casablanca, ENSET Institute, Mohammedia Morocco since 1991. He received the B.S. degree in Applied Electronics in 1991 from the ENSET Institute. He has his DEA diploma in Information Processing from Ben M'sik University of Casablanca in 1994. He received the Ph.D. in Parallel Architectures Application and Image Processing from the Ain Chock University of Casablanca in 1998. He currently serves as a full professor in the Department of Electrical Engineering at ENSET of Mohammedia. His research topics are various and large in multiple domains of medical image processing and electrical engineering, energy management systems, power and energy systems control and smart grids. He has published over than 74 papers in distinguished scientific journals and more than 60 papers in international conferences. He supervised and delivered more than 15 Ph.D. theses. He worked closely in national research programs with IRESEN under the grant "Green INNO Project/UPISREE". He can be contacted at email: raihani@enset-media.ac.ma.