

Deep Learning for Arabic Question Classification: Leveraging BERT and Hybrid Neural Networks

Somia Khedimi¹, Abdelghani Bouziane², Djelloul Bouchiha³

^{1,2,3}Department of Computer Science, Institute of Science, University centre of Naama, Algeria
EEDIS Lab., UDL, Sidi Bel Abbes, Algeria

Article Info

Article history:

Received Feb 7, 2025

Revised Nov 2, 2025

Accepted Nov 16, 2025

Keyword:

Arabic question classification
Natural Language Processing (NLP)

Deep learning

CNN

BiLSTM

Pre-trained language models

ABSTRACT

This paper presents a deep learning approach for Arabic question classification, leveraging the strengths of pre-trained language models and advanced neural network architectures to address the unique challenges of Arabic text processing. The proposed methodology employs BERT and Word2Vec to generate contextualized and semantic-rich representations of Arabic questions, effectively capturing their linguistic intricacies and morphological complexity. These embeddings are fed into a hybrid classification framework combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks, enabling the extraction of both spatial and sequential features from the input. Experimental results demonstrate the model's effectiveness, achieving an Accuracy of 85.12%, along with high Precision, Recall, and F1-score metrics. These findings highlight the potential of integrating pre-trained Arabic-specific language models with hybrid deep learning architectures, providing a robust solution for Arabic question classification. This work contributes to advancing Arabic natural language processing, offering a strong foundation for the development of high-performance question-answering systems and related applications.

Copyright © 2025 Institute of Advanced Engineering and Science.
All rights reserved.

Corresponding Author:

Somia Khedimi
Department of Computer Science,
Institute of Science, University centre of Naama, Algeria
Email: khedimi@cuniv-naama.dz

1. INTRODUCTION

A Question Answering System (QAS) is a key application in the field of natural language processing (NLP) that aims to provide accurate answers to natural language queries posed by users. A typical QAS consists of three core modules: question processing, document retrieval, and answer extraction. The question processing module plays a pivotal role in formulating a structured query passed to the document retrieval module. Within this module, question classification is a critical task that directly impacts the overall accuracy of the QAS [1]. Question Classification (QC) involves categorizing a question into one of several predefined classes based on a specific taxonomy [1, 2]. Two widely used taxonomies are: Li and Roth's taxonomy [3], which classifies questions into six coarse-grained classes and 50 fine-grained subclasses; and Bloom's taxonomy [4], which organizes assessment questions into six cognitive levels: Remember, Understand, Apply, Analyze, Evaluate, and Create.

Moreover, QC is considered necessary in QAS because an accurate question classification helps narrow down the search space, thereby improving response time and enhancing the precision of a QAS [5, 6]. For example, correctly identifying the question class for "في أي قارة تقع الجزائر؟" (Translating to "In which continent is Algeria located?") as *Location* would direct the QAS to prioritize passages containing geographical information.

However, despite the remarkable progress in NLP, primarily for English due to the application of deep learning techniques, particularly Long Short-Term Memory (LSTM) networks [7] and transformer-based

language models, classifying questions remains a challenging task, especially for languages with limited resources, such as Arabic, which is a Semitic language spoken by over 420 million people [8]. Arabic's linguistic features, such as the use of diacritics that affect both pronunciation and meaning, and its complex morphology, which enables single words to convey entire sentences (e.g., "فستعلمون," meaning "you will know"), significantly complicate text processing. Furthermore, the absence of capitalization, unlike in many other languages, complicates the identification of proper nouns and sentence boundaries. Adding to these difficulties is the scarcity of high-quality tools and datasets for Arabic, which poses a major obstacle to developing robust NLP systems.

To address these issues, this paper presents a deep learning-based approach for Arabic question classification, adopting Li and Roth's taxonomy as the classification framework. The proposed method harnesses the strengths of pre-trained language models and advanced neural network architectures to address the complexities inherent in the rich morphology and linguistic diversity of the Arabic language. Using an Arabic dataset, the model first generates numerical representations of the input text through either BERT (ArabicBERT/AraBERT) or Word2Vec (AraVec). These representations are then processed by a hybrid classification module combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory networks (BiLSTM) to classify the text. A comparative analysis of different model combinations reveals that the integration of ArabicBERT with CNN-BiLSTM achieves the best performance.

The remainder of this paper is organized as follows: Section 2 reviews the related work in question classification, while Section 3 presents the components of the proposed approach, detailing the dataset, representation module, and classification module. Section 4 provides the experimental results and evaluates the performance of the model. Finally, Section 5 concludes the study and outlines recommendations for future research.

2. RELATED WORKS

Research on question classification has advanced significantly, employing deep learning and machine learning techniques to address challenges across different languages. Below, we summarize key studies in this domain.

Aiming at classifying Turkish questions, [9] developed a model using Word2Vec embeddings with Continuous Bag of Words (CBOW) and Skip-gram techniques. They combined these embeddings with CNN, LSTM, and GRU architectures. The model was trained on a Turkish question dataset derived from an English dataset. The best performance was achieved using CNN with Skip-gram, attaining an accuracy of 93.7%.

In a separate study, [6] focused on classifying Indonesian questions using the BiLSTM algorithm and one-hot encoding for word representations. Through comparative experiments, they demonstrated that BiLSTM outperformed LSTM and RNN models, achieving an accuracy of 90%. This study highlights the effectiveness of sequential modeling for question classification tasks.

The Chinese questions were classified by [12], who utilized ERNIE embeddings to classify Chinese questions. They developed two classifiers: highway-CNN and a hybrid model integrating DCU (Didactic Composite Units), Highway Networks, and BiLSTM. Tested on the NLPCC-QA 2018 and FUDAN datasets, the models achieved accuracies of 97% and 83%, respectively.

In another study, [14] proposed a machine learning-based Arabic question classification model using TF-IDF, Bag of Words, and Named Entity Recognition for feature extraction. Logistic Regression outperformed other classifiers, achieving 82% accuracy on the NSURL 2019 Kaggle competition dataset, highlighting its robustness for Arabic text classification. In a different research, [1] introduced two deep learning-based approaches for Arabic question classification. The first used embeddings from AraBERT, BERT, or ELMo, while the second combined these embeddings with Word2Vec. A *Softmax* layer was employed for classification. Using a dataset of 3173 questions collected from TREC, CLEF, and school textbooks, AraBERT achieved the highest accuracy of 94.2%.

In the medical domain, [10] built a multilabel classification model for Arabic COVID-19-related questions using DeBERTa for word embeddings and BiLSTM for classification. The dataset was sourced from the ATTIBI FAQ website. Despite the complexity of the task, the proposed model achieved an accuracy of 71%, showcasing the potential of transformer-based embeddings for Arabic text classification. In a similar study, [15] created two datasets for Arabic medical question classification: an unlabelled dataset for domain-specific Word2Vec embeddings and a labelled dataset for training and testing. Their proposed model, using BiLSTM, achieved a classification accuracy of 87.2%, outperforming LSTM.

[11] concerned with the agricultural question leveraging BERT, where they introduced a QC model for Chinese agricultural queries by combining BERT embeddings with a Deep Pyramid Convolutional Neural Network (DPCNN). Using a dataset of 100284 question-answer pairs from Chinese agricultural websites, the model achieved an outstanding accuracy of 99.07%, outperforming BiLSTM and TextCNN architectures.

In the educational domain, [13] created an Arabic dataset for assessment questions based on Bloom's taxonomy. They proposed a modified TF-IDF feature extraction method that assigns higher weights to question tools. For classification, algorithms such as SVM, KNN, Decision Trees, Naive Bayes, and Logistic Regression were tested, with Naive Bayes achieving the best accuracy of 78%. Similarly, [2] combined a modified TF-IDF approach with ArELMo embeddings to classify Arabic assessment questions according to Bloom's taxonomy. The classification module integrated BiGRU, CNN, and Attention layers. Trained on the dataset introduced by Alammary [13], the model achieved an accuracy of 84.26%.

A summary of the reviewed studies is presented in Table 1, which includes the languages, datasets used, embeddings, classifiers, and accuracies.

Table 1. Summary of question classification research

Reference	Language	Dataset	Embedding	Classifier	Accuracy
[9]	Turkish	Turkish dataset based on English Q&A	Word2Vec (CBOW, Skip-gram)	CNN, LSTM, GRU	93.7%
[6]	Indonesian	Indonesian question dataset	One-hot encoding	BiLSTM, LSTM, RNN	90%
[10]	Arabic	ATTIBI FAQ dataset	DeBERTa	BiLSTM	71%
[11]	Chinese	Chinese agricultural websites	BERT	DPCNN	99.07%
[12]	Chinese	NLPCC-QA 2018, FUDAN	ERNIE	BiLSTM, Highway-CNN	97%
[13]	Arabic	Arabic assessment questions dataset	Modified TF-IDF	SVM, KNN, DT, NB, LG	78%
[14]	Arabic	NSURL 2019 Kaggle competition Dataset by Alammary (2021)	TF-IDF	Logistic Regression	82%
[2]	Arabic	Arabic assessment questions dataset	ArELMo	BiGRU, CNN, Attention	84.26%
[15]	Arabic	ALTTIBI society datasets	Word2Vec	LSTM, BiLSTM	87.2%
[1]	Arabic	TREC, CLEF, school books dataset	AraBERT, ELMo, Word2Vec	Softmax Layer	94.2%

3. PROPOSED MODEL ARCHITECTURE

The proposed model for Arabic question classification integrates multiple components designed to handle the unique linguistic features of the Arabic language. As shown in Figure 1, the classification process relies on three primary components: the dataset, the representation module, and the classification module. The representation module generates embeddings using either BERT-based models (ArabicBERT or AraBERT) or Word2Vec (AraVec), while the classification module employs CNN and/or BiLSTM neural networks to classify the text. Below, we describe each component in detail and outline the general architecture of the model.

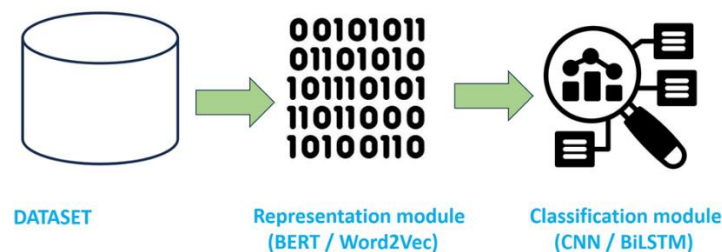


Figure 1. General architecture of the proposed Arabic question classification model

3.1. Dataset

The dataset used in this study is derived from the English question dataset created by UIUC, which is categorized according to the Li and Roth taxonomy [3]. The original dataset consists of a training set of 5452 questions and a test set of 500 questions. To adapt this dataset for Arabic, all questions were translated into

Arabic, resulting in a combined dataset of 5952 questions. The questions are hierarchically classified into six coarse-grained classes: ABBREVIATION (ABBR), ENTITY (ENTY), DESCRIPTION (DESC), HUMAN (HUM), LOCATION (LOC), and NUMERIC (NUM), which are further subdivided into 50 fine-grained classes.

Figure 2 presents the distribution of the coarse classes, while Table 2 illustrates a sample of the dataset. This hierarchical structure allows the model to address both general and specific classification tasks, making it suitable for a wide range of question-answering applications in Arabic.

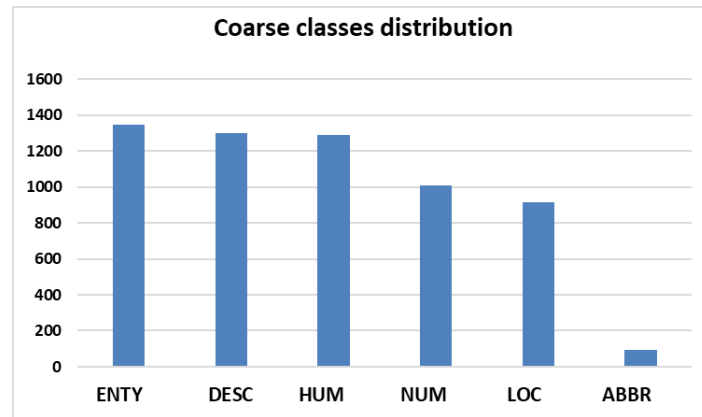


Figure 2. Distribution of coarse classes in the used dataset.

Table 2. A sample from the dataset.

Class	Question
LOC:other	ما هو أكبر مبنى مكاتب في العالم؟
HUM:ind	من الذي حول كل ما لمسه إلى ذهب؟
ENTY:dismed	ما هو المصطلح الطبي لسرطان الدم؟
HUM:ind	من يلعب في فريق البيسبول تشارلي براون؟

3.2. Representation module

The representation module is responsible for encoding the Arabic text into dense, numerical vectors that capture semantic and contextual information. Two main approaches are used for this purpose:

3.2.1. BERT-based models

Bidirectional Encoder Representations from Transformers (BERT) is a state-of-the-art transformer model introduced by Devlin, Chang [16]. It uses a multi-layer bidirectional transformer encoder to process token sequences. BERT applies two pretraining tasks:

- *Masked Language Model (MLM)*: Predicts masked tokens in a sequence to generate contextual representations.
- *Next Sentence Prediction (NSP)*: Determines whether one sentence logically follows another, enabling the model to capture sentence relationships.

For Arabic text, several fine-tuned BERT-based models are utilized:

- *ArabicBERT*: Trained on 8.2 billion words from Standard Modern Arabic and dialectal Arabic, it supports multiple model sizes (mini, small, base, large) and excels in various NLP tasks [17].
- *AraBERT*: Pretrained on approximately 24 GB of Arabic text, AraBERT incorporates Farasa for segmentation and SentencePiece for tokenization. It performs well in tasks like question answering, named entity recognition, and sentiment analysis [18].
- *Additional models*: Other BERT variants, such as JABER, SABER, ARBERT, and MARBERT, are also optimized for Arabic text processing [19, 20].

3.2.2. Word2Vec-based models

Word2Vec [21] is a neural network model that generates dense, continuous word embeddings by learning word relationships from large corpora. It uses two architectures:

- *Continuous Bag-of-Words (CBOW)*: Predicts a target word based on its surrounding context.
- *Skip-Gram (SG)*: Predicts the context based on a target word, favoring rare words.

For Arabic, AraVec [22] is an adaptation of Word2Vec. It is trained on Arabic corpora from Twitter, Arabic Wikipedia, and other sources, using both CBOW and SG architectures to generate high-quality embeddings.

3.3. Classification module

The classification module processes the embeddings generated by the representation module and assigns questions to their respective classes. It combines two neural network architectures:

3.3.1. Bidirectional Long Short-Term Memory (BiLSTM)

BiLSTM [23], an advanced variant of LSTM networks [7], captures long-term dependencies in sequential data by processing it in both forward and backward directions. This bidirectional capability enhances the model's understanding of context, making BiLSTM highly effective for natural language processing tasks.

3.3.2. Convolutional Neural Networks (CNN)

CNNs are powerful architectures originally designed for image processing, but have proven effective for text classification. They consist of convolutional layers to extract features, pooling layers for dimensionality reduction, and fully connected layers for classification. CNNs are valued for their ability to capture local patterns in data and their parameter efficiency, making them suitable for analyzing n-grams in text.

By combining CNN and BiLSTM, the proposed model leverages the strengths of both architectures. CNN extracts local spatial features, while BiLSTM captures long-range dependencies, resulting in a robust classification framework for Arabic questions.

4. EXPERIMENTS AND RESULTS

To evaluate the performance of the CNN-BiLSTM model for Arabic question classification, a series of experiments was conducted. These experiments aimed to assess the effectiveness of the proposed fusion model compared to standalone CNN and BiLSTM architectures. Each of these models was tested with three different embedding techniques: AraVec, AraBERT, and ArabicBERT.

The dataset was split into three subsets: 70% for training, 20% for validation, and 10% for testing. Preprocessing steps included normalization and the removal of diacritics, numbers, and punctuation. However, stop words were retained due to their critical role in question classification tasks. After preprocessing, the questions were tokenized and encoded using the selected embedding models, resulting in the following vector dimensions:

- *AraBERT* and *ArabicBERT* produced embeddings with a shape of (15, 768), where 15 represents the maximum sequence length.
- *AraVec* generated embeddings with dimensions of (15, 100).

The resulting embeddings were fed into the CNN-BiLSTM model, whose architecture is detailed below and visualized in Figure 3:

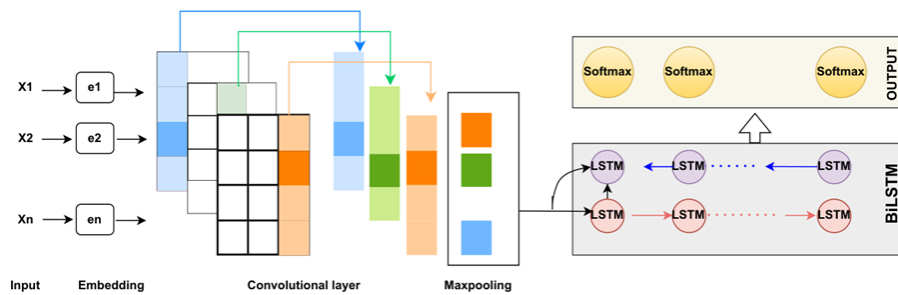


Figure 3. Architecture of the CNN-BiLSTM model.

- *Convolutional layer*: Extracts local features from the input sequences using a set of filters.
- *Max-pooling layer*: Reduces the dimensionality of the feature maps by sliding a window over the input and selecting the maximum value in each region.
- *BiLSTM layer*: Captures bidirectional context, allowing the model to understand dependencies from both past and future words in the sequence.
- *Dropout layer*: Mitigates overfitting by randomly dropping a fraction of the layer's activations during training.
- *Fully connected dense layer*: Outputs a probability distribution over the 50 fine-grained classes by applying a *Softmax* activation function.

To evaluate the performance of our deep learning model for Arabic question classification, we employed four widely used evaluation metrics: Accuracy, Precision, Recall, and F1-score. These metrics allow

for a detailed assessment of the model's ability to classify Arabic questions into the correct hierarchical classes. Below are the definitions of these metrics in the context of Arabic question classification:

Accuracy: Represents the proportion of Arabic questions that were correctly classified into their respective coarse or fine-grained classes out of the total number of questions. For example, if the model correctly identifies that "في أي قارة تقع الجزائر؟" (Translating to "In which continent is Algeria located?") belongs to the *LOCATION* class, this contributes to the model's accuracy.

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

TP (True Positive) is the number of Arabic questions that the model correctly classifies into their actual class, *TN* (True Negative) is the number of Arabic questions that the model correctly identifies as not belonging to a given class, *FP* (False Positive) is the number of Arabic questions that the model incorrectly classifies into a specific class when they actually belong to a different class, and *FN* (False Negative) is the number of Arabic questions that the model fails to classify into their correct class, instead assigning them to a different class.

Precision: Measures the proportion of Arabic questions correctly classified into a specific class (e.g., *ENTITY* or *NUMERIC*) among all questions predicted to belong to that class. For instance, if the model predicts a question as belonging to the *ENTITY* class, Precision evaluates how many of these predictions are correct.

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

Recall: Assesses the model's ability to identify all Arabic questions that truly belong to a specific class. For example, Recall evaluates how well the model identifies all questions that should be classified under the *HUMAN* class, regardless of incorrect predictions in other classes.

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

F1-score: Combines Precision and Recall to provide a balanced measure of the model's performance, especially when there is an uneven class distribution. For instance, if the model achieves high Recall but relatively low Precision for the *DESCRIPTION* class, the F1-score provides a single metric reflecting this balance.

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

These metrics were calculated for all models and embedding combinations tested in this study. They provide an in-depth understanding of the models' strengths and weaknesses, enabling us to assess their overall performance and suitability for Arabic question classification tasks.

Tables 3,4, and 5 present a detailed comparison of the models based on key evaluation metrics. This analysis demonstrates the effectiveness of combining ArabicBERT embeddings with the CNN-BiLSTM architecture for Arabic question classification tasks.

Table 3. Evaluation of the Aravec model

Model	Precision	F1-score	Recall
CNN	67,98	68,39	70,44
BiLSTM	70.41	71,68	74.42
CNN-BiLSTM	75,37	74,34	76,10

Table 4. Evaluation of the AraBERT model

Model	Precision	F1score	Recall
CNN	81,79	80,66	80,92
BiLSTM	81.34	80.62	81.13
CNN-BiLSTM	83,32	81,88	81,76

Table 5. Evaluation of the ArabicBERT model

Model	Precision	F1score	Recall
CNN	84.28	82.93	83.43
BiLSTM	84.87	83.15	83.22
CNN-BiLSTM	86.32	85.17	85.11

The experimental results highlight the performance of the CNN-BiLSTM model compared to standalone CNN and BiLSTM models. The proposed CNN-BiLSTM consistently achieved the best performance across all evaluation metrics when paired with any of the three embedding models. Among the combinations, the ArabicBERT-CNN-BiLSTM model delivered the highest performance, achieving an Accuracy of 85.12%, a Precision of 86.32%, an F1-score of 85.17%, and a Recall of 85.11%.

The results are further visualized in the histogram shown in Figure 4, which illustrates the Accuracy achieved by each model.

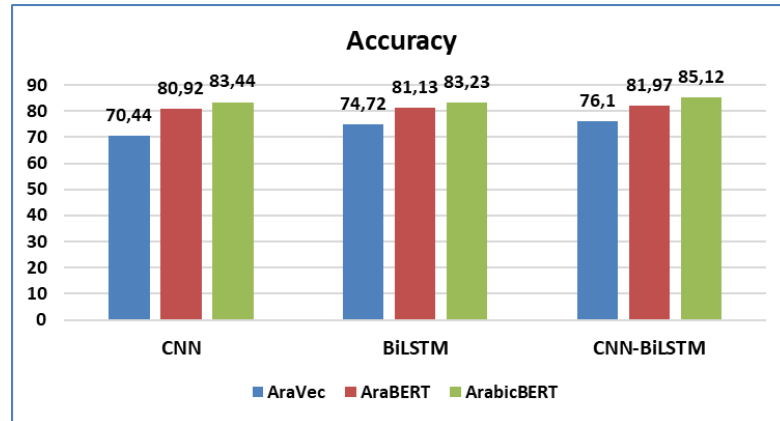


Figure 4. Histogram illustrating test Accuracy for each model.

Table 6 presents the classification report for the test set, which comprises 477 examples, using the ArabicBERT-CNN-BiLSTM model. The unbalanced distribution of classes in the dataset impacts the overall model performance. For instance, in the entire dataset, there are only 11 examples for both ENTY: symbol and ENTY:instru classes. Consequently, the model fails to predict examples for these classes, which negatively impacts the model's effectiveness.

Table 6. classification report of test set.

Class	Precision	Recall	F1-score	Class	Precision	Recall	F1-score
ABBR:abb	100%	100%	100%	HUM:desc	100%	100%	100%
ABBR:exp	60%	50%	55%	HUM:gr	80%	50%	62%
DESC:def	88%	82%	85%	HUM:ind	89%	91%	90%
DESC:desc	69%	78%	73%	HUM:title	100%	50%	67%
DESC:manner	100%	95%	98%	LOC:city	86%	100%	92%
DESC:reason	92%	69%	79%	LOC:country	92%	85%	88%
ENTY:animal	89%	80%	84%	LOC:mount	100%	100%	100%
ENTY:body	100%	100%	100%	LOC:other	97%	90%	94%
ENTY:color	100%	100%	100%	LOC:state	86%	100%	92%
ENTY:cremat	94%	88%	91%	NUM: code	100%	100%	100%
ENTY:currency	100%	100%	100%	NUM: count	93%	93%	93%
ENTY:dismed	86%	75%	80%	NUM:date	100%	100%	100%
ENTY:event	57%	80%	67%	NUM:dist	100%	100%	100%
ENTY:food	73%	89%	80%	NUM:money	86%	100%	92%
ENTY:instru	0%	0%	0%	NUM:other	80%	80%	80%
ENTY:lang	100%	100%	100%	NUM:perc	100%	100%	100%
ENTY:letter	50%	100%	67%	NUM:period	88%	100%	93%
ENTY:other	37%	56%	44%	NUM: speed	100%	100%	100%
ENTY:plant	100%	100%	100%	NUM:temp	100%	100%	100%
ENTY:product	60%	75%	67%	NUM:volsize	100%	100%	100%
ENTY:sport	83%	100%	91%	NUM:weight	100%	100%	100%
ENTY:substance	100%	60%	75%	ENTY:termeq	67%	75%	71%
ENTY:symbol	0%	0%	0%	ENTY:veh	100%	50%	67%
ENTY:techmeth	75%	100%	86%	ENTY:word	100%	100%	100%
Weighted average precision		86%					
Weighted average recall		85%					
Weighted average F1-score		85%					

5. CONCLUSION AND PERSPECTIVE

This study presented a deep learning-based approach for Arabic question classification, leveraging the strengths of pre-trained language models and advanced neural network architectures. The proposed methodology utilized Arabic-specific pre-trained models, including AraVec, AraBERT, and ArabicBERT, to

generate contextualized embeddings that capture the rich linguistic and morphological features of the Arabic language. These embeddings were combined with a fusion of Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) networks to classify questions hierarchically according to the widely recognized Li and Roth taxonomy [3].

Experimental results validated the effectiveness of the proposed approach in two significant aspects. First, BERT-based models, particularly ArabicBERT, demonstrated a superior ability to comprehend the nuances of Arabic, highlighting the importance of pretraining on diverse and linguistically rich corpora. Second, the fusion of CNN and BiLSTM architectures proved to be highly effective in extracting both local and sequential features, enabling accurate classification of questions into coarse and fine-grained classes. This approach achieved promising results, with the ArabicBERT-CNN-BiLSTM model delivering the highest performance across all evaluation metrics, including Accuracy, Precision, Recall, and F1-score. These findings indicate that integrating advanced embedding techniques with hybrid deep learning models has the potential to significantly enhance Arabic question-answering systems.

Looking ahead, several avenues for future research are identified. Expanding the dataset with more diverse and domain-specific Arabic questions can enhance the model's ability to generalize and perform well across varied contexts. Incorporating colloquial Arabic and regional dialects could further improve the system's robustness and practical applicability. Additionally, future studies could focus on optimizing model hyperparameters, such as sequence length, learning rates, and dropout rates, to achieve better performance. Exploring advanced architectures, such as transformers and attention mechanisms, may yield further improvements in accuracy and efficiency. The practical implementation of the proposed approach in real-world Arabic question-answering systems would provide valuable insights into its utility and help refine its design. Moreover, extending the model to handle multilingual scenarios, including code-switched Arabic that combines Arabic with other languages, presents an exciting direction for expanding its capabilities. Together, these efforts can push the boundaries of Arabic natural language processing and contribute to the development of more robust, versatile question-answering systems.

REFERENCES

- [1] Hamza, A., "Embedding arabic questions by feature-level fusion of word representations for questions classification: It is worth doing?", *Computer and Information Sciences*, 2022.
- [2] Balla, H., M.L., Salvador, and S.J. Delany, "Arabic Question Classification using Deep Learning", in *CCRIS'22: 2022 3rd International Conference on Control, Robotics and Intelligent System*, 2022.
- [3] Li, X. and D. Roth, "Learning question classifiers", in the *19th international conference , Association for Computational Linguistics*, 2002
- [4] Bloom, B.S., et al., "Taxonomy of educational objectives: The classification of educational goals", Vol. 1, David McKay Company, INC. USA, EDWARDS BROS., ANN ARBOR, MICHIGAN, 1956.
- [5] Loni, B., "A Survey of State-of-the-Art Methods on Question Classification", 2011.
- [6] Anhar, R., T.B. Adji, and N. Akhmad Setiawan, "Question Classification on Question-Answer System using Bidirectional-LSTM", in *2019 5th International Conference on Science and Technology (ICST), IEEE*, 2019.
- [7] Hochreiter, S. and J. Schmidhuber, "Long short-term memory", *Neural computation*, vol. 9, no. 8, 1997, pp. 1735-1780.
- [8] Al-dihaymawee, D.T.M., A.A. Merzah, and H.M. Abdul Ridha, "The Story of Arabic Language: Historical Linguistics Study", *Tasnim International Journal for Human, Social and Legal Sciences*, vol. 3, No. 1, 2024, pp. 572-582.
- [9] Zulqarnain, M., et al., "A comparative analysis on question classification task based on deep learning approaches", *PeerJ Computer Science*, vol. 7, 2021, pp. e570.
- [10] Al-Smadi, B.S., "DeBERTa-BiLSTM: A multi-label classification model of Arabic medical questions using pre-trained models and deep learning", *Computers in Biology and Medicine*, vol. 170, 2024, pp. 107921.
- [11] Guo, X., et al., "Efficient Agricultural Question Classification With a BERT-Enhanced DPCNN Model", *IEEE Access*, vol. 12, 2024, pp. 109255-109268.
- [12] Liu, G., et al., "Chinese Question Classification Based on ERNIE and Feature Fusion", *Springer International Publishing*, 2020.
- [13] Alammary, A.S., "Arabic Questions Classification Using Modified TF-IDF", *IEEE Access*, vol 9, 2021, pp. 95109-95122.
- [14] Malkawi, R., S. Alsrahan, and A.A. Saifan, "Arabic questions classification machine learning algorithms", *Journal of Theoretical and Applied Information Technology* , Vol. 100, no. 20, 2022.
- [15] Faris, H., et al., "Classification of Arabic healthcare questions based on word embeddings learned from massive consultations: a deep learning approach", *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 4, 2022. pp. 1811-1827.
- [16] Mikolov, T., et al., "Efficient Estimation of Word Representations in Vector Space", *arXiv*, 2013.
- [17] Devlin, J., et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *arXiv*, 2019.

- [18] Safaya, A., M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media", in *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, International Committee for Computational Linguistics, 2020.
- [19] Antoun, W., F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding", *arXiv*, 2021.
- [20] Ghaddar, A., et al., "JABER and SABER: Junior and Senior Arabic BERT", *arXiv*, 2022.
- [21] Abdul-Mageed, M., A. Elmadany, and E.M.B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic", in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, vol. 1, Association for Computational Linguistics, 2021.
- [22] Soliman, A.B., 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, 2017, pp. 256-265.
- [23] Schuster, M. and K.K. Paliwal, Bidirectional recurrent neural networks, *IEEE transactions on Signal Processing*, vol. 45, no. 11, 1997, pp. 2673-2681.

BIOGRAPHIES OF AUTHORS



Somia KHEDIMI is a teacher of computer science in a Secondary high school. She received her Master's degree in Computer Science from the University Centre of Naama in 2018. She has been pursuing a Ph.D. in Computer Science at the same university since 2023. Her research focuses on artificial intelligence and its applications, particularly in the field of natural language processing. She can be contacted at email: khedimi@cuniv-naama.dz.



Abdelghani BOUZIANE is an Associate Professor of Computer Science at the University Centre of Naama, specializing in Artificial Intelligence (AI), Machine Learning, Deep Learning, and Natural Language Processing. He earned his Ph.D. in Computer Science from the University of Sidi BelAbb'es, Algeria, in 2019. Since 2012, he has been actively involved in integrating AI technologies into both research and teaching. At the University Centre of Naama, he serves as the Head of Field in the Computer Science Department and Chief of Doctoral Training in AI. His research primarily focuses on AI and Natural Language Processing. He can be contacted at email: bouziane@cuniv-naama.dz.



Djelloul BOUCHIHA received his Engineer degree in Computer Science from Sidi Bel Abbes University, Algeria, in 2002, followed by an M.Sc. in Computer Science from the same university in 2005, and a Ph.D. in 2011. From 2005 to 2010, he served as a Lecturer in the Department of Computer Science at Saida University, Algeria. In September 2013, he was appointed Associate Professor at the University Centre of Naama, Algeria, and has been a Full Professor since January 2018. His research interests include the Semantic Web, Software Engineering, Knowledge Management, Natural Language Processing, Artificial Intelligence, and Machine Learning. He can be contacted at email: bouchiha@cuniv-naama.dz.