Deep Learning for Arabic Question Classification: Leveraging AraELECTRA and CNNs

Abstract. This paper presents a deep learning approach for Arabic question classification, leveraging the capabilities of the AraELECTRA language model to generate word representations and a convolutional neural network for classification. The dataset employed in this study, translated from English to Arabic, categorizes questions following the Li and Roth taxonomy. The results demonstrate an accuracy of 85.32%, showcasing the potential of this approach to contribute significantly to the development of Arabic question-answering systems and support researchers in improving their accuracy.

Keywords: Natural Language Processing (NLP), Arabic question classification, question-answering systems, AraELECTRA, convolutional neural networks (CNN), deep learning, Li and Roth taxonomy.

1 Introduction

Question classification (QC) is a critical component of question analysis in question-answering systems (QAS). A QAS processes a user's query and generates an appropriate response, and QC plays a vital role in this process by assigning questions to predefined categories [1]. This classification step significantly enhances the accuracy of QAS by narrowing the search space and enabling a better understanding of the user's intent, which ultimately leads to more precise responses [2].

Questions can be categorized based on specific taxonomies. For example, Bloom's taxonomy [3] organizes educational questions according to cognitive levels such as remembering, understanding, and applying. On the other hand, Li and Roth's taxonomy [4]commonly used in open-domain question classification, categorizes questions into types such as number, human, location, description, and abbreviation. Traditionally, machine learning algorithms like Support Vector Machines (SVM) and Decision Trees have been employed for QC tasks, as demonstrated in studies by [5] and [6].

Recently, deep learning methods, particularly Long Short-Term Memory (LSTM) [7] networks and transformer-based language models, have driven significant advancements in Natural Language Processing (NLP), especially for English. However, extending these advancements to Arabic, a Semitic language spoken by over 420 million people [8], presents unique challenges. Arabic's linguistic characteristics, such as the use of diacritics, symbols that influence pronunciation and meaning, and its rich morphology, which allows single words to encode complete sentences (e.g., "فستعلمون" meaning "you will know"), make processing Arabic text particularly complex. Additionally, the absence of capitalization, unlike in many other languages, makes identifying proper nouns and sentence boundaries more difficult. Furthermore, the limited availability of high-quality tools and datasets for Arabic significantly hinders the development of robust NLP systems.

To address these challenges, this paper introduces a deep learning-based approach to Arabic question classification. The proposed method leverages AraELECTRA [9] , a state-of-the-art pretrained language model for Arabic, alongside Convolutional Neural Networks (CNN), which are well-suited for classification tasks. The combination of these techniques aims to improve the accuracy and robustness of Arabic question classification, contributing to advancements in Arabic NLP and QAS.

The remainder of this paper is organized as follows. Section 2 reviews existing research on question classification. Section 3 presents the proposed approach. Section 4 details the conducted experiments and the results obtained. Finally, Section 5 concludes the study and provides recommendations for future work.

2 Related work

Several studies have addressed question classification across various domains and languages, utilizing diverse machine learning and deep learning approaches.

Dake et al. (2023) [10] developed an automatic question classification module aimed at assisting instructors in managing increasing student interactions on e-learning platforms. They employed machine learning algorithms such as Support Vector Machines (SVM), Random Forest (RF), and Decision Trees (DT). For feature extraction, they used N-gram models (Unigram, Bigram, Trigram) with weighting schemes including Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), and Term Presence (TP). Their dataset, comprising 1096 questions categorized into seven classes, achieved the highest performance with the AdaBoost SVM ensemble algorithm, obtaining an accuracy of 78.55% and an F1-score of 0.782 using Unigram features.

Mao et al. (2021) [11]proposed a novel approach for classifying Chinese medical questions by integrating semantic features extracted through Long Short-Term Memory (LSTM) networks and topic features derived from Latent Dirichlet Allocation (LDA). These features were further refined by a Convolutional Neural Network (CNN). The study collected questions from Chinese websites and classified them into seven topics. Their LSTM-CNN model achieved an impressive F1-score of 99.01%.

Mohammed and Omar (2020) [12]introduced a method for classifying educational questions according to Bloom's Taxonomy cognitive levels. Their approach combined Term Frequency-Part of Speech Inverse Document Frequency (TFPOS-IDF) with Word2Vec for vector representations. Two datasets of labeled open-ended questions were utilized, and classification was performed using SVM, Logistic Regression (LR), and K-Nearest Neighbors (KNN). The best results were achieved by the SVM algorithm, which obtained an F1-score of 89.7%.

Mohasseb and Kanavos (2023) [13] proposed a grammar-based framework for question classification called GQCC, which classifies questions as factoid or non-factoid. The study employed ensemble learning methods, including bagging and boosting, which combine classifiers such as SVM, RF, KNN, DT, and Naive Bayes (NB). The best performance was achieved by a bagged Decision Tree, with an accuracy of 89%.

Faris et al. (2022) [14] developed a medical question classification model using Word2Vec-based word embeddings trained on an unlabelled dataset from the ALTTIBI website. The embeddings were then classified into 15 medical specialties using LSTM and BiLSTM neural networks. The proposed model achieved test accuracies of 87.1% with LSTM and 87.2% with BiLSTM.

Al-Smadi (2024) [15] created an Arabic dataset composed of COVID-19-related questions collected from the ALTTIBI website. The study developed a DeBERTa-BiLSTM model for question classification, achieving an accuracy of 71%.

Malkawi et al. (2022) [6] utilized TF-IDF for feature extraction on an Arabic question dataset provided by the NSURL 2019 Kaggle competition. Classification was performed using Naive Bayes (NB), SVM, and Logistic Regression (LR) algorithms. The LR algorithm outperformed SVM and NB, achieving an accuracy of 82%.

Table 1 provides a summary of the reviewed research, highlighting the language, feature extraction methods, classifiers, and results.

Table 1. Summary of related work on question classification

Contribution	Language	Embeddings	Classifier	Evaluation
[10]	English	N-gram, TF-IDF, TP	SVM, RF, DT	Accuracy=78.55%
[11]	Chinese	LSTM- CNN	LSTM-CNN	F1-score =99.01%
[15]	Arabic	DeBERTa	BiLSTM	Accuracy=71%
[12]	English	TFPOS-IDF,word2vec	SVM, LR, KNN	F1-score= 89.7%
[13]	English	GQCC	SVM, RF, KNN, DT	Accuracy=89%
[6]	Arabic	TF-IDF	SVM, NB, LG	Accuracy=82%
[14]	Arabic	Word2Vec	LSTM, BiLSTM	Accuracy=87.2%

3 Model overview

The goal of this study is to design an efficient model for Arabic question classification. This involves leveraging advanced language models and classification algorithms. Specifically, we utilize the AraELECTRA language model [9] for word embeddings and Convolutional Neural Networks (CNN) for classification. Comparative experiments are conducted to evaluate the performance of this approach against traditional methods, such as Term Frequency-Inverse Document Frequency (TF-IDF) combined with Support Vector Machines (SVM), as well as the AraVec language model [16]. Next, we describe the key components of our proposed model:

3.1 Language models and feature extraction

Feature extraction is an essential step in question classification. Traditional methods like TF-IDF measure the importance of terms within a document relative to a collection of documents. TF-IDF is calculated as the product of Term Frequency (TF), which

counts a word's occurrence in a document, and Inverse Document Frequency (IDF), which down-weights common words across documents. In contrast, modern language models like AraELECTRA and AraVec offer a more sophisticated way of representing textual data. AraELECTRA is an Arabic adaptation of the ELECTRA model, which employs a unique Replaced Token Detection (RTD) strategy. This model is trained on 77 GB of Arabic text, enabling it to learn meaningful contextual representations. AraVec, on the other hand, is a Word2Vec-based language model that uses either Continuous Bag-of-Words (CBOW) or Skip-Gram (SG) architectures and is trained on large Arabic corpora, including texts from Twitter, Wikipedia, and the Web.

3.2 Classification algorithms:

The classification component of the model plays a crucial role in mapping feature representations to specific categories. CNNs are employed for this purpose, as they are highly effective in processing sequential or grid-structured data like text. A CNN typically comprises convolutional layers for extracting local features, pooling layers for dimensionality reduction, dropout layers to prevent overfitting, and fully connected layers for classification. In addition, SVM is used as a baseline classifier[17]. This algorithm constructs an optimal hyperplane to separate data points into different classes. For non-linear problems, SVM uses kernel functions like linear, polynomial, and Radial Basis Function (RBF) to transform data into higher-dimensional spaces where linear separation becomes feasible.

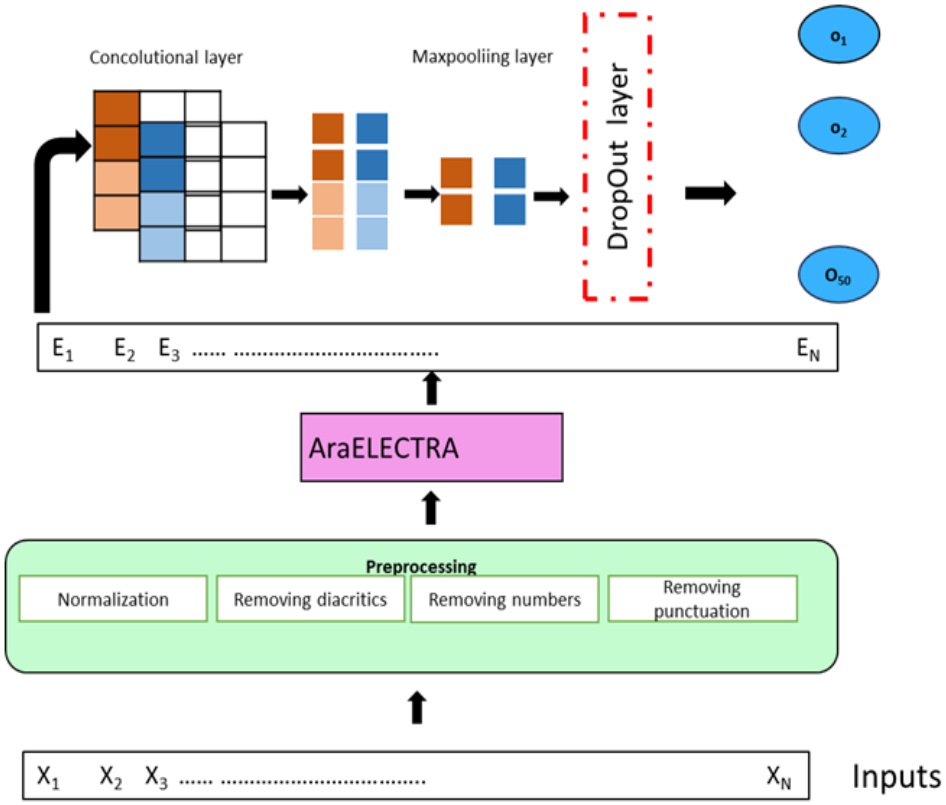
3.3 Dataset:

The scarcity of Arabic datasets for question classification necessitated the use of a translated version of the English UIUC[1] dataset. This dataset includes a training set of 5452 questions and a test set of 500 questions. The questions are categorized according to Li and Roth's taxonomy [4] which defines six coarse classes (ENTITY [ENTY], ABBREVIATION [ABBR], HUMAN [HUM], DESCRIPTION [DESC], NUMERIC [NUM], and LOCATION [LOC]) and fifty fine-grained classes.

The proposed model processes questions through a pipeline designed to optimize classification accuracy. As depicted in Figure 1, the process begins with normalization, where diacritics, numbers, and punctuation are removed. Question tools are retained, as they are crucial for identifying question types. Preprocessed questions are then fed into the AraELECTRA model to generate embeddings of size 15768, where 15 represents the maximum sequence length, and 768 is the dimensionality of the embedding. These embeddings are subsequently passed through a CNN composed of two convolutional layers with 100 and 80 filters, respectively. A max-pooling layer follows, which reduces the dimensionality by selecting the maximum value from each region. A dropout layer with a rate of 0.5 is applied to prevent overfitting, and a dense layer with a Softmax activation function produces the final classification.

[1] https://cogcomp.seas.upenn.edu/Data/QA/QC/

Fig. 1. Overall architecture of the proposed model.



4 Experiments and results

This section presents the experiments conducted and the corresponding results.

In the first experiment, we used AraELECTRA in combination with a convolutional neural network (CNN). The architecture of this model is illustrated in Figure 1. This model outperformed all other models, achieving an accuracy of 85.35%, an F1-score of 85.25%, precision of 86.17%, and recall of 83.35%.

For the second experiment, we replaced the CNN with a Support Vector Machine (SVM). The SVM model received word embedding vectors generated by AraELECTRA and employed a linear kernel to find the best hyperplane. This configuration yielded an accuracy of 64.82%.

In the third experiment, we used the Aravec language model, which encodes the input questions into vectors of shape (100, 15), where 100 represents the embedding dimension and 15 is the maximum sequence length. These vectors were fed into both a CNN and an SVM. The Aravec-CNN model performed better than the AraVec-SVM model, achieving an accuracy of 76.52%.

For the final experiment, we utilized TF-IDF for feature extraction, setting the maximum number of features to 300. The resulting vectors, of shape (300,), were processed once by the CNN and once by the SVM. In this experiment, SVM outperformed CNN, achieving an accuracy of 65.07%, while CNN reached an accuracy of 63.10%.

Tables 1,2 and 3 summarize the results of each experiment, and Figure 2 illustrates the test accuracy for each model.

Table 2. Evaluation metrics TF-IDF models

Model	Precision	F1 score	Recall
SVM	65%	63%	66%
CNN	61.21%	60.97%	63.1%

Table 3. Evaluation metrics for Aravec models

Model	Precision	F1 score	Recall
SVM	75%	73%	73%
CNN	76.37%	75.64%	76.51%

Table 4. Evaluation metrics for AraELECTRA models

Model	Precision	F1 score	Recall
SVM	65%	64%	65%
CNN	86.17%	85.25%	85.32%

The results clearly demonstrate that the combination of deep learning techniques, particularly AraELECTRA with CNN, is the most effective approach for question classification. As shown in Table 2 and Figure 2, the AraELECTRA pretrained language model surpasses both the Aravec and TF-IDF models, owing to its ability to generate contextualized word representations. In contrast, both the AraVec and TF-IDF models generate static word vectors. Furthermore, CNN outperforms SVM when combined with AraELECTRA (vector shape: (15,768)) and AraVec (vector shape: (15,100)). However, SVM is more effective when combined with TF-IDF (vector shape: (300,)). These findings highlight the power of deep neural networks in handling high-dimensional data.

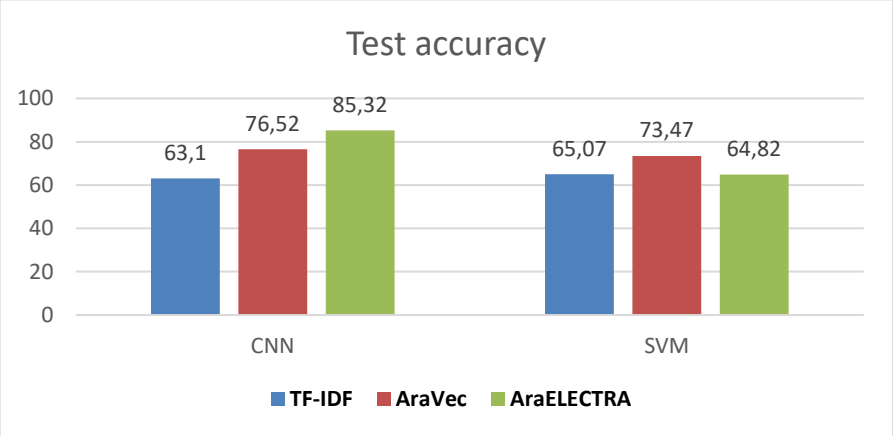


Fig. 2. Test accuracy for each model

5 Conclusion and perspectives

In this study, we proposed a novel approach for classifying Arabic questions using advanced deep learning techniques, which represent the current state-of-the-art in natural language processing. The approach leverages the AraELECTRA language model for word embeddings, demonstrating superior performance compared to the AraVec language model, which is based on machine learning, and TF-IDF, a statistical feature extraction method. AraELECTRA's ability to generate contextualized word representations played a significant role in enhancing classification accuracy.

Additionally, convolutional neural networks (CNNs) proved highly effective in classification tasks, achieving the best performance and surpassing support vector machines (SVMs) in most scenarios. This highlights the capability of CNNs to handle complex, high-dimensional representations such as those produced by the AraELECTRA model.

The findings of this study are promising for advancing Arabic text classification, particularly in the domain of question classification. These results pave the way for further developments in Arabic natural language processing applications, including question-answering systems, educational tools, and automated customer support solutions. Future work could explore optimizing model parameters, expanding datasets to include domain-specific questions, and integrating the proposed approach into broader question-answering pipelines.

References

1. 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. ACM.
2. 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)*. 2019. IEEE.
3. Bloom, B.S., et al., *Taxonomy of Educational.* Objectives: Handbook I-Cognitive Domain/BS Bloom, MD Englehart, EJ Furst, WH Hill, DR Krathwohl. New York: David McKay Co Inc, 1956.
4. Li, X. and D. Roth. *Learning question classifiers*. in *the 19th international conference*. 2002. Association for Computational Linguistics.
5. Alammary, A.S., *Arabic Questions Classification Using Modified TF-IDF*. IEEE Access, 2021. **9**: p. 95109-95122.
6. MALKAWI, R., S. ALSRAHAN, and A.A. SAIFAN, *ARABIC QUESTIONS CLASSIFICATION MACHINE LEARNING ALGORITIMS.* Journal of Theoretical and Applied Information Technology, 2022. **100**(20).
7. Hochreiter, S., *Long Short-term Memory.* Neural Computation MIT-Press, 1997.
8. AL-dihaymawee, D.T.M., A.A. Merzah, and H.M.A. Ridha, *The Story of Arabic Language: Historical Linguistics Study.* Tasnim International Journal for Human, Social and Legal Sciences, 2024. **3**(1): p. 572-582.
9. Antoun, W., F. Baly, and H. Hajj, *AraELECTRA: Pre-Training Text Discriminators for Arabic Language Understanding*. 2021, arXiv.
10. Dake, D.K., et al., *Instructor-assisted question classification system using machine learning algorithms with N-gram and weighting schemes.* Discover Artificial Intelligence, 2023. **3**(1): p. 29.
11. Mao, S., L.-L. Zhang, and Z.-G. Guan, *An LSTM&Topic-CNN Model for Classification of Online Chinese Medical Questions.* IEEE Access, 2021. **9**: p. 52580-52589.
12. Mohammed, M. and N. Omar, *Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF*. International Journal on Advanced Science, Engineering and Information Technology, 2018. **8**(4-2): p. 1679-1685.
13. Mohasseb, A. and A. Kanavos, *Grammar-Based Question Classification Using Ensemble Learning Algorithms*, in *Web Information Systems and Technologies*, M. Marchiori, F.J. Domínguez Mayo, and J. Filipe, Editors. 2023, Springer Nature Switzerland: Cham. p. 84-97.
14. 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, 2022. **13**(4): p. 1811-1827.
15. 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, 2024. **170**: p. 107921.
16. 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, 2017. **117**: p. 256-265.

17. Chapelle, O., P. Haffner, and V.N. Vapnik, *Support vector machines for histogram-based image classification.* IEEE Transactions on Neural Networks, 1999. **10**(5): p. 1055-1064.