Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 12, No. 3, September 2024, pp. 690∼699
ISSN: 2089-3272, DOI: 10.52549/ijeei.v12i3.5659

BERT-BiLSTM model for hierarchical Arabic text classification

Benamar HAMZAOUI[1,*], Djelloul BOUCHIHA[1,*], Abdelghani BOUZIANE[1,*], Noureddine DOUMI[2,*]
[1]Department of Computer Science, Institute of Science, University centre of Naama, Algeria
[2]Department of Computer Science, University of Saida, Algeria
[*]EEDIS Lab., UDL, Sidi Bel Abbes, Algeria

Article Info

Article history:

Received Jun 19, 2024
Revised Sep 19, 2024
Accepted Sep 26, 2024

Keywords:

Text classification
Hierarchical classification
BERT (Bidirectional Encoder Representations from Transformers)
Deep learning
BiLSTM (Bidirectional Long Short-Term Memory)
Arabic language

ABSTRACT

Text classification is a fundamental task in natural language processing (NLP) aimed at categorizing text documents into predefined categories or labels. Leveraging artificial intelligence (AI) tools, particularly deep learning and machine learning, has significantly enhanced text classification capabilities. However, for the Arabic language, which lacks comprehensive resources in this domain, the challenge is even more pronounced. Hierarchical text classification, which organizes categories into a tree-like structure, presents added complexity due to inter-category similarities and connections across different levels. In addressing this challenge, we propose a deep learning model based on BERT (Bidirectional Encoder Representations from Transformers) and BiLSTM (Bidirectional Long Short-Term Memory). Experimental evaluations demonstrate the effectiveness of our approach compared to existing methods, yielding promising results. Our study contributes to advancing text classification methodologies, particularly in the context of Arabic language processing.

Corresponding Author:

Djelloul BOUCHIHA
Department of Computer Science, Institute of Sciences, University centre -SALHI Ahmed-Naama, Algeria
Email: bouchiha@cuniv-naama.dz

1. INTRODUCTION

Artificial intelligence (AI) has revolutionized numerous domains by enabling machines to perform tasks that traditionally required human intelligence [1]. Within AI, machine learning and deep learning stand out as powerful methodologies driving innovation across various applications. In particular, natural language processing (NLP) tasks have witnessed significant advancements, with text classification being a cornerstone application.

Text classification is the process of assigning textual data to specific categories or labels based on its content. This technique is widely used in various applications, including sentiment analysis, spam filtering, and topic identification. Its importance spans diverse domains, including information retrieval, document organization, and automated decision-making systems.

Motivated by the vast Arabic-speaking community, our focus lies in addressing the challenges of text classification for the Arabic language. Despite the sizable population that communicates in Arabic, the language still faces a shortage of resources, particularly in the realm of hierarchical datasets necessary for complex classification tasks.

Hierarchical classification presents unique challenges compared to flat classification, primarily due to the inherent similarities between categories and sub-categories. The intricate relationships within hierarchical structures necessitate specialized approaches to effectively capture the nuances of classification.

In this paper, we propose a deep-learning-based solution tailored to hierarchical Arabic text classification. Our methodology begins with the acquisition and preprocessing of an Arabic dataset, which is subsequently divided into training and testing sets. The classification pipeline leverages state-of-the-art techniques, starting with the application of BERT for text preprocessing and encoding, followed by the utilization of a BiLSTM model for classification.

Evaluation results demonstrate the efficacy of our approach, with promising performance metrics indicating its potential for real-world applications. Furthermore, a comparative study highlights the superiority of deep learning models over traditional machine learning approaches in the context of hierarchical Arabic text classification.

The rest of this paper is organized as follows: Section 2 gives an overview of related work in the field of text classification and Arabic language processing. Section 3 details our proposed deep-learning-based model architecture, including the BERT and BiLSTM components. Section 4 presents the experimental setup, dataset, evaluation results, and a comparative analysis with existing methods. Finally, Section 5 concludes the paper and outlines avenues for future research.

2. RELATED WORK

Various approaches have been proposed for Arabic text classification, using a range of classification methods on different datasets. Below is a brief review of the main approaches from the past six years:

Sundus et al. proposed a supervised feed-forward model, utilizing a machine learning technique for logistic regression [2]. Their study demonstrated significant improvements in classification accuracy and training time compared to traditional logistic regression methods.

Galal et al. focused on using convolutional neural networks (CNN) for Arabic text classification, which excelled in various natural language processing (NLP) tasks [3]. They developed a novel algorithm that incorporated additional Arabic letters and word embedding distances to cluster related Arabic words effectively.

El-Alami et al. proposed a technique based on Bag-of-Concepts and deep Autoencoder representations for Arabic text classification [4]. They utilized the Chi-Square method to recognize the most informative features and integrated explicit semantics from Arabic WordNet.

Alhawarat and Aseeri suggested a Superior Arabic Text Categorization Deep Model, which is a new framework using a Multi-Kernel CNN with N-gram word embedding [5].

Almuzaini and Azmi examined how different stemming approaches impact classification accuracy in large Arabic texts using deep learning [6]. They used their model to classify Arabic news writings from the Saudi Press Organization.

Aljedani et al. proposed a hierarchical multi-label Arabic text categorization (HMATC) model incorporating machine learning techniques [7]. They optimized the Hierarchy Of Multilabel classifiER (HOMER) algorithm by exploring various combinations of multi-label classifiers, clustering techniques, and cluster numbers to enhance hierarchical classification.

El Rifai et al. automatically analyzed tagged news articles using vocabulary features, testing deep learning approaches on two datasets [8].

Alsukhni aimed to create an Arabic news display for users, employing Multilayer Perceptron (MLP) and Recurrent Neural Network (RNN) with to develop two models in Python [9].

Elghannam conducted a project in two phases, automatically assigning multiple labels to hotel reviews based on highlighted aspects [10]. The data instances were tagged as multi-label using collected seed key clusters.

Rifai's research aimed to automatically tag input documents with multiple tags based on vocabulary features [11]. A large dataset was compiled and made publicly available, and classifiers were wrapped in a OneVsRest classifier for this purpose.

Bouchiha et al. implemented and compared various classifiers using different feature extraction techniques (e.g., Doc2Vec, BoW) and machine learning algorithms (e.g., Decision tree, Naïve Bayes) to determine the best Arabic text classification results, using the DataSet for Arabic Classification (DSAC) [12].

Bdeir and Ibrahim proposed a framework for the multi-label classification of Arabic tweets, employing word embedding techniques and CNN/RNN algorithms [13]. Their dataset included Arabic tweets collected from Twitter.

The next table summarizes the mentioned Arabic text classification approaches:

Table 1. Summary of Arabic text classification approaches

Reference	Text classification technique	Dataset
[3]	CNNNorm, CNNGStem	Arabic news dataset
[2]	Feed-forward neural network model and Logistic Regression	Khaleej-2004 and an inhouse corpus from online Arabic news
[5]	Word embedding, CNN multi-kernel, and n-gram	15 public datasets
[4]	Restricted Boltzmann Machines autoencoder, MLP and SVM	OSAC corpus
[6]	Stemming strategies, deep learning approaches, and word embedding techniques	Arabic News Texts and Saudi Press Agency
[7]	HMATC model with a machine learning technique	Raw dataset describing the Islamic field
[8]	Shallow and deep learning algorithms	Two datasets were created from Arabic news portals
[9]	Recurrent Neural Network with Long Short-Term Memory, and Multilayer Perceptron	Mowjaz Multitopic Labelling Task Competition
[10]	Various multi-label classification learning approaches	Hotel reviews
[11]	Logistic Regression XGBoost, and deep learning techniques	Multi-tagged articles from various Arabic news portals
[12]	Feature extraction techniques with Machine learning algorithms	DataSet for Arabic Classification (DSAC)
[13]	Word embedding techniques and CNN/RNN algorithms	Arabic tweets collected from Twitter

Other surveys on Arabic text classification, such as those by Wahdan et al. [14] and Abdulghani and Abdullah [15], are available. However, these surveys do not focus on dataset levels (flat and hierarchical).

Most previous works have not been tested using hierarchical datasets or evaluated with metrics dedicated to hierarchical structures. To address these limitations, we proposed a BERT-BiLSTM model tested on a hierarchical Arabic dataset and evaluated with a metric that considers hierarchical dependencies between categories.

3. PROPOSED MODEL

This section describes the proposed BERT-BiLSTM model for hierarchical Arabic text classification. It begins by introducing BERT and BiLSTM, followed by an explanation of the general architecture of the proposed model.

3.1. BERT

BERT, or Bidirectional Encoder Representations from Transformers [16], is a Transformer Encoder stack trained to provide numerical representations of text in the form of vectors. It consists of two main stages: Preprocessing and Encoding.

In BERT, the preprocessing stage prepares Arabic text for the encoding process. This involves tokenizing the text into sub-word units using WordPiece tokenization, which helps handle complex and rare words by breaking them down into meaningful components. Special tokens like [CLS] and [SEP] are added to mark the start and end of sentences, while padding ensures all sequences are of equal length. An attention mask is generated to distinguish real tokens from padding, and in some cases, normalization is applied to standardize characters or remove unnecessary diacritics in the Arabic text.

In the encoding stage, BERT uses a transformer architecture to generate dense vector representations of the text. It processes the input bidirectionally, allowing each word to be understood in the context of the entire sequence, which is crucial for Arabic's rich morphology and complex sentence structures. BERT creates contextualized embeddings for each token, meaning that even the same word will have different representations depending on its context. These embeddings are then used in downstream tasks like text classification, where the [CLS] token's final state typically serves as the representation for classification.

3.2. BiLSTM

Long Short-Term Memory (LSTM) networks excel in sequence prediction tasks by effectively capturing long-term dependencies within data sequences [17], excel in sequence prediction tasks by effectively capturing prolonged dependencies within data sequences. BiLSTMs, an extension of LSTM networks introduced by Schuster and Paliwal [18], enhance LSTM's ability to capture long-range dependencies by processing input sequences in both forward and backward directions.

BiLSTM has been used to address several problems, such as detecting Plant Disease Relations (PDR) [19] and classifying families of protein sequences [20]. In this paper, we leverage the power of BiLSTM for hierarchical classification of Arabic texts.

3.3. General architecture

To classify hierarchical Arabic texts, we propose a deep learning solution integrating BERT and BiLSTM. The classification process comprises two phases: Training and Evaluation. The training phase involves learning the model to classify text into corresponding categories, while the evaluation phase assesses the trained model's performance.

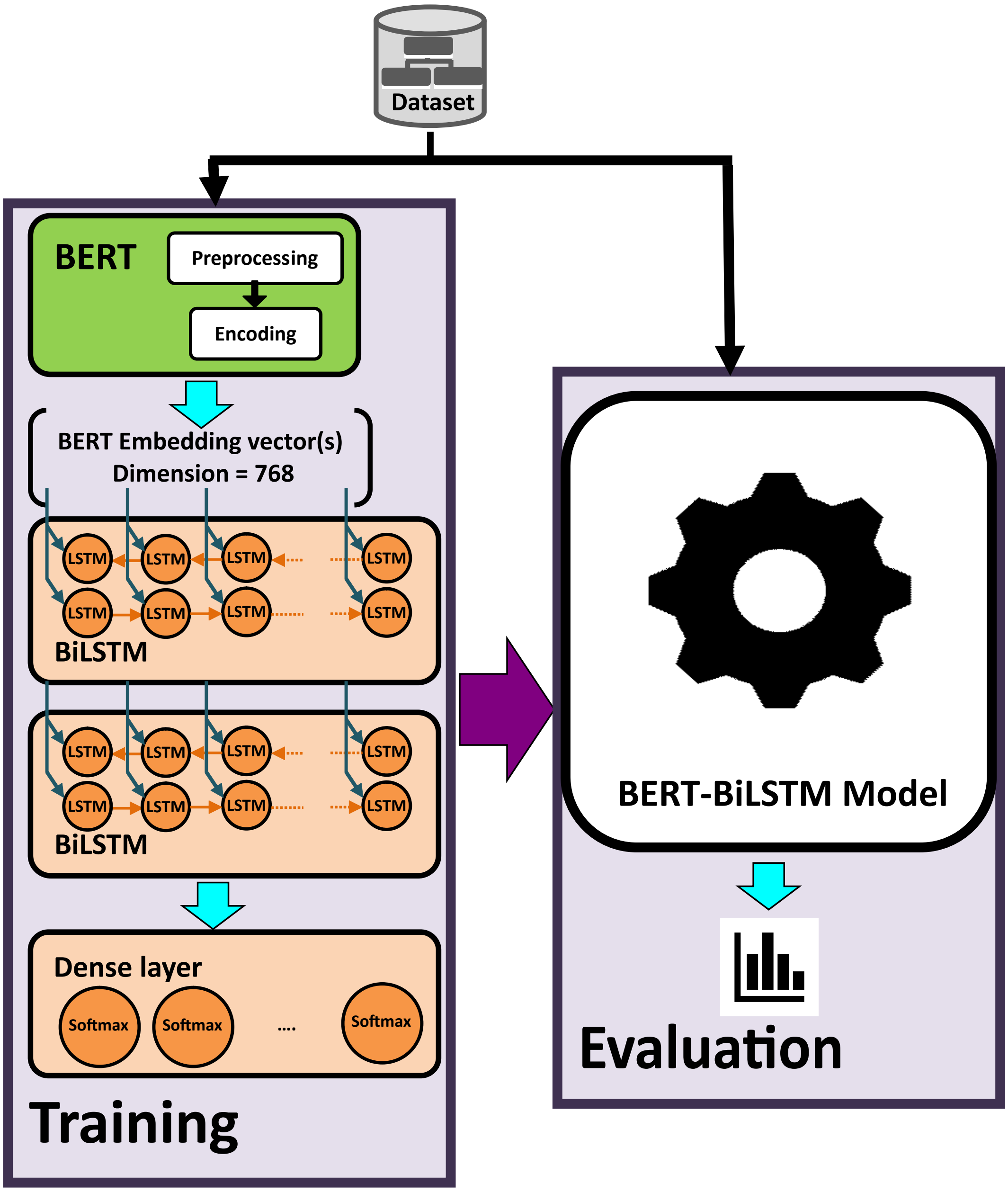

Figure 1. BERT-BiLSTM model architecture

As depicted in Figure 1, the input hierarchical Arabic dataset is divided into two sets: training (70%) and test (30%). During the training phase, BERT preprocesses and encodes the Arabic text as vectors. BERT, based on transformers, is adept at representing text as embedding vectors. The preprocessing step includes tasks such as removing stop-words, numbers, punctuations and non-Arabic words; It also includes tokenization and normalization. The encoding step generates a 768-dimensional from a text.

Subsequently, the BiLSTM model, designed for sequence classification, processes the encoded sequences. The model architecture includes two bidirectional LSTM layers and a dense layer. This architecture effectively captures complex sequence dependencies inherent in text classification tasks.

Once the training phase is completed, the model undergoes evaluation using the test set. The evaluation results, as discussed in the following section, demonstrate the model's performance.

4. EXPERIMENTS

This section details the implementation of our proposed model, the dataset used, the selected Arabic BERT version, the BiLSTM for sequence classification, and the evaluation and comparison results that demonstrate the model's efficiency.

4.1. Implementation

Our model is implemented in Python, utilizing built-in modules that facilitate the construction of various system components. The source code for our model is publicly available on GitHub[1], providing a valuable resource for the AI and NLP communities.

4.2. Dataset

We utilized the WiHArD (Wikipedia-based Hierarchical Arabic Dataset) as our input dataset [21]. The WiHArD dataset is derived from the Arabic Wikipedia category tree, a hierarchical system that organizes articles into categories and subcategories, facilitating navigation through related topics. It begins with broad themes like Math, Culture, and History, which further branch into specific subjects such as Tourism, Events, or Algebra. This structure is crucial for organizing content, as each article is classified under relevant themes, connecting related topics. Its well-defined hierarchy and extensive coverage of Arabic subjects make it a valuable resource for natural language processing tasks like hierarchical text classification.

WiHArD[2] is available on the Elsevier Mendeley Data Platform and comprises over 6000 files organized into three classes at "Level 1" and nine classes at "Level 2" (see Figure 2):

- **Level 1:** Culture (ثقافة), History (تاريخ) and Math (رياضيات), which provide general concepts related to these fields.
- **Level 2:** Clothes (ملابس), Food_drinks (طعام و شراب), Tourism (سياحة), Events (أحداث), Inventions (اختراعات), Monuments (أثار), Algebra (جبر), Analysis (تحليل) and Geometry (هندسة) presenting specific concepts related to sub-fields.

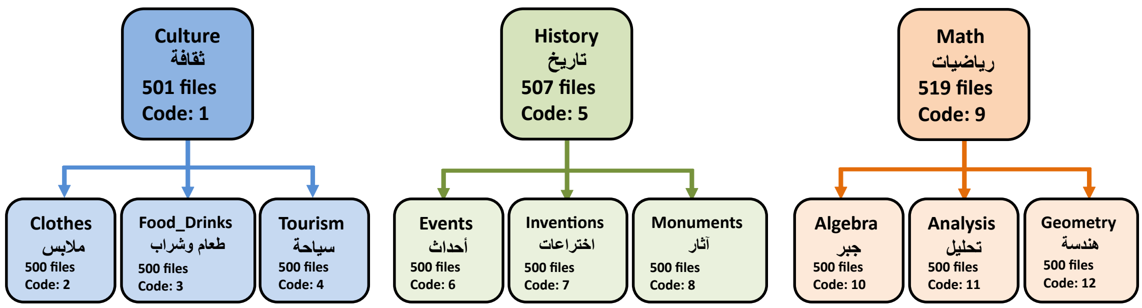

Figure 2. WiHArD taxonomy [21]

[1] https://github.com/khouloud-1/BERT-BiLSTM
[2] https://data.mendeley.com/datasets/kdkryh5rs2/2

4.3. BERT for Arabic language

Several BERT versions tailored for the Arabic language exist. To identify the best-performing model, we conducted a comparison study using various Arabic BERT models and datasets, including ANT (Arabic News Text) [22], the Khaleej dataset [23], and OSAC (Open Source Arabic Corpora) [24]. The BERT versions tested were ArabicBERT [25], AraBERT [26], ARBERT and MARBERT [27], and mBERT [16].

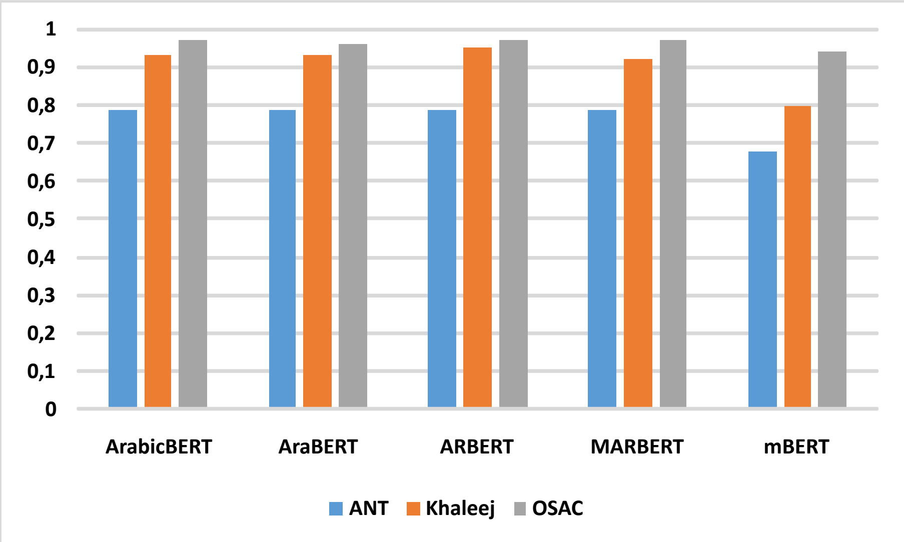

Figure 3. Arabic text classification accuracy (Arabic BERT model/Arabic dataset)

The graph in Figure 3 shows the classification accuracy of each dataset-Arabic BERT combination. Accuracy is defined as the fraction of the total number of texts that are correctly classified [28]:

$$Accuracy = \frac{(TPos + TNeg)}{Num} \tag{1}$$

Where TPos (true positives) is the number of texts correctly classified as positive, TNeg (true negatives) is the number of texts correctly classified as negative, and Num is the total number of texts.

The results indicate that ARBERT consistently delivers the best performance across different datasets.

4.4. BiLSTM for sequence classification

As explained in Section 3, the BiLSTM model is structured as a Sequential model with the following components:

- **Model Initialization:** The model is initialized as a Sequential model.
- **Input Layer:** Accepts input sequences of length 768.
- **First BiLSTM Layer:** A bidirectional LSTM layer with 300 units and return_sequences=True to return outputs for each timestep.
- **Dropout Layer:** A dropout layer with a 0.2 rate to prevent overfitting.
- **Second BiLSTM Layer:** Another bidirectional LSTM layer with 300 units and return_sequences=False.This setup is typical for classification tasks where a single output prediction is required from the sequence input.
- **Another Dropout Layer:** Another dropout layer to reduce overfitting.
- **Dense Output Layer:** A dense layer with 12 units and a softmax activation function. This layer outputs the probability distribution over 12 classes.

- **Model Compilation:** Compiled with the categorical crossentropy loss function and the Adam optimizer.
- **Model Training:** Trained over 50 epochs, with optional validation and early stopping mechanisms to monitor and prevent overfitting.

4.5. Evaluation results

We evaluated our model using the hierarchical F-score (HF-score), which adapts the flat F-score for hierarchical classification by enhancing the classifier's assessment precision [29]. The HF-score is based on hierarchical precision (hP) and hierarchical recall (hR). It is computed as follows:

$$hP = \frac{\Sigma_i |\hat{C}_i \cap \hat{C}'_i|}{\Sigma_i |\hat{C}'_i|} \tag{2}$$

$$hR = \frac{\Sigma_i |\hat{C}_i \cap \hat{C}'_i|}{\Sigma_i |\hat{C}_i|} \tag{3}$$

Where $\hat{C}_i$ contains the true class and its parent nodes for a single instance, and $\hat{C}'_i$ contains the predicted class and its parents for a single instance i.

Now, we can compute the $HF - score$ as follows:

$$hF_\beta = \frac{(\beta^2 + 1) \cdot hP \cdot hR}{(\beta^2 \cdot hP + hR)}, \beta \in [0, +\infty[\tag{4}$$

Where β is used to control the weight of precision over the weight of recall. In our experiments, we set $\beta = 1$ to give equal weight to precision and recall.

The classification results were:

Hierarchical measures	HPrecision	=	0.7748724489795918
	HRecall	=	0.7692307692307693
	HF-score	=	0.7720413026211279

4.6. Comparative study

We compared our deep learning model with a supervised machine learning model for hierarchical Arabic text classification. The machine learning classification process involves two phases: Training and Evaluation [12] (see Figure 4).

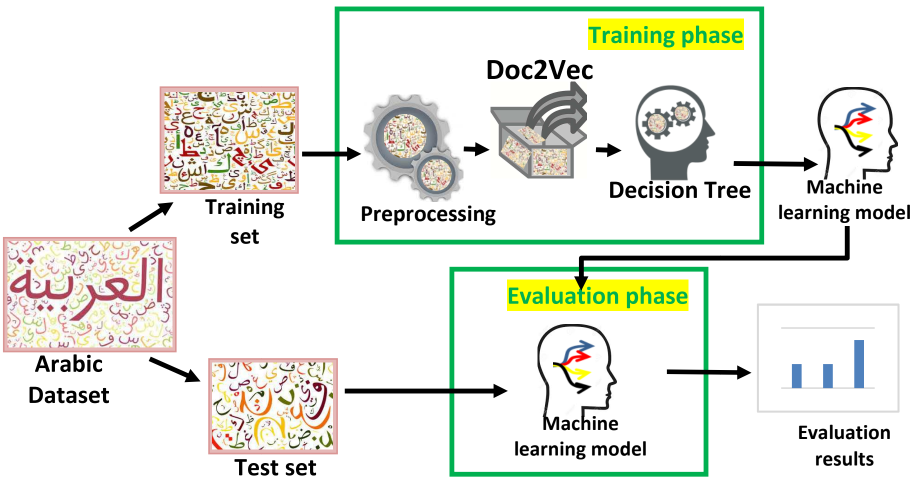

Figure 4. Flowchart of the machine learning Arabic text classification process [12]

The classification process begins with the examination of a labeled dataset, which is divided into two parts: the training set and the test set. Typically, the training set encompasses 70% of the entire dataset, with the remaining 30% allocated to the test set. The model is trained using the training set, and its performance is then evaluated using the test set.

After processing, the texts in the training set are input into the Doc2Vec module. This module converts the Arabic texts into numerical vectors. Doc2Vec is a natural language processing method that employs a neural network to capture relationships between words within a large dataset [30]. It aims to represent the text as a vector [31]. Thus, several texts yield a feature matrix.

The feature matrix serves as the input for the Decision Tree machine learning algorithm, which, after training, becomes a machine learning model. The Decision Tree predicts the class of an instance by traversing from the root node of the tree to a leaf [32]. The resulting machine learning model is then evaluated using the test set.

The graph below shows the comparison results of the hierarchical Arabic text classification process in terms of HPrecision, HRecall, and HF-score between our BERT-BiLSTM model and Doc2Vec-DecisionTree model described above.

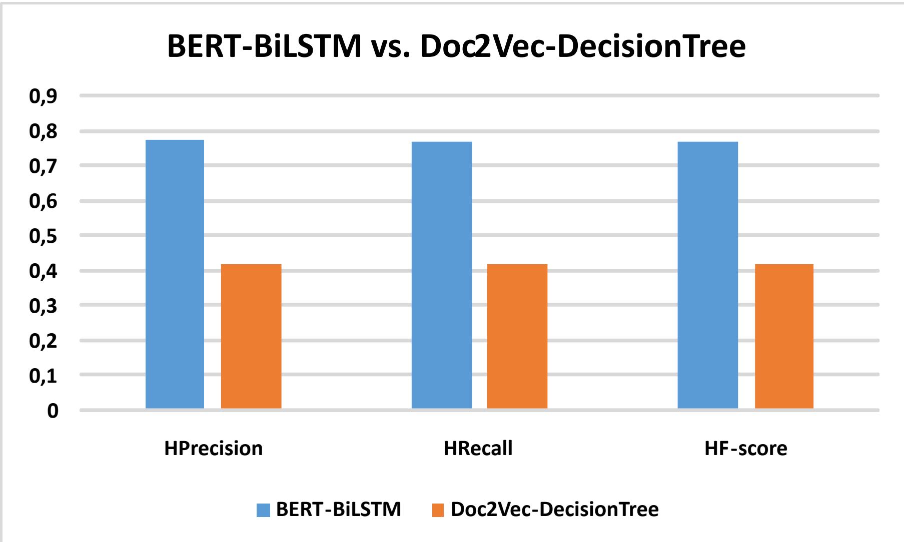

Figure 5. Hierarchical Arabic Text classification comparison results

The results, as illustrated in Figure 5, show that our BERT-BiLSTM model outperforms Doc2Vec-DecisionTree model in terms of HPrecision, HRecall, and HF-score. Nevertheless, these results could be further improved by using a dataset with longer texts, as BERT and BiLSTM typically perform better with longer sequences. Unlike the short texts of the WiHArD dataset, long texts typically contain more contextual information and features that can be extracted. In contrast, short texts can be noisy, often lack sufficient context, and contain less frequent words [33].

5. CONCLUSION AND FUTURE WORK

In this paper, we presented a deep-learning-based approach for hierarchical Arabic text classification, leveraging the power of BERT and BiLSTM models. Our methodology demonstrated promising results in tackling the challenges of categorizing Arabic text into hierarchical structures. As summarized below, our study contributes to the advancement of text classification methodologies, particularly in the context of the Arabic language.

Firstly, our approach showcased the effectiveness of deep learning techniques, specifically BERT and BiLSTM, in addressing the complexities of hierarchical classification tasks. By integrating these models, we achieved notable improvements in classification HF-score and robustness.

However, while our results are encouraging, there remain opportunities for further enhancement. Notably, the dataset used in this study primarily comprises short texts, which may limit the performance of our model, particularly given that BiLSTM tends to excel with longer sequences. Thus, future research endeavors could focus on curating a new hierarchical dataset containing longer texts to better harness the capabilities of BiLSTM and further improve classification results.

Additionally, our comparative analysis with other machine learning models underscores the superiority of deep learning approaches in hierarchical Arabic text classification. By outperforming traditional machine learning methods, our model demonstrates its efficacy and potential for practical applications across various domains.

In conclusion, our study underscores the importance of leveraging advanced deep learning techniques for hierarchical Arabic text classification. Moving forward, the development of new hierarchical datasets and continued refinement of deep learning models hold promise for further enhancing classification accuracy and advancing the field of Arabic language processing.

REFERENCES

[1] A. L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development*, vol. 3, pp. 210-229, 1959.

[2] K. Sundus, F. Al-Haj, and B. Hammo, "A deep learning approach for arabic text classification," in *2019 2nd International Conference on New Trends in Computing Sciences (ICTCS)*, 2019, pp. 1-7.

[3] M. Galal, M. M. Madbouly, and A. El-Zoghby, "Classifying Arabic text using deep learning," *Journal of Theoretical and Applied Information Technology*, vol. 97, pp. 3412-3422, 2019.

[4] F.-Z. El-Alami, A. El Mahdaouy, S. O. El Alaoui, and N. En-Nahnahi, "A deep autoencoder-based representation for Arabic text categorization," *Journal of Information and Communication Technology*, vol. 19, pp. 381-398, 2020.

[5] M. Alhawarat and A. O. Aseeri, "A superior Arabic text categorization deep model (SATCDM)," *IEEE Access*, vol. 8, pp. 24653-24661, 2020.

[6] H. A. Almuzaini and A. M. Azmi, "Impact of stemming and word embedding on deep learning-based Arabic text categorization," *IEEE Access*, vol. 8, pp. 127913-127928, 2020.

[7] N. Aljedani, R. Alotaibi, and M. Taileb, "HMATC: Hierarchical multi-label Arabic text classification model using machine learning," *Egyptian Informatics Journal*, vol. 22, pp. 225-237, 2021/09/01/ 2021.

[8] H. El Rifai, L. Al Qadi, and A. Elnagar, "Arabic text classification: the need for multi-labeling systems," *Neural Computing and Applications*, vol. 34, pp. 1135-1159, 2022.

[9] B. Alsukhni, "Multi-Label Arabic Text Classification Based on Deep Learning," in *Proceedings of the 2021 12th International Conference on Information and Communication Systems*, ICICS, 2021.

[10] F. Elghannam, "Multi-Label Annotation and Classification of Arabic Texts Based on Extracted Seed Keyphrases and Bi-Gram Alphabet Feed Forward Neural Networks Model," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, pp. 1-16, 2022.

[11] H. E. Rifai, L. Al Qadi, and A. Elnagar, "Arabic Multi-label Text Classification of News Articles," in *Advanced Machine Learning Technologies and Applications: Proceedings of AMLTA 2021*, 2021, pp. 431-444.

[12] D. Bouchiha, A. Bouziane, and N. Doumi, "Machine Learning for Arabic Text Classification: A Comparative Study," *Malaysian Journal of Science and Advanced Technology*, pp. 163-173, 2022.

[13] A. M. Bdeir and F. Ibrahim, "A framework for arabic tweets multi-label classification using word embedding and neural networks algorithms," in *Proceedings of the 2020 2nd International Conference on Big Data Engineering*, 2020, pp. 105-112.

[14] A. Wahdan, M. Al-Emran, and K. Shaalan, "A systematic review of Arabic text classification: areas, applications, and future directions," *Soft Computing*, vol. 28, pp. 1545-1566, 2024/01/01 2024.

[15] F. A. Abdulghani and N. A. Abdullah, "A survey on Arabic text classification using deep and machine learning algorithms," *Iraqi Journal of Science*, pp. 409-419, 2022.

[16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 1 (Long and Short Papers), Minneapolis, Minnesota, 2019, pp. 4171-4186.

[17] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735-1780, 1997.

[18] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *IEEE transactions on Signal Processing*, vol. 45, pp. 2673-2681, 1997.

[19] S. Riyanto, I. S. Sitanggang, T. Djatna, and T. D. Atikah, "Plant-Disease Relation Model through BERT-BiLSTM-CRF Approach," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 12, pp. 113-124, 2024.

[20] R. Roslidar, N. Brilianty, M. J. Alhamdi, C. N. Nurbadriani, E. Harnelly, and Z. Zulkarnain, "Improving Bi-LSTM for High Accuracy Protein Sequence Family Classifier," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 12, pp. 40-52, 2024.

[21] D. BOUCHIHA, A. BOUZIANE, N. DOUMI, BERBOUCHI, F. Omar;, A. A. KEBIR, N. MEBARKI, and B. A. BENAMEUR, "WiHArD: Wikipedia based Hierarchical Arabic Dataset," V2 ed. *Mendeley Data*, 2023.

[22] A. Chouigui, O. B. Khiroun, and B. Elayeb, "ANT Corpus: An Arabic News Text Collection for Textual Classification," *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, pp. 135-142, 2017.

[23] M. Abbas and K. Smaïli, "Comparison of Topic Identification methods for Arabic Language," in *International Conference on Recent Advances in Natural Language Processing - RANLP 2005*, Borovets, Bulgaria, 2005.

[24] M. Saad and W. M. Ashour, "OSAC: Open Source Arabic Corpora," 2010.

[25] A. Safaya, M. Abdullatif, and D. Yuret, "KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media," Barcelona (online), 2020, pp. 2054-2059.

[26] W. Antoun, F. Baly, and H. Hajj, "AraBERT: Transformer-based Model for Arabic Language Understanding," Marseille, France,

2020, pp. 9-15.

[27] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, "ARBERT & MARBERT: Deep Bidirectional Transformers for Arabic," *ArXiv*, vol. abs/2101.01785, 2020.

[28] C. E. Metz, "Basic principles of ROC analysis," *Seminars in Nuclear Medicine*, vol. 8, pp. 283-298, 1978/10/01/ 1978.

[29] K. Svetlana, M. Stan, and N. Richard, A. Fazel F."Learning and evaluation in the presence of class hierarchies: application to text categorization," in *Conference of the Canadian Society for Computational Studies of Intelligence*, 2006, pp. 395-406.

[30] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[31] Q. Le and T. Mikolov, "Distributed Representations of Sentences and Documents," in *Proceedings of the 31st International Conference on Machine Learning*, Beijing China, 2014, pp. 1188–1196.

[32] S. Shalev-Shwartz and S. Ben-David, Understanding machine learning: From theory to algorithms: Cambridge university press, 2014.

[33] S. F. C. Haviana and B. S. W. Poetro, "Deep Learning Model for Sentiment Analysis on Short Informal Texts," *Indonesian Journal of Electrical Engineering and Informatics (IJEEI)*, vol. 10, pp. 82-89, 2022.

BIOGRAPHIES OF AUTHORS

Benamar HAMZAOUI received his Engineer degree in Computer Science from Sidi Bel Abbes University, Algeria, in 1997, and his Master's degree in Computer Science from the same university in 2022. Since 2023, he has been pursuing a Ph.D. in Computer Science, with a focus on Artificial Intelligence. His research interests include image processing, databases and database management systems (DBMS), artificial intelligence, machine learning, deep learning, and natural language processing (NLP). He can be contacted at email: hamzaoui@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;.

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 Bel Abbès, 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.

Noureddine DOUMI is currently an Assistant Professor in the Computer Science Department at the University of Saida - Dr. Moulay Tahar. He received his Doctorate in Science in Computer Science from the University of Sidi Bel Abbès in 2017. He is a member of the EEDIS Lab at UDL-SBA and an active developer in the Unitex/GramLab project at the University of Paris-Est Marne-La-Vallée. His research interests include Arabic Natural Language Processing (NLP), linguistic resources, finite-state machines, and machine learning. He can be contacted at email: noureddine.doumi@univ-saida.dz.