computer sciences and mathematics forum

MDPI

Proceeding Paper

A Novel Deep Learning ArCAR System for Arabic Text Recognition with Character-Level Representation †

Abdullah Y. Muaad [1,2,*], Mugahed A. Al-antari [3,*], Sungyoung Lee [4] and Hanumanthappa Jayappa Davanagere [1,*]

[1] Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore 570006, India
[2] IT Department, Sana'a Community College, Sana'a 5695, Yemen
[3] Department of Artificial Intelligence, Sejong University, Sejong 30019, Korea
[4] Department of Computer Science and Engineering, College of Software, Kyung Hee University, Suwon-si 17104, Korea; sylee@oslab.khu.ac.kr
* Correspondence: abdullahmuaad9@gmail.com (A.Y.M.); en.mualshz@sejong.ac.kr (M.A.A.-a.); hanumsbe@gmail.com (H.J.D.)
† Presented at the 1st International Electronic Conference on Algorithms, 27 September–10 October 2021; Available online: https://ioca2021.sciforum.net/.

Abstract: AI-based text classification is a process to classify Arabic contents into their categories. With the increasing number of Arabic texts in our social life, traditional machine learning approaches are facing different challenges due to the complexity of the morphology and the delicate variation of the Arabic language. This work proposes a model to represent and recognize Arabic text at the character level based on the capability of a deep convolutional neural network (CNN). This system was validated using five-fold cross-validation tests for Arabic text document classification. We have used our proposed system to evaluate Arabic text. The ArCAR system shows its capability to classify Arabic text in character-level. For document classification, the ArCAR system achieves the best performance using the AlKhaleej-balance dataset in terms of accuracy equal to 97.76%. The proposed ArCAR seems to provide a practical solution for accurate Arabic text representation, both for understanding and as a classifications system.

Keywords: deep learning ArCAR system; Arabic character-level representation; Arabic text document classification; Arabic sentiment analysis

Citation: Muaad, A.Y.; Al-antari, M.A.; Lee, S.; Davanagere, H.J. A Novel Deep Learning ArCAR System for Arabic Text Recognition with Character-Level Representation. *Comput. Sci. Math. Forum* **2022**, 2, 14. https://doi.org/10.3390/IOCA2021-10903

Academic Editor: Frank Werner

Published: 26 September 2021

1. Introduction

Natural Language Processing (NLP) is one of the most important topics which came from the combination of linguistics and artificial intelligence, etc. NLP is an interesting topic for humans to make interactions with machines. NLP's purpose is to process textual content and extract the most useful information so that we can make better decisions in our daily lives.

There are about 447 million native Arabic speakers and dialects in the world [1,2]. The Arabic language is the main language of 26 Arab countries (i.e., Arab countries) which possesses many difficulties compared to English. Arabic text analytics are incredibly significant with respect to making our lives easier in many domains such as document text categorization [3], Arabic sentiment analysis [4], and detection of email spam. In fact, the Arabic text faces many challenges as mentioned in [5] such as stemming, dialects, phonology, orthography, and morphology. Each level of the classification method necessitates a significant amount of labor and attention from the user, especially with preprocessing text which requires various steps due to the difficulties of Arabic text. Until today most of the representation techniques for the classification of Arabic text have depended on words rather than characters while at the same time the difficulty of stemming Arabic words is still a big challenge. For that reason, we attempted to determine a representation for Arabic

text which would decrease these difficulties. Stemming Arabic words is still a big challenge which, requires an understanding of the word's root which is not easy in many cases.

Due to these challenges, we developed a new Arabic text computer-aided representation and classification system that understands and recognizes Arabic at the character level to classify Arabic documents. This paper will aid in the representation of Arabic text while at the same time assisting the classification.

2. Related Works

The work which has been done for Arabic text representation and classification is much less compared to English text. Little research on the analysis of Arabic text classification had been done but it has been shown to give different results when working with Arabic text. The most important technique for Arabic text classification is usually representation and classification, so in this section, we will survey the most important steps for that reason. In this section, we will conduct a brief literature review focusing on two key stages: representation such as paper [6,7] and classification such as paper [8] as follows:

2.1. Representation

The authors in [8] introduced Term Class Weight-Inverse Class Frequency (TCW-ICF) as a new representation approach for Arabic text. Using their representation, the most promising features of Arabic texts can be retrieved.

Etaiwi et al. introduced an Arabic text categorization model based on a graph-based semantic representation model [7]. Their accuracy, sensitivity, precision, and F1-score, for their work increased by 8.60 percent, 30.20 percent, 5.30 percent, and 16.20 percent, respectively.

To improve Arabic text representation, Almuzaini et al. presented a framework that combined document embedding representation (doc2vec) with sense disambiguation. They then used the OSAC corpus dataset to conduct their work experiments. In terms of F-measure, they were able to attain a text categorization accuracy of 90% [9].

Oueslati et al. implemented Deep CNN to Arabic sentiment analysis (SA) text in 2020. They used character-level features to represent Arabic text for sentiment analysis. As a result, their effort has several limitations, such as the absence of all characters and a large number of Arabic characters, which lead to misunderstandings of the Arabic text [10].

As a result, we're quite enthusiastic to look for a better option for representing Arabic text in order to overcome these challenges.

2.2. Classification

The most crucial phase in the classification of the various contextual Arabic materials into a valid category is the classification itself. Here we survey some of the recent work.

The authors in [11] implemented a fuzzy classifier to improve Arabic document classification performance. Their results were equal to a precision of 60.16%, recall 62.66%, and f-measure 61.18%.

The first character-level deep learning ConvNet for English text classification was proposed by Zhang et al. [12]. They employed eight large-scale datasets to validate their model and had the lowest testing errors across the board.

In 2020, Daif et al. presented AraDIC [6], the first deep learning framework for Arabic document classification based on image-based characters

Ameur et al. suggested a hybrid CNN and RNN deep learning model for categorizing Arabic text documents using static, dynamic, and fine-tuned word embedding [3]. The most meaningful representations from the space of Arabic word embedding are automatically learned using a deep learning CNN model.

Due to this survey of the classification algorithm for Arabic text, we concluded that we should use Python 3.7 programming to complete our project. We also employed machine learning technologies.

3. Proposed Model

Figure 1 shows the proposed framework for Arabic text classification at the character level with two types of algorithms; (1) traditional machine learning, (2) Deep learning using CNN as we mention in Figure 2. Our proposed approach can be used to recognize Arabic documents

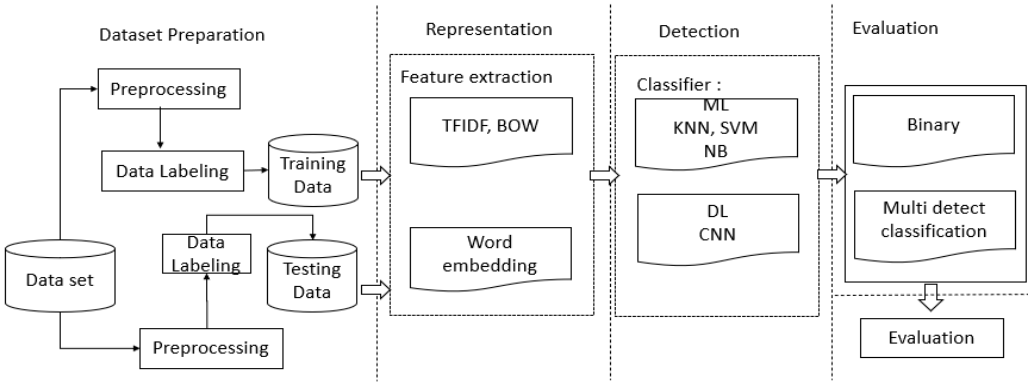

Figure 1. Arabic document classification using machine learning.

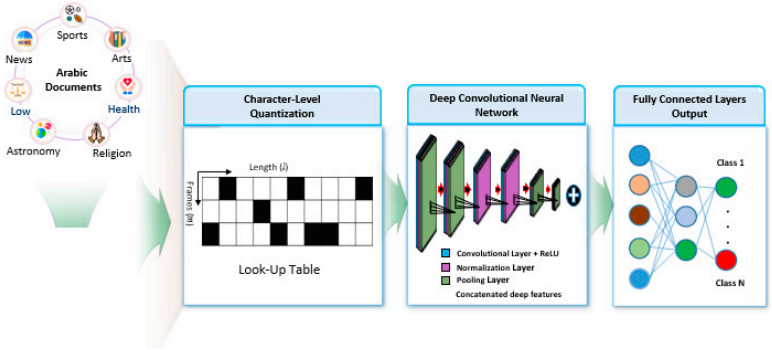

Figure 2. Arabic document classification using deep learning.

3.1. Architecture

The proposed machine learning for Arabic text classification based on different types of representation is presented in Figure 1.

3.2. Machine Learning

This model utilizes two different types of representation TFIDF and BOW.

3.3. Deep Learning

We proposed a deep learning model for Arabic text classification based on CNN. The represented text was at character level as shown in Figure 2 with an Arabic documents classification of accuracy equal to 97. The beauty of this model is that we can avoid preprocessing steps by representing text in character level which at the same time enables better accuracy.

4. Experimental Analysis

We used Python programming to complete our work. We also employed machine learning technologies and data analysis known as scikit-learn[2], TensorFlow, and Kera's. We used a classification system based on CNN and character level representation to classify Arabic text.

4.1. Dataset

This dataset is gathered from all articles published in the news portal from 2008 to 2018. The collected text dataset exceeds a volume of 4 GB and most of the articles published on the websites were not categorized and had a vague label. As a result, there were seven categories populated with a reasonable number of articles under each category to serve the text classification tasks. The dataset was balanced by restricting the number of articles in each category to around 6500, as shown in Table 1

Table 1. Data Distribution Per Class for Alkhaleej Corpus.

Class Name	No. of Documnet
Finance	6500
Sports	6500
Culture	6500
Technology	6500
Politics	6500
Medical	6500
Religion	6500

4.2. Implementation Environment

We utilized a PC with the following characteristics to carry out all of the experiments in this study: One NVIDIA GeForce GTX 1080 GPU and an Intel R Core(TM) i5 K processor with 8 GB RAM and a 3.360 GHz clock. The described system is built with a Python 3.7 with TensorFlow and Kera's back-end libraries on a Windows operating system.

4.3. Evaluation Metrics

To evaluate our proposed ArCAR, we used the following metrices as in [13]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{1}$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \tag{2}$$

$$\text{F1} - \textit{Measure} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}}, \tag{3}$$

$$\text{Overall Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}, \tag{4}$$

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative detections, respectively. A multidimensional confusion matrix was utilized to generate all of these properties. Finally, we used the weighted-class technique to determine the evaluation for each dataset to avoid having test sets that were uneven across all classes or indices [14,15].

5. Results and Discussion

The algorithms such as MNB, BNB, Logistic Regression, SGD Classifier, SVC, and linear SVC are implemented herein using Python with Anaconda [Jupyter notebook]. The proposed methods use Python-based machine learning tools such as NLTK, pandas, and scikit-learn to investigate performance indicators. Meanwhile, for deep learning models such as CNN, additional libraries like as Kera's and TensorFlow were used. The results and discussions concerning the various techniques incorporated are highlighted in the subsequent sections.

5.1. Machine Learning

For this work, the proposed system was evaluated using Khaleej datasets with machine learning. As shown in Table 2, the best performance was achieved using Linear SVC with

Accuracy 93 with TFIDF representation. At the same time, the best accuracy with BOW representation was SGD Classifier.

Table 2. Accuracy for Alkhaleej with and without preprocessing.

Classifiers	BOW without Pre	BOW with Pre	TFIDF without Pre	TFIDF with Pre
Multinomial NB	88	88	64	58
Bernoulli NB	61	73	61	73
Logistic Regression	92	92	90	91
SGD Classifier	91	91	93	92
SVC	90	91	90	92
Linear SVC	92	91	93	92

5.2. Our Proposed Deep Learning

For this work, the proposed system was evaluated using Khaleej datasets with deep learning. As shown in Table 3 and Figure 3, the best performance was achieved using CNN with overall accuracy, F1 measure score, precision, and recall, of 97.47%, 93.23%, 92.75%, and 92%, respectively.

Table 3. Result of the proposed system in deep learning.

Metrics	Accuracy	F Measure-Score	Precision	Recall
AlKhaleej data	97.47	92.63	92.75	92

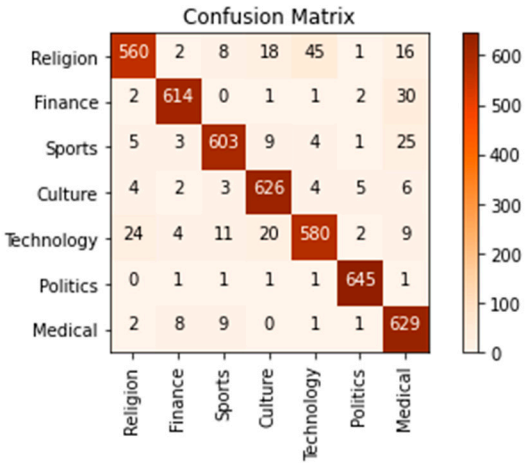

Figure 3. Averaged multiclass confusion matrices AlKhaleej data set.

6. Conclusions

This paper provides a new deep learning strategy for character-level Arabic text classification in Arabic text data. We used datasets in the multiclass problem to demonstrate our system's dependability and capability regardless of the number of classes in our technique, which encodes Arabic text at the character level to avoid preprocessing restrictions like stemming. Simultaneously, we compared our results to those of five machine learning techniques to show that our model outperformed them all. The following are our future plans to increase the performance of the planned system: The problems of multi-label text categorization and Arabic data augmentation need to be handled.

Author Contributions: Conceptualization, A.Y.M. and M.A.A.-a.; methodology, A.Y.M. and M.A.A.-a.; software, A.Y.M.; validation, A.Y.M. and M.A.A.-a.; formal analysis, A.Y.M.; investigation, H.J.D. and A.Y.M.; resources, A.Y.M. and H.J.D.; data curation, A.Y.M.; writing—original draft preparation, A.Y.M. and M.A.A.-a.; writing—review and editing, A.Y.M. and M.A.A.-a.; visualization, M.A.A.-a.;

supervision, S.L.; project administration, M.A.A.-a.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by University of Mysore and the Ministry of Science and ICT (MSIT), South Korea, through the Information Technology Research Center (ITRC) Support Program under Grant IITP-2021-2017-0-01629, and in part by the Institute for Information & Communications Technology Promotion (IITP), through the Korea Government (MSIT) under Grant 2017-0-00655 and IITP-2021-2020-0-01489 and Grant NRF-2019R1A2C2090504.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest: There are no conflict of interest associated with publishing this paper.

References

1. Hakak, S.; Kamsin, A.; Tayan, O.; Idris, M.Y.I.; Gilkar, G.A. Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges. *Inf. Process. Manag.* **2019**, *56*, 367–380. [CrossRef]
2. Elnagar, A.; Al-Debsi, R.; Einea, O. Arabic text classification using deep learning models. *Inf. Process. Manag.* **2020**, *57*, 102121. [CrossRef]
3. Ameur, M.; Belkebir, R.; Guessoum, A. Robust Arabic Text Categorization by Combining Convolutional and Recurrent Neural Networks. *ACM Trans. Asian Low-Resour. Lang. Inf. Process. (TALLIP)* **2020**, *19*, 1–16. [CrossRef]
4. Harrat, S.; Meftouh, K.; Smaili, K. Machine translation for Arabic dialects (survey). *Inf. Process. Manag.* **2019**, *56*, 262–273. [CrossRef]
5. Bounhas, I.; Soudani, N.; Slimani, Y. Building a morpho-semantic knowledge graph for Arabic information retrieval. *Inf. Process. Manag.* **2020**, *57*, 102124. [CrossRef]
6. Daif, M.; Kitada, S.; Iyatomi, H. AraDIC: Arabic Document Classification using Image-Based Character Embeddings and Class-Balanced Loss. *arXiv* **2020**, arXiv:2006.11586.
7. Etaiwi, W.; Awajan, A. Graph-based Arabic text semantic representation. *Inf. Process. Manag.* **2020**, *57*, 102183. [CrossRef]
8. Almuzaini, G.H.A.; Azmi, A.M. Impact of stemming and word embedding on deep learning-based Arabic text categorization. *IEEE Access* **2020**, *8*, 127913–127928. [CrossRef]
9. Oueslati, O.; Cambria, E.; HajHmida, M.B.; Ounelli, H. A review of sentiment analysis research in Arabic language. *Future Gener. Comput. Syst.* **2020**, *112*, 408–430. [CrossRef]
10. Al-Taani, A.T.; Al-Sayadi, S.H. Classification of Arabic Text Using Singular Value Decomposition and Fuzzy C-Means Algorithms. In *Applications of Machine Learning*; Springer: Berlin, Germany, 2020; pp. 111–123.
11. Zhang, X.; Zhao, J.; LeCun, Y. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*; NeurIPS Proceedings; Palais Des Congrès De Montréal: Montréal, QC, Canada, 2015; pp. 649–657.
12. Muaad, A.Y.; Jayappa, H.; Al-antari, M.A.; Lee, S. ArCAR: A Novel Deep Learning ComputerAided Recognition for Character-Level Arabic Text Representation and Recognition. *Algorithms* **2021**, *14*, 216. [CrossRef]
13. Al-Antari, M.A.; Hua, C.H.; Bang, J.; Lee, S. Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest X-ray images. *Appl. Intell.* **2020**, *51*, 2890–2907. [CrossRef] [PubMed]
14. Hanumanthappa, J.; Muaad, A.Y.; Bibal Benifa, J.V.; Chola, C.; Hiremath, V.; Pramodha, M. IoT-Based Smart Diagnosis System for HealthCare. In *Sustainable Communication Networks and Application*; Karrupusamy, P., Balas, V.E., Shi, Y., Eds.; Lecture Notes on Data Engineering and Communications Technologies; Springer: Singapore, 2022; Volume 93. [CrossRef]
15. Muaad, A.Y.; Hanumanthappa, J.; Al-antari, M.A.; JV, B.B.; Chola, C. AI-based Misogyny Detection from Arabic Levantine Twitter Tweets. *Algorithms* **2021**, *14*, 4–11.