IOCA 2021

MDPI

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

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

[1] Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore 570006, India; Abdullahmuaad9@gmail.com (A.Y.M.); hanumsbe@gmail.com (H.J.);
[2] Sana'a Community College, Sana'a 5695, Yemen
[3] Department of Computer Science and Engineering, College of Software, Kyung Hee University, Suwon-si 17104, Korea, sylee@oslab.khu.ac.kr, en.mualshz@khu.ac.kr (M. A. A.),
* Correspondence

Abstract:

AI-based text classification is a process of classify Arabic contents into their category. With the increasing number of Arabic texts in our social life, traditional machine learning approaches are facing different challenges due to the complexity of morphology and the variation delicate of the Arabic language. In this work, proposed model to represent and recognize Arabic text at the character level based on the capability of a deep convolutional neural network (CNN). This system is validated using five-fold cross-validation tests Arabic text document classification. We have been used to evaluate our proposed system for 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%, respectively. The proposed ArCAR seems to provide a practical solution for accurate Arabic text representation, understanding, and classifications system.

Keywords: keyword; Keywords: Deep Learning ArCAR System; Arabic Character-level Representation; Arabic Text Document Classification; Arabic Sentiment Analysis.

Citation

Published: date

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

Natural Language Processing (NLP) is one of the most important topics which came from combination of linguistics and artificial intelligence etc. NLP interested topic for human to make interactions with machine. 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 in the world [1] [2] and its dialects. The Arabic language the main language of 26 Arab countries (i.e., Arab countries) which have a lot of difficulties compare to English. Arabic text analytics are incredibly significant in order to make 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 they mention 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 require difficult steps due to difficulties of Arabic text. Until today most of representation techniques for classification Arabic text depend on words rather than character at the same time difficult of stemming Arabic word still big challenge for that reason, we are trying to find representation for Arabic text which will decrease these difficulties. Stemming of Arabic word still a big challenge which, require understanding the word's root which not easy for many cases.

Due to these challenges, we developed 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 representing of Arabic text. At the same time will aid in the classification of Arabic text.

2. Related Works

The work which has been done for the Arabic text representation and classification is very less comparing to the English text. Few researches on analysis of Arabic text classification had been done and it showed different result to work 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 and classification as follows:

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 are retrieved.

Etaiwi et al. introduced an Arabic text categorization model based on a graph-based semantic representation model in [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. They present 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, this effort has several limitations, such as the absence of all characters and a large number of Arabic characters, which will lead to misunderstandings for 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.

Classification

The most crucial phase in classification the various contextual Arabic materials into a valid category is classification here we will survey some of last work
The authors in [11] implemented fuzzy classifier to improve Arabic document classification performance. Their result equal to precision 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. in [12]. They employed eight large-scale datasets to validate their model, and they 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 in [3]. The

most meaningful representations from the space of Arabic word embedding are automatically learned using a deep learning CNN model.

During to this survey in classification algorithm for Arabic text we conclude that. We used Python 3.7 programming to complete our project. We also employed machine learning technologies.

3. Proposed Model

Figure 3.1 shows the proposed framework for Arabic text classification at the character level with tow type of algorithms (1) traditional machine learning (2) Deep learning using CNN as we mention in figure 3.2. Our proposed approach can be used to recognize Arabic documents

3.1. Architecture

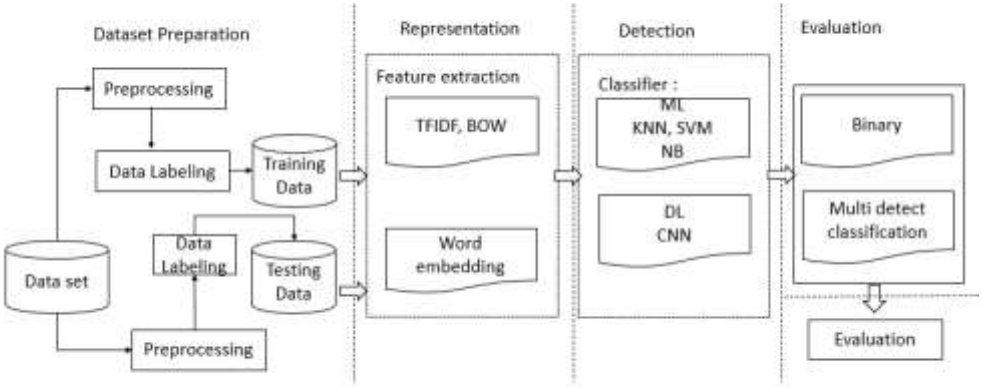

Figure 1. Arabic document classification using machine learning

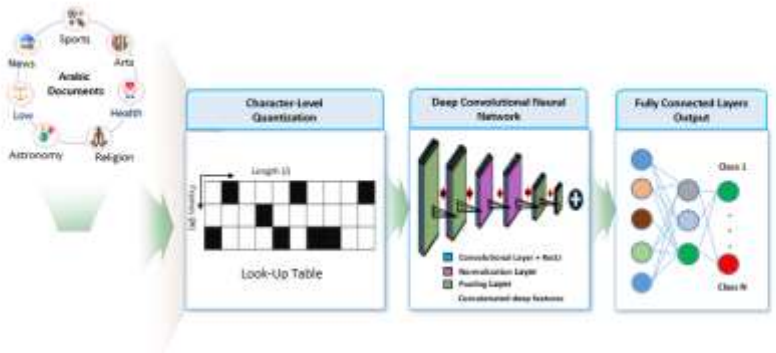

Figure 2. Arabic document classification using deep learning

3.2. Machine learning

The proposed machine learning for Arabic text classification based on different type representation is presented in **Figure 1.** This model utilize two different type of representation TFIDF and BOW.

33. Deep learning

We proposed deep learning model for Arabic text classification based on CNN. The represented text was at character level. As We have shown in **Figure 1.** Arabic documents classification with accuracy equal to 97. The beauty of this model that we can avoid preprocessing steps by represent text in character level which at the same time help us to get 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 representaion to classify Arabic text.

4.1. Dataset

This dataset is scraped from all articles published in the news portal from 2008 to 2018. The collected text dataset exceeds the volume of 4GB and most of the articles published on the websites were not categorized and had a vague label. As a result, it can be mentioned in seven categories and populated under each category with a reasonable number of articles to serve the text classification tasks. The dataset is balanced by restricting the amount of articles in each category to around 6,500, as shown in Table 1

Table 1. Data Distribution Per Class for Alkhaleej Corpus

Class Type	AlKhaleej
Finance	6,500
Sports	6,500
Culture	6,500
Technology	6,500
Politics	6,500
Medical	6,500
Religion	6,500

4.2. Implementation Environment

We utilize 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 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 use the following metrices as in [13]

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad (1)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad (2)$$

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

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

where TP, TN, FP, and FN stand for true positive, true negative, false positive, and false negative detections, respectively. A multidimensional confusion matrix is 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. indices [14]

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 have been used. The results and discussions concerning to various techniques incorporated are highlighted in the subsequent sections.

5.1. Machine Learning

For this work, the proposed system is evaluated using Khaleej datasets with machine learning. As shown in **TABLE 2**, the best performance is achieved using Linear SVC with Accuracy 93 with TFIDF representation. At the same time the best accuracy with BOW representation is 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 is evaluated using Khaleej datasets with deep learning. As shown in **TABLE 3**, the best performance is 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

6. Conclusion

This paper provides a new deep learning strategy for character-level Arabic text classification in Arabic text data. We use 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 compare our results to those of five machine learning techniques to show that our model outperforms them all. The following are future plans to increase the performance of the planned system: The problem of multi-label text categorization and Arabic data augmentation could be handled.

ACKNOWLEDGEMENT

This work was supported in part by 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.

References

1. S. Hakak, A. Kamsin, O. Tayan, M. Y. I. Idris, and G. A. Gilkar, "Approaches for preserving content integrity of sensitive online Arabic content: A survey and research challenges," Information Processing & Management, vol. 56, pp. 367-380, 2019.
2. A. Elnagar, R. Al-Debsi, and O. Einea, "Arabic text classification using deep learning models," *Information Processing & Management,* vol. 57, p. 102121, 2020.
3. M. Ameur, R. Belkebir, and A. Guessoum, "Robust Arabic Text Categorization by Combining Convolutional and Recurrent Neural Networks," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),* vol. 19, pp. 1-16, 2020.
4. S. Harrat, K. Meftouh, and K. Smaili, "Machine translation for Arabic dialects (survey)," *Information Processing & Management,* vol. 56, pp. 262-273, 2019.
5. I. Bounhas, N. Soudani, and Y. Slimani, "Building a morpho-semantic knowledge graph for Arabic information retrieval," *Information Processing & Management,* vol. 57, p. 102124, 2020.
6. M. Daif, S. Kitada, and H. Iyatomi, "AraDIC: Arabic Document Classification using Image-Based Character Embeddings and Class-Balanced Loss," *arXiv:2006.11586,* 2020.
7. W. Etaiwi and A. Awajan, "Graph-based Arabic text semantic representation," *Information Processing & Management,* vol. 57, p. 102183, 2020.
8. Guru
9. 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.
10. O. Oueslati, E. Cambria, M. B. HajHmida, and H. Ounelli, "A review of sentiment analysis research in Arabic language," *Future Generation Computer Systems,* vol. 112, pp. 408-430, 2020.
11. A. T. Al-Taani and S. H. Al-Sayadi, "Classification of Arabic Text Using Singular Value Decomposition and Fuzzy C-Means Algorithms," in *Applications of Machine Learning*, ed: Springer, 2020, pp. 111-123.
12. X. Zhang, J. Zhao, and Y. LeCun, "Character-level convolutional networks for text classification," in *Advances in neural information processing systems*, 2015, pp. 649-657.
13. Muaad, A.Y.; Jayappa, H.; Al-antari, M.A.; Lee, S. ArCAR: A Novel Deep Learning ComputerAided Recognition for CharacterLevel Arabic Text Representation and Recognition. Algorithms 2021, 14, 216. https://doi.org/10.3390/a14070216.

14. M. A. Al-antari, C.-H. Hua, J. Bang, and S. Lee, "Fast deep learning computer-aided diagnosis of COVID-19 based on digital chest x-ray images," *Applied Intelligence,* pp. 1-18, 2020.
15. Hanumanthappa .J et,al. "IoT-Based Smart Diagnosis System for Health Care"2021/7/29, 3rd International Conference on Sustainable Communication Networks and Application ICSCN 2021,ICSCN-2021 Springer Conference.