Umm Al-Qura University
Faculty of Computer Science and Information Systems
Computer Science Department

Arabic Poetry Classification based on Era using Machine Learning with Word and Document Embedding

Represented by:
Manal Anetallah Alsahafi

Supervised by:
Dr.Ashwag Maghraby

17-5-2022
Semester II, 2021/2022

Abstract

Arabic Poetry classification is still one of the recent fields in Arabic natural language processing. This work applied two different pre-trained embedding models AraVec and Doc2Vec Embeddings with text classification. Indeed, this study proposed two of the machine learning model classifiers: Logistic Regression (LR) and Support Vector Machines (SVM). As there were a few public datasets for Arabic Poetry, we constructed a dataset from the Adab website containing the Islamic and Modern eras. Then, we started with a preprocessing phase of text to be compatible with the classifier. In the first experiment, LR and SVM classifiers were implemented with Doc2Vec, which achieved the highest accuracy and F1 score of 0.95 each. In the second experiment, the two classifiers implemented with AraVec reported less accuracy and an F1 score of 0.80 each.

1. INTRODUCTION

Over the years, there was a development in Arabic poetry in many aspects as the topics, the rhythm, and the terms used in the poems changed. This can be an indication that the poetry is in an era that is different from the previous one [1]. There are a few studies in the field of classification of Arabic poetry based on their eras by using machine learning [2,3]. The main aim of this paper was to classify Arabic poetry based on machine learning into Islamic and Modern eras using the Logistic Regression (LR) and the Support Vector Machines (SVM) models. By using embeddings to represent words as vectors in a continuous space, capturing many syntactic and semantic relations among them. So, the performance was compared with two different representations of a text: AraVec as a pre-trained word embedding model and Doc2Vec as Document embedding. The results were compared with the previous studies. Our investigation is on how effective the LR Model is and the SVM model that uses the Semantic Embedding models on categorizing Arabic poems and contributes to comparing their performances.

This research is organized as follows; section 2 defined the embedding techniques and machine learning algorithms, section 3 related works are reviewed and the gap that is filled by this paper, then a dataset collecting, cleaning, and preprocessing are described in section 4 and the architecture of the used classifiers and the pre-trained embedding models is presented in section 5. Finally, after training and testing the classifiers, the results are analyzed and discussed in section 6 and the conclusions with future work are reported in section 7.

2. BACKGROUND

This section introduced the concept of word embedding which was used to represent text as real-value vector numbers to make it ready to feed the machine learning models that have been used in this study, the Support Vector Machines, and Logistic Regression models.

2.1 WORD EMBEDDING

The phrase "word Embedding" was first coined by the study by Bengio et al. [4]. After that, Mikolov et al. [5] introduced the concept of the Word2Vec toolkit which became the vanguard of studies and contributed to its large use. It was used and tuned to generate embeddings without any effort and losing time. They proposed two one of-a-kind model architectures for representing phrases in a multidimensional vector space namely the continuous bag-of-words (CBOW) model and the skip-gram model. The CBOW pursuits to analyze the embeddings with the aid of using predicting the middle phrase in a context given the alternative phrases withinside the context without regard to their order withinside the sentence. The Skip-Gram is the opposite of the CBOW because it pursues to are expecting the encompassing context phrases given the middle phrase. Figure 1. Describe the architectures of these models.

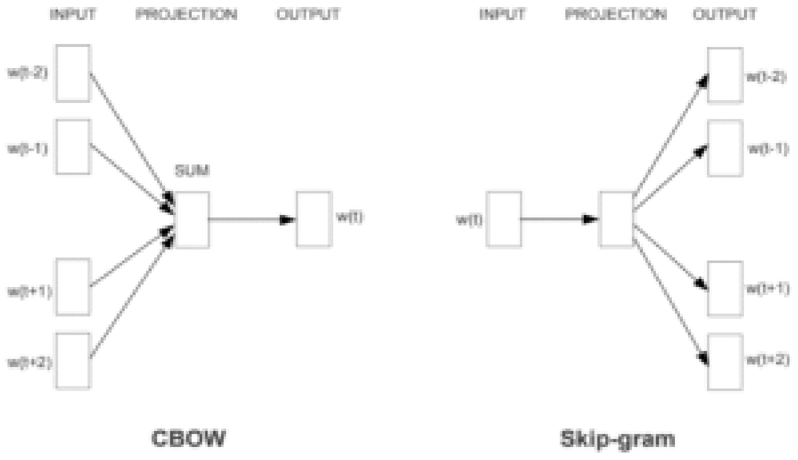


FIGURE 1. THE ARCHITECTURES OF CBOW AND SKIP-GRAM

2.3 MACHINE LEARNING ALGORITHMS

In our approach, two Machine Learning algorithms were selected for the classification of Arabic poetry. These algorithms have been proven successful in the classification of Arabic text. The first algorithm is Support Vector Machines (SVM), and the second is Logistic Regression (LR).

2.3.1 SUPPORT VECTOR MACHINES

SVM is a computationally kernel-primarily based totally set of rules for regression and binary information class purposes. It has a higher generalization overall performance in comparison to different Machine Learning techniques which include Artificial Neural Networks (ANNs)[6]. SVM has thus far been exceptional in fixing numerous real-global information mining predictive issues like time collection prediction, textual content categorization, photograph processing, and sample recognition [7].

2.3.2 LOGISTIC REGRESSION

Logistic Regression (LR) is a well-known statistical algorithm that has the advantage of yielding a probability model that can be useful in many applications. It is a discriminative model that is suitable for binary classification [8].

3. RELATED WORKS

Arabic text classification has appeared in many studies with different methods, where Support Vector Machine (SVM) was one of the algorithms used in the past research [9], such as El Mahdaouy et al. (2016) classify the Arabic text using the SVM classifier based on Doc2vec and Glove document embeddings to generate text representations. Document and word vectors were less sensitive to learning parameters. For increasing greatly, the performance of Arabic Text Categorization, El-Alami, and Alaoui (2018) proposed a method that used SVM and Naïve Bayes model for classifying Arabic text into multi-classes based on Doc2vec with the Word Sense Disambiguation (WSD) to choose the nearest concept for the ambiguous terms.

These studies showed the power of integrating the embedding technique with the machine learning model in Arabic text classification, but they were not applied to Arabic poetry. To show the possibility to distinguish, with decent accuracy, poems from different eras, El Gharbat et al. (2019) applied different classification algorithms to classify the poems into Abbasid and Andalusian eras based on identifying discriminant features. SVM classifier obtained the highest accuracy of 70.50%. Despite the difference in the method of extracting features from the text in relation to the two previous studies that adopted embedding, it confirmed that the SVM model gives the best results in classifying Arabic poems in their correct era. This research contributes to investigating and comparing the performance of the SVM model and the Logistic Regression (LR) Model to classify Arabic poetry to find the outperformed model of one of them for Arabic Poetry classification.

4. RESEARCH METHOD

Figure 2 demonstrates the key phases we followed. In the beginning, choosing the dataset that will use; after that, all the steps of data preprocessing were applied, including the embedding models (Doc2vec and AraVec). Two machine learning algorithms (SVM and LR) were used in training and testing.

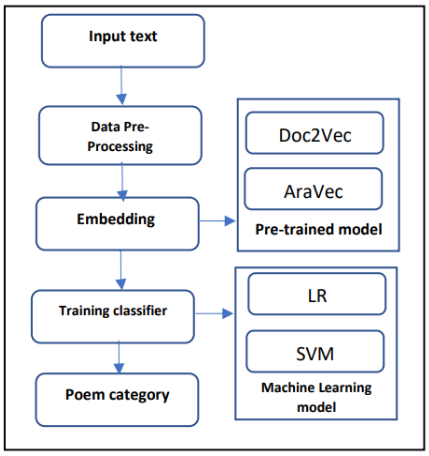


FIGURE2 . BLOCK DIAGRAM OF THE PROPOSED METHOD

4.1 ARABIC POETRY CORPUS BUILDING

Due to the limited number of free available datasets in Arabic poetry, the dataset was scraped from the Adab website, which contains 5979 poems in Arabic by 149 poets. The poems are categorized by their chronological origins as follows: Islamic era and Modern era. The pre-trained embedding model encodes with a fixed-length vector of integers for each document [10]. So, there was no need for dividing the poems into fragments. Details of the dataset can be found in Table 1.

TABLE 1. NUMBER OF POEMS OF EACH ERA

Era	Number of poems
Islamic	2990
Moder n	2989
Total	5979

4.2 DATA PREPROCESSING

At this stage, several pre-processing steps have been applied to prepare the dataset for the classification phase. The importance of this stage is in giving precision to the classification task.

4.2.1 Cleaning process

The cleaning process consists of removing:

1. Arabic diacritics and punctuation marks.
2. Non-Arabic characters such as numbers and English characters.
3. Empty lines and one or more spaces.

4.2.2 Tokenization process

The tokenization process is related to dividing the text into an array of words known as tokens. The tokenization depends mostly on the white spaces between words to retrieve the tokens. In this work, we divided the poems into sentences according to punctuation marks that are repeated in the dataset like '--', '...', and '؟ .'?Then, tokenize the words of each sentence.

4.2.3 Removing Stop words

Arabic stop words were collected from the previously published list and added to the list of stop words that have been discovered while the pre-processing for poetry texts.

4.2.4 Stemming

Stemming is the process of getting the root of the words, usually, in the Arabic language, the roots of most the words consist of 3 letters [11]. In our work, we did not apply any stemming due to the large difference in the shapes of words and we want to keep the meaning of words for the embedding phase.

4.3 DOC2VEC EMBEDDING MODELS

Doc2vec is an NLP tool for representing documents as vectors, which is a generalization of the word2vec method. The main purpose of Doc2Vec is to convert a sentence (or paragraph) into a vector[10]. In the Distributed Bag of Words version of Paragraph Vector (PV-DBOW), each paragraph of the poetry dataset changed into a completely unique vector, represented with matrix D and each word changed into a specific vector, represented with matrix W. To predict the next word in a context the paragraph and word vectors are averaged or concatenated. The other model namely Paragraph Vector with distributed memory as opposed to Distributed Memory version of Paragraph Vector (PV-DM) where the paragraph vector was used to predict the words. These techniques are shown in Figure 3.

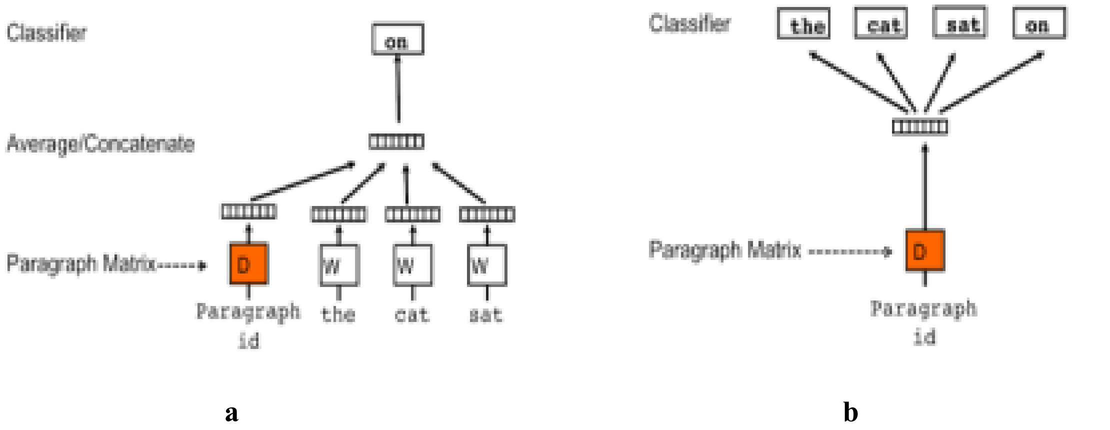


FIGURE 3. NEURAL NETWORK ARCHITECTURE FOR(A) PV-DBOW AND (B) PV-DM

After preparing the dataset to be compatible with Doc2vec, the embeddings were trained in a supervised manner on the labeled corpus using the two models PV-DM and PV-DBOW. In this study, the PV-DBOW was given the best accuracy with the classifiers. This conclusion was reached after the experiment, and the result was presented in the next section.

4.4 ARAVEC EMBEDDING MODEL

The pre-trained AraVec model provides distributed word representation using the CBOW and Skip-Gram technique of the Word2Vec model with different versions. These models were built on top of three different Arabic content; Tweets, World Wide Web pages, and Wikipedia Arabic articles [12]. In this work, AraVec 2.0 was chosen that built on the CBOW technique with Web content. Also, the model that loaded had the same Dimension of Doc2vec which was set to 100.

5. CLASSIFIERS

After the dataset was collected, cleaned, preprocessed, and embeddings were extracted and trained, the supervised models were trained to build SVM and LR classifiers.

Doc2vec with Classifiers: The default parameters of the Doc2vec model were used, only the Dimension of the vectors changed to 100 and the number of epochs increased to 60 to be more compatible with the provided dataset. Afterward, these representations were fed to the classifier. In the LR the optimizer was changed to liblinear as the size of our dataset and the rest parameters were still at default. The SVM was applied with no changing in its parameters.

AraVec with Classifiers: To train the model on our dataset, the custom class should be built to set the new parameters to the model. The maximum number of tokens in the sentence was set to be 1200 and the two classifiers were built as the same in Doc2vec.

6. EXPERIMENTS AND RESULTS DISCUSSION

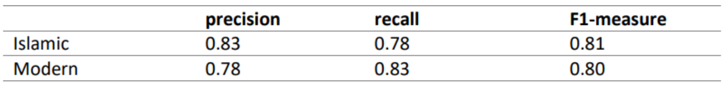
The work was done with the genism Python library for NLP used consists of implementations for Doc2vec embeddings and word2vec to load and use the AraVec model. Then, the classifiers were built and tested in Python using google colab notebook, and these machine configurations: OS: Windows 11, CPU Speed: 3.20 GHz, Processor: Intel Core i7, RAM: 16GB. To compare the performances of the different classifiers for all eras of Arabic Poems, different parameters such as precision, recall, and f-measure were measured.

The performance of the proposed method is presented in Tables 2 to 5 as described below. The first type of machine learning algorithm used was Logistic Regression (LR) with AraVec. Table 2 illustrates the precision, recall, and f-measure for this algorithm.

The maximum value for precision was for the Islamic class, while for the recall, the maximum value was for the Modern class. F-measure was highest in the Islamic class. The results for this algorithm were compared to the results of other machine learning algorithms. Table 3 presents the results of the LR algorithm with Doc2vec. From the results, the maximum value for precision was for the Modern class, while the recall value and F-measure were all for the Islamic class. This result was also compared to the results of the other machine learning frameworks.

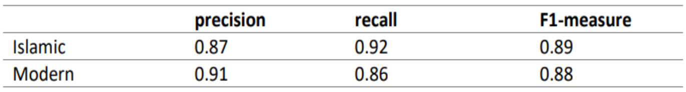
Table 4 illustrates the result of the classification process using Support Vector Machine (SVM) algorithm with AraVec. From the results, the maximum value of precision, recall, and f-measure were the same in the LR with AraVec. The results for this algorithm were compared to the results of other machine learning algorithms. Table 5 illustrates the result of the classification process using the SVM algorithm with Doc2vec. Precision, recall, and f-measure were all the same results that had been get from LR with Doc2vec. This result was also compared to the results of the other machine learning frameworks. From Table 6, the SVM algorithm was found to have the maximum precision, recall, and f-measure values with the pretrained Doc2vec model. On the other hand, AraVec model with the two classifiers had the same score.

TABLE2 . CLASSIFICATION OF OUR DATASET USING LR WITH ARAVEC



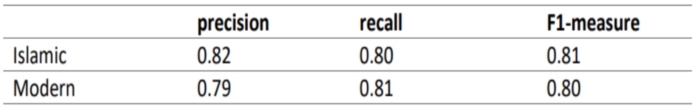
	precision	recall	F1-measure
Islamic	0.83	0.78	0.81
Modern	0.78	0.83	0.80

TABLE3 . CLASSIFICATION OF OUR DATASET USING LR WITH DOC2VEC



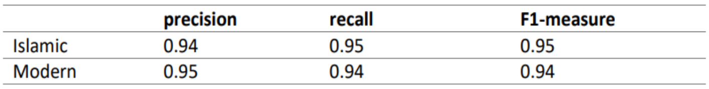
	precision	recall	F1-measure
Islamic	0.87	0.92	0.89
Modern	0.91	0.86	0.88

Table4 . Classification of our Dataset using SVM with AraVec



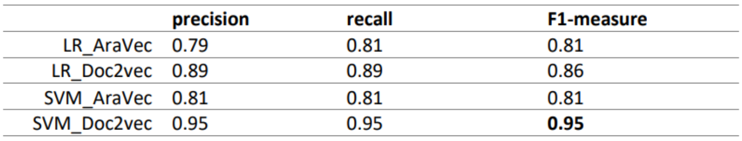
	precision	recall	F1-measure
Islamic	0.82	0.80	0.81
Modern	0.79	0.81	0.80

Table5 . Classification of our Dataset using SVM with Doc2vec



	precision	recall	F1-measure
Islamic	0.94	0.95	0.95
Modern	0.95	0.94	0.94

Table6 . Average results of using Two Machine Learning Algorithms with Two different Embedding Models



	precision	recall	F1-measure
LR_AraVec	0.79	0.81	0.81
LR_Doc2vec	0.89	0.89	0.86
SVM_AraVec	0.81	0.81	0.81
SVM_Doc2vec	0.95	0.95	**0.95**

7. CONCLUSIONS AND FUTURE WORK

In this paper, Support Vector Machine and Logistic Regression were used for the classification of Arabic poems. The machine learning algorithms proved to be good tools for text classification. From the comparison of the result of the precision, recall, and f-measure for all types of Arabic poems, the best result was found when using SVM with Doc2vec. This method of classification can be further improved for the other eras of Arabic poetry. A good approach for future work can be building a large public dataset for Arabic Poetry to use the Embedding from the transformer model.

REFERENCES

[1] Abbas, Mourad & Lichouri, Mohamed. (2019). Classification of Arabic Poems: from the 5th to the 15th Century.

[2] M. Gharbat, H. Saadeh and R. Q. Al Fayez, "Discovering The Applicability of Classification Algorithms With Arabic Poetry," 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), 2019, pp. 453-458, doi: 10.1109/JEEIT.2019.8717387.

[3] Orabi, Mariam & El Rifai, Hozayfa & Elnagar, Ashraf. (2020). Classical Arabic Poetry: Classification based on Era. 10.1109/AICCSA50499.2020.9316520.

[4] Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, “A Neural Probabilistic Language Model,” J. Mach. Learn. Res., vol. 3, pp. 1137- 1155, #mar# 2003.

[5] T. Mikolov, K. Chen, G. Corrado and J. Dean, “Efficient estimation of word representations in vector space,” arXiv preprint arXiv:1301.3781, 2013.

[6] Z Chen, Z Qi, B Wang, L Cui, F Meng, Y Shi. Learning with label proportions based on nonparallel support vector machines. Knowledge-Based Systems. 2017; 119: 126-141.

[7] X Zhang, S Ding, Y Xue. An improved multiple birth support vector machine for pattern classification. Neurocomputing. 2017; 225: 119-128.

[8] Zhang J, Jin R, Yang Y and Hauptmann A. “Modified logistic regression: an approximation to SVM and its applications in large-scale text categorization”. In: Proc Twentieth Int Conf Machine Learning (ICML 2003), Washington, DC USA, August, 2003; pp. 21–24.

[9]Wahdan, Ahlam & Hantoobi, Sendeyah & Salloum, Said & Shaalan, Khaled. (2020). A systematic review of text classification research based on deep learning models in Arabic language. 6629-6643. 10.11591/ijece.v10i6.pp6629-6643.

[10] Mahdaouy, Abdelkader & Gaussier, Eric & Ouatik El Alaoui, Said. (2016). Arabic Text Classification Based on Word and Document Embeddings. 10.1007/978-3-319-48308-5_4. [11] Abdelkader El Mahdaouy, Eric Gaussier, Saïd El Alaoui Ouatik. Arabic Text Classification Based on Word and Document Embeddings. Advances in Intelligent Systems and Computing , 533, Springer International Publishing, pp 32-41, 2016, Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2016, 978-3-319-48307-8. [12] Le, Q. &amp; Mikolov, T.” Distributed Representations of Sentences and Documents.”. Proceedings of the 31st International Conference on Machine Learning.32(2):1188-1196 . (2014)

[13] Wahbeh, Abdullah & Al-Kabi, Mohammed & Al-Radaideh, Qasem & Al-Shawakfa, Emad & Alsmadi, Izzat. (2011). The Effect of Stemming on Arabic Text Classification: An Empirical Study. International Journal of Information Retrieval Research. 1. 10.4018/IJIRR.2011070104. [14] Abu Bakr Soliman, Kareem Eisa, and Samhaa R. El-Beltagy, “AraVec: A set of Arabic Word

Embedding Models for use in Arabic NLP”, in proceedings of the 3rd International Conference on Arabic Computational Linguistics (ACLing 2017), Dubai, UAE, 2017.