International Journal of Electrical and Computer Engineering (IJECE)
Vol.8, No.6, December 2018, pp. 4352~4355
ISSN: 2088-8708, DOI: 10.11591/ijece.v8i6.pp4352-4355

A Survey of Arabic Text Classification Models

Ahed M. F. Al Sbou
Department of Computer Science, Faculty of Information Technology, Al_Hussein Bin Talal Unversity, Jordan

Article Info

Article history:

Received Nov 25, 2017
Revised Feb 17, 2018
Accepted Mar 2, 2018

Keyword:

Arabic language processing
Arabic text categorization
Arabic text mining
Classification algorithms
Clustering algorithms
Natural languages processing
Text classification

ABSTRACT

There is a huge content of Arabic text available over online that requires an organization of these texts. As result, here are many applications of natural languages processing (NLP) that concerns with text organization. One of the is text classification (TC). TC helps to make dealing with unorganized text. However, it is easier to classify them into suitable class or labels. This paper is a survey of Arabic text classification. Also, it presents comparison among different methods in the classification of Arabic texts, where Arabic text is represented a complex text due to its vocabularies. Arabic language is one of the richest languages in the world, where it has many linguistic bases. The researche in Arabic language processing is very few compared to English. As a result, these problems represent challenges in the classification, and organization of specific Arabic text. Text classification (TC) helps to access the most documents, or information that has already classified into specific classes, or categories to one or more classes or categories. In addition, classification of documents facilitate search engine to decrease the amount of document to, and then to become easier to search and matching with queries.

Corresponding Author:

Ahed M. F. Al Sbou,
Department of Computer Science, Faculty of Information Technology,
Al Hussein Bin Talal Unversity,
Rawdat Al-Amir Rashid, Ma'an, Jordan.
Email: ahed_alsbou@ahu.edu.jo

1. INTRODUCTION

The Arabic language is one of the most common languages with more than 420 million speakers over the world. Unlike English, Arabic doesn't have upper cases. It also differs from other natural languages due to the presence of diacritics which represent a small vowel letters such as "fatha, kasra, damma, sukun, shadda, and tanween". The Arabic language's orthographic system is based on diacritics effect, where each specific type of diacritics produces different words with different meanings. This language has specific letters known as Arabic vowels (waw, yaa, alf) that require a special system of morphology and grammars. What also distinguishes Arabic is the huge amount of vocabularies and concepts [1].

Although the Arabic texts are viewed as the most difficult ones, there are few studies on the processing of Arabic texts for reasons related to the linguistic characteristics of the Arabic language Due to the strict linguistic characteristics of Arabic texts and the limitations of studies on processing it [2], this study will deal with a variety of NLP applications that have recently emerged to manipulate languages such as Arabic, English, and Urdu. One of these applications is text classification (TC), which aims to make a set of documents from unstructured documents. This structured set of texts includes a description of the content of documents. TC is a process of classifying the textual document into groups based on subject's similarity or other features [3].

This paper is organized as follows. The present section offers a brief introduction to the topic and the design of the paper. The second section briefly describes the related works in the area of Arabic texts classification. The third section displays the Arabic text challenges. The fourth section provides a brief

explanation about the common techniques and algorithms used in Arabic texts classification. The last section provides a summary of the paper and suggests future work.

2. RELATED WORK

There are many classification Algorithms that have been applied to Arabic Texts. In their application of Naive Bayes (NB) algorithm to classify 1500 Arabic text documents, El-Kourdi et al find five major categories whose results indicated that the accuracy was around 68.78% [4]. Sawaf et al conducted another study based on collecting data from the Arabic NEWSWIRE corpus by using statistical methods. The results were 62.7% [5].

El-Halees et al also classified 300 Arabic documents by applying different algorithms such as vector space model (VSM), K-Nearest Neighbor algorithms (KNN), and Naïve Bayes (NB). The accuracy of the classification was 74.41% [6]. The same accuracy was obtained in Al-Zoghby's study which includes CHARM algorithm to classify Arabic text documents from 5524 records [7].

Other studies applied by Mesleh to classify Arabic document through using Support Vector Machine (SVMs) with Chi Square feature. He conducted an experimental study on 1445 online Arabic corpus that involves Al-Nahar, Al-hayat, Al-Jazeera, Al-Ahram, and Al-Dostor to be classified into 9 categories. The F-measure result was 88.11% [8]. Harrag et al developed Arabic TCs through using Hybrid approach with tree algorithm factor to select the features. The data was collected from several Arabian scientific encyclopedia in many fields. The accuracy was 91% and 93% for literary and scientific corpus, respectively [9].

3. ARABIC TEXT CHALLENGES

Natural languages such as Arabic language have been processed through different methods. This language which has several textual features requires a specific categorical environment of its morphology, concepts, and ontology. Arabic language is one of the most complex natural languages. It comprises 28 characters [1]. The characters in this language are written in different forms based on their positions in the word. The characters may come in the front, middle, or last part of the word [10]. TC seeks to collect similar documents into specific categories that assign the categories of Arabic texts, and manipulate the relative categories that have been produced from other text classifications [11].

The system of retrieving information from the large amount of Arabic texts accessible on the web is very challenging. The retrieval task of query to all relevant documents is very important to the users, too. Therefore, the TC to access the different categories makes the processes of query easier and then can attain the information needed from them [12]. Further, Arabic texts include some problematic issues due to the nature of language. To the best of my knowledge, the studies on Arabic language are very limited, in which there is a lack of Arabic corpus, language tools, and comprehensive studies on preprocessing Arabic texts. All these problems refer to diverse areas of challenges to categorize the specific Arabic textual data into a closed category.

4. TEXT CLASSIFICATION

TC includes different phases. The first phase starts from preprocessing the text to remove the punctuations, stop words, and normalization. The second and third phases include TC and evaluating the classified text [7], [11], [13]. TC is the best mechanism to manage and organize the data. It helps machine to access the data categories and text labels using predefined process [14], [15]. This mechanism can be used to classify a group of documents into kinds of documents using several features such as contents, authors, or publisher [16]. The core goal of TC is to convert unstructured text into organized or structured that can be used in different NLP applications such as summarization or retrieval [10], [17].

There are two methods utilized in TC: machine learning in which the text can be classified by using a set of training documents, and rule-based TC which allows the usage of experts, or engineer's knowledge to classify the text [18]. Furthermore, the TC can be used in several applications of computer science such as spam or e-mail filtering, or as an accessible tool for interesting information in particular documents [4], [9].

4.1. Common Models, and Algorithms of Arabic Text Classification

Different algorithms are used to classify the Arabic documents. In this section, we will focus on the following models: Naïve Bayesian algorithm (NB), K-Nearest Neighbor algorithm (KNN), Support Vector Model (SVM), Artificial Neural Network (ANN).

4.1.1. Naïve Bayesian Algorithm

NB is a machine learning technique used to classify text into predefined categories based on the similar features. NB has been applied to improve the processing and manipulating of texts or information from different sources. This algorithm represents a probabilistic method. In other words, NB classifier assumes that the absence of class feature is unrelated to the absence of other features. NB is commonly used to classify documents due to that is given a good performance in classification, NB computes the probability of documents that related to classify them into different classes, and then assigns them to the specific class with the highest probability [19].

Like many other models, NB has numerous advantages. It is generally considered the most powerful model used in this field. NB is understandable and very simple in implementation. As for the disadvantages, NB suffers some limitations such as it needs occurrence of class, because depends on probability, whereas the probability in usually depends on frequency.

4.1.2. K-Nearest Neighbor Algorithms

K-Nearest neighbor (KNN) is another type of machine learning algorithm of TC. It represents a non-parametric technique to classify documents or objects depend on closed class or training feature. It includes k value that is always a positive value; KNN the object has been classified to close neighbor class. KNN attempts to classify the object that is most vote of its neighbor [16]. KNN is possible when training data is large, and very large. Yet, there are some disadvantages of KNN including the necessity of defining k-parameter value where k represents a nearest neighbor's number. Also, using this model is so expensive in comparison with other algorithms [20].

4.1.3. Support Vector Model (SVM)

VSM is one of the supervised learning models that have been applied for TC. It classifies the different objects and documents into a finite dimensional space. VSM is also used to analyze data, texts, and documents in order to compute the similarity among them [21]. VSM shows different helpful aspects as an important model used in computer science. First, this method defends on a linear algebra, where it doesn't contain any complex algebra equation [8].

The other advantage is the efficiency of weights ascribed to concepts or terms. This model also shows a special sense of ease in comparison with other methods. It makes the machine compute the similarity among documents [22]. However, VSM contains some limitations that prevent some researchers to use it. The difficulty of using synonyms in Arabic represents a massive challenging area, where Arabic language has many synonyms for each word, or concept. Other limitations that it's assume that the terms are statistically independent. While most of Arabic terms have a strong relationship with other terms.

4.1.4. Artificial Neural Network

ANN is one of machine learning of information processing likes human brain. It has been applied in different computer areas such as classification and pattern recognition. It consists of a set of inputs and adaptive weight, non-linear function, and outputs [23]. Different advantages and disadvantages are worth mentioning for using artificial neural networks. It represents one of the easy models to use. Also, it is usually appropriate for complex problems or large texts. The disadvantages of this model include the less recommendation of using with simpler solutions or small texts. This method also needs loading the training data.

5. CONCLUSION AND FUTURE WORK

This paper is a survey of the importance of TC, as well as the current methods used in NLP field. In this research, we have discussed the traditional TC models that are used to classify the Arabic texts, corpus, and documents into different categories. The future work needs more efforts to build and develop a new standard model of Arabic TC. This model must be more efficient than the current traditional methods. The other important task that need to improvement in this model is language dialects. In other words, due to different Arabic dialects this model must be compatible with these Arabic language dialects. Also, it can be applied in any Arabic texts.

ACKNOWLEDGEMENT

We would like to thank Al_Hussein bin Talal University (AHU) for providing us a good scientific environment to produce this simple work.

REFERENCES

[1] Duwairi, R.M. (2007). Arabic text categorization. *Int. Arab J. Inf. Technol.*, 4(2), 125-132.
[2] Fauzi, M.A., Arifin, A.Z., & Yuniarti, A. (2017). Arabic Book Retrieval using Class and Book Index Based Term Weighting. *International Journal of Electrical and Computer Engineering (IJECE)*, 7(6), 3705-3711.
[3] El-Halees, A. (2006). *Mining Arabic association rules for text classification.* Paper presented at the Proceedings of the first international conference on Mathematical Sciences. Al-Azhar University of Gaza, Palestine.
[4] Caballero, Y., Bello, R., Alvarez, D., & Garcia, M.M. (2006). *Two new feature selection algorithms with Rough Sets Theory.* Paper presented at the IFIP International Conference on Artificial Intelligence in Theory and Practice.
[5] El Kourdi, M., Bensaid, A., & Rachidi, T.-e. (2004). *Automatic Arabic document categorization based on the Naïve Bayes algorithm.* Paper presented at the Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages.
[6] Al-Zoghby, A., Eldin, A.S., Ismail, N.A., & Hamza, T. (2007). *Mining Arabic text using soft-matching association rules.* Paper presented at the Computer Engineering & Systems, 2007. ICCES'07. International Conference on.
[7] Al-Harbi, S., Almuhareb, A., Al-Thubaity, A., Khorsheed, M., & Al-Rajeh, A. (2008). Automatic Arabic text classification.
[8] Mesleh, A. (2008). Support vector machines based Arabic language text classification system: feature selection comparative study *Advances in Computer and Information Sciences and Engineering* (pp. 11-16): Springer.
[9] Dharmadhikari, S.C., Ingle, M., & Kulkarni, P. (2011). Empirical studies on machine learning based text classification algorithms. *Advanced Computing*, 2(6), 161.
[10] Khan, A., Baharudin, B., Lee, L.H., & Khan, K. (2010). A review of machine learning algorithms for text-documents classification. *Journal of advances in information technology*, 1(1), 4-20.
[11] Ababneh, J., Almomani, O., Hadi, W., El-Omari, N.K.T., & Al-Ibrahim, A. (2014). Vector space models to classify Arabic text. *International Journal of Computer Trends and Technology (IJCTT)*, 7(4), 219-223.
[12] Mesleh, A. (2007). Chi square feature extraction based svms arabic language text categorization system. *Journal of Computer Science*, 3(6), 430-435.
[13] Khorsheed, M.S., & Al-Thubaity, A.O. (2013). Comparative evaluation of text classification techniques using a large diverse Arabic dataset. *Language resources and evaluation*, 47(2), 513-538.
[14] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1), 1-47.
[15] Khreisat, L. (2009). A machine learning approach for Arabic text classification using N-gram frequency statistics. *Journal of Informetrics*, 3(1), 72-77.
[16] Alaa, E. (2008). A comparative study on arabic text classification. *Egypt. Comput. Sci. J*, 2.
[17] Mesleh, A. (2008). Support Vector Machine Text Classifier for Arabic Articles: Ant Colony Optimization-Based Feature Subset Selection. *The Arab Academy for Banking and Financial Sciences*.
[18] Sebastiani, F. (2005). Text categorization *Encyclopedia of Database Technologies and Applications* (pp. 683-687): IGI Global.
[19] Abu-Errub, A. (2014). Arabic Text Classification Algorithm using TFIDF and Chi Square Measurements. *International Journal of Computer Applications*, 93(6).
[20] Al-Shalabi, R., Kanaan, G., & Gharaibeh, M. (2006). *Arabic text categorization using KNN algorithm.* Paper presented at the the Proc. of Int. multi conf. on computer science and information technology CSIT06.
[21] Jackson, P., & Moulinier, I. (2007). *Natural language processing for online applications: Text retrieval, extraction and categorization* (Vol. 5): John Benjamins Publishing.
[22] Sawaf, H., Zaplo, J., & Ney, H. (2001). Statistical classification methods for Arabic news articles. *Natural Language Processing in ACL2001, Toulouse, France*.
[23] Harrag, F., El-Qawasmeh, E., & Pichappan, P. (2009). *Improving Arabic text categorization using decision trees.* Paper presented at the Networked Digital Technologies, 2009. NDT'09. First International Conference on.

BIOGRAPHY OF AUTHOR

Ahed Al-Sbou is a lecturer in the Information Technology School of Computer Science at the University of Al-Hussein Bin Talal where he has been a faculty member since 2014. He holds the master degree in computer science. Ahed completed his Master degree from Al-Balqa Applied University, Salt, Jordan in 2012 and his B.S. degree in computer science from Al-Hussein Bin Talal University, Ma'an, Jordan in 2006. His research interests lie in computer science are in the area of programming languages, ranging from theory to design to implementation, Data base, Data Mining, Natural languages Processing (NLP), and information systems. Ahed has worked as a computer lab supervisor (2006-2014) at Al-Hussein Bin Talal University.

.