Arabic Text Clustering

Subjects: Computer Science, Artificial Intelligence

Contributor: Souad Larabi-Marie-Sainte , Mashael Bin Alamir , Abdulmajeed Alameer

Arabic text clustering is an essential topic in Arabic Natural Language Processing (ANLP). Its significance resides in various applications, such as document indexing, categorization, user review analysis, and others.

natural language processing | text clustering | Arabic text

1. Introduction

Clustering text documents is an important field in the area of Natural Language Processing (NLP) as it simplifies the tedious process of categorizing specific documents among millions of resources, especially when metadata such as key phrases, titles, and labels are not available. Text clustering is valuable for different applications, including topic extraction, spam filtering, automatic document categorization, user reviews analysis, and fast information retrieval.

The process of clustering text written in natural languages is complicated, especially for the Arabic language. One of the complications in Arabic is the language's morphological complexity. For instance, a word in Arabic can be written in several forms that might exceed ten forms [1]. Ambiguity is also another major complication in the Arabic language, which is caused by the richness and complexity of Arabic morphology [1][2]. There are various other factors in the Arabic language causing difficulty in text clustering. Among these factors are the different dialects for different regions. Texts from different regions may exhibit significant linguistic variations. Moreover, in the Arabic language, the ordering of words in a sentence provides quite different interpretations for that sentence [3][4].

Several Arabic text clustering techniques have been proposed by researchers to encounter these challenges. Among the various techniques, it has been concluded that the K-Means clustering algorithm is the most widely applied, and that is due to its simplicity and efficiency in comparison with other clustering algorithms [2][5][6][7]. However, the initiation process of K-Means weakens its accuracy results. The initiation starts with plotting the centers of the clusters randomly and then assigning documents to the nearest center. If the initiation process is inaccurate, then the clustering will be imprecise [8]. Researchers proposed the use of K-Means++, which is an improved algorithm for the initialization process of K-Means [9]. However, the experiments show that even with this smart initialization process, the accuracy of the clustering is low compared to other techniques. Researchers also proposed the use of other clustering techniques, such as Suffix Tree clustering [10] and Self-Organizing Maps (SOM) [11]. Suffix Tree clustering has a limitation of overlapping documents in different clusters [12], while SOM clustering techniques demonstrated high effectiveness in clustering text even with high-dimensional datasets [13][14][15][16][17].

2. Arabic Text Clustering

Alharwat and Hegazi demonstrated the issue of data mining and data with high dimensions [18]. To overcome the addressed problem, the authors applied modeling techniques to the documents before clustering them. The scholars used the Modern Standard Arabic (MSA) dataset [19], which has several versions with different preprocessed articles. The outcome of the study showed that normalized data provided better quality in clustering than unnormalized ones. With normalization, the purity of their clusters was 0.933, and the F1-score was 0.8732. Similar to Alharwat and Hegazi, Al-Azzawy et al. used K-Means to cluster an Arabic dataset corpus which contains 20 documents related to news and short anecdotes [20]. The highest clustering scores for the precision, recall, and F1-measure were 98%, 88%, and 93%, respectively. Mahmood and Al-Rufaye also addressed the problem of the high dimensionality of documents by minimizing the dimensionality of documents using the Term Frequency (TF), Inverse Document Frequency (IDF), and Term Frequency–Inverse Document Frequency (TF-IDF) feature selection approaches [21]. Following that, K-Means and K-Medoids were used for the clustering. The authors implemented their experiment on a 300-document corpus they built. The authors reported that K-Medoids provided more accurate results than K-Means; the first scored 60%, 78%, and 67% for the precision, recall, and F1-measure, respectively, while the second scored 80%, 83%, and 81%, respectively. Another group of researchers used K-Means clustering along with the TF-IDF and Binary Term Occurrence (BTO) feature selection approaches [22]. The scholars used a dataset that contains 1121 Arabic tweets. The outcome of their work showed that the BTO feature selection approach outperformed the TF-IDF. The literature for clustering Arabic text using K-Means shows high variation in the performance scores for clustering Arabic text, which could be attributed to the instability and inconsistency of the K-Means clustering algorithm.

To overcome the limitations of the K-Means random initiation of cluster centroids, researchers used PSO-optimized K-Means to cluster Arabic text [23][24][25]. The use of Particle Swarm Optimization (PSO) contributes to selecting the initial seeds of K-Means. A group of researchers implemented their algorithm for the purpose of Quran verses theme clustering [23], whereas another group [24][25] used three different datasets, named BBC, CNN, and OSAC [26]. The outcome of these research papers demonstrated the effectiveness of applying optimization methods for enhancing the accuracy of the clustering models used.

Another work on clustering Arabic documents was based on the sentiment orientation and context of words in the data corpus [5]. The authors used the Brown clustering algorithm on user reviews of several topics, such as news, movies, and restaurants. The data in the research were collected from several sources [27][28][29][30]. The evaluation results of the approach showed that the subjectivity and polarity of the clustering documents provided rates of 96% and 85%, respectively. The evaluation results indicated that the number of clusters also affects the accuracy rates, showing that fewer clusters provide better results.

In another work [2], the authors used a combination of Markov Clustering, Fuzzy-C-Means, and Deep Belief Neural Networks (DBN) in an attempt to cluster Arabic documents. Two datasets were used in the study; the first was acquired from the Al-Jazeera news website with 10,000 documents and the second from a Saudi Press Agency [31] with 6000 documents. The clustering precision, recall, and F1-measure resulted in 91.2%, 90.9%, and 91.02%,

respectively. The model that was used was highly impacted by the feature selection of the root words leading to imprecise clustering results.

Al-Anzi and Abuzeina [11] used Expectation-Maximization (EM), SOM, and K-Means algorithms to cluster Arabic documents. They built a corpus of 1000 documents extracted from a Kuwaiti newspaper website called Alanba [32]. The documents cover different topics, such as health, technology, sports, politics, and others. The authors then compared the evaluation of the three clustering algorithms. They reported that SOM obtained the highest accuracy between the three algorithms with a rate of 93.4%. From the study, it appears that the use of SOM in clustering Arabic text is promising.

The Bond Energy Algorithm (BEA) was also used by researchers to cluster Arabic text [33]. The results of the study showed that the BEA algorithm outperforms K-Means clustering in terms of precision, recall, and the F1-score.

In the broader field of text clustering, researchers also proposed the use of prototype-based models for text clustering [34]. The results of the work showed that it outperforms K-Means clustering.

To conclude, most of the current work on Arabic text clustering used K-Means clustering because it is a simple model and can be applied easily. However, the mechanism that K-Means follows has limitations. For instance, K-Means first initiates centers of clusters and then assigns documents to these clusters. If the initiation process of K-Means is not well formulated, then the risk of incorrect clustering arises. Moreover, techniques that integrate K-Means clustering with Particle Swarm Optimization [25] have promising results. This shows that optimization contributes positively to clustering models. In addition, previous work showed that the use of SOM provided better clustering results than K-Means for Arabic text [11]. **Table 1** presents a summary of the recent work regarding Arabic text clustering.

Table 1. Arabic text clustering related work comparison.

Ref.	Model	Dataset	Purity	F1-Score	Precision	Recall	Accuracy
[18]	K-Means	MSA	93.3%	87.32%	87.13%	87.52%	-
[20]	K-Means	Own corpus	-		93%	98%	88%
[25]	K-Means + (PSO)	BBC, CNN, OSAC	50%	47%	33%	-	-
[5]	Brown clustering algorithm	Own corpus	85%	-	-	-	-
[24]	K-Means	Arabic tweets	76.4%	-	-	-	-
[22]	TF-IDF + BTO	Arabic tweets	-	-	-	-	-
[21]	K-Medious	Own corpus	-	67%	60%	78%	-

Ref.	Model	Dataset	Purity	F1-Score	Precision	Recall	Accuracy
[2]	Markov + Fuzzy-C-Means + DBN	Own corpus	-	91.02%	91.02%	90.9%	-
[10]	Suffix Tree	Own corpus	-	81.11%	80.3%	83.75%	-
[11]	SOM	Own corpus	-	-	-	-	93.4%

1. Farghaly, A.; Shaalan, K. Arabic natural language processing: Challenges and solutions. ACM Trans. Asian Lang. Inf. Process. (TALIP) 2009, 8, 14.
2. Jindal, V. A Personalized Markov Clustering and Deep Learning Approach for Arabic Text Categorization. In Proceedings of the ACL 2016 Student Research Workshop, Berlin, Germany, 7–12 August 2016; pp. 145–151.
3. Habash, N.Y. Introduction to Arabic natural language processing. Synth. Lect. Hum. Lang. Technol. 2010, 3, 1–187.
4. Wenchao, L.; Yong, Z.; Shixiong, X. A novel clustering algorithm based on hierarchical and K-means clustering. In Proceedings of the Control Conference, Zhangjiajie, China, 26–31 July 2007; pp. 605–609.
5. Alotaibi, S.; Anderson, C. Word Clustering as a Feature for Arabic Sentiment Classification. IJ Educ. Manag. Eng. 2017, 1, 1–13.
6. Goodfellow, I.; Bengio, Y.; Courville, A.; Bengio, Y. Deep Learning; MIT Press: Cambridge, UK, 2016; Volume 1.
7. Kriesel, D. A Brief Introduction on Neural Networks. 2007. Available online: https://www.dkriesel.com/en/science/neural_networks (accessed on 26 July 2023).
8. Mahdavi, M.; Abolhassani, H. Harmony K-means algorithm for document clustering. Data Min. Knowl. Discov. 2009, 18, 370–391.
9. Arthur, D.; Vassilvitskii, S. K-Means++: The Advantages of Careful Seeding. In Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA'07, New Orleans, LA, USA, 7–9 January 2007; pp. 1027–1035.
10. Sahmoudi, I.; Lachkar, A. Towards a linguistic patterns for arabic keyphrases extraction. In Proceedings of the Information Technology for Organizations Development (IT4OD), Fez, Morocco, 30 March–1 April 2016; pp. 1–6.
11. Al-Anzi, F.S.; AbuZeina, D. Big data categorization for arabic text using latent semantic indexing and clustering. In Proceedings of the International Conference on Engineering Technologies and Big Data Analytics (ETBDA 2016), Bangkok, Thailand, 21–22 January 2016; pp. 1–4.

12. Pujari, P.S.; Waghmare, A. A Review of Merging based on Suffix Tree Clustering. In Proceedings of the National Conference on Advances in Computing, Roorkee, India, 13–15 February 2020.
13. Cottrell, M.; Olteanu, M.; Rossi, F.; Villa-Vialaneix, N. Self-organizing maps, theory and applications. Rev. Investig. Oper. 2018, 39, 1–22.
14. Yoshioka, K.; Dozono, H. The classification of the documents based on Word2Vec and 2-layer self organizing maps. Int. J. Mach. Learn. Comput. 2018, 8, 252–255.
15. Yang, H.C.; Lee, C.H.; Wu, C.Y. Sentiment discovery of social messages using self-organizing maps. Cogn. Comput. 2018, 10, 1152–1166.
16. Gunawan, D.; Amalia, A.; Charisma, I. Clustering articles in bahasa indonesia using self-organizing map. In Proceedings of the 2017 International Conference on Electrical Engineering and Informatics (ICELTICs), Banda Aceh, Indonesia, 18–20 October 2017; pp. 239–244.
17. Liu, Y.C.; Liu, M.; Wang, X.L. Application of Self-Organizing Maps in Text Clustering: A Review; IntechOpen: London, UK, 2012; Volume 10.
18. Alhawarat, M.; Hegazi, M. Revisiting K-Means and Topic Modeling, a Comparison Study to Cluster Arabic Documents. IEEE Access 2018, 6, 42740–42749.
19. Abuaiadh, D. Dataset for Arabic Document Classification. 2014. Available online: http://diab.edublogs.org/dataset-for-arabic-document-classification (accessed on 26 July 2023).
20. Al-Azzawy, D.S.; Al-Rufaye, F.M.L. Arabic words clustering by using K-means algorithm. In Proceedings of the New Trends in Information & Communications Technology Applications (NTICT), Baghdad, Iraq, 7–9 March 2017; pp. 263–267.
21. Mahmood, S.; Al-Rufaye, F.M.L. Arabic text mining based on clustering and coreference resolution. In Proceedings of the Current Research in Computer Science and Information Technology (ICCIT), Sulaymaniyah, Iraq, 26–27 April 2017; pp. 140–144.
22. Al-Rubaiee, H.; Alomar, K. Clustering Students' Arabic Tweets using Different Schemes. Int. J. Adv. Comput. Sci. Appl. 2017, 8, 276–280.
23. Bsoul, Q.; Atwan, J.; Salam, R.A.; Jawarneh, M. Arabic Text Clustering Methods and Suggested Solutions for Theme-Based Quran Clustering: Analysis of Literature. J. Inf. Sci. Theory Pract. (JISTaP) 2021, 9, 15–34.
24. Abuaiadah, D.; Rajendran, D.; Jarrar, M. Clustering Arabic tweets for sentiment analysis. In Proceedings of the Computer Systems and Applications (AICCSA), Hammamet, Tunisia, 30 October–3 November 2017; pp. 449–456.
25. Daoud, A.S.; Sallam, A.; Wheed, M.E. Improving Arabic document clustering using K-means algorithm and Particle Swarm Optimization. In Proceedings of the Intelligent Systems Conference (IntelliSys), London, UK, 7–8 September 2017; pp. 879–885.

26. Saad, M.K.; Ashour, W.M. OSAC: Open source arabic corpora. In Proceedings of the 6th International Conference on Electrical and Computer Systems (EECS'10), Lefke, North Cyprus, 25–26 November 2010; Volume 10.
27. Newspaper, E.S. Electronic Sabq Newspaper. 2020. Available online: https://sabq.org/ (accessed on 26 July 2023).
28. Souq Online Shopping. 2022. Available online: https://saudi.souq.com/sa-en/ (accessed on 26 July 2023).
29. Al-Subaihin, A.A.; Al-Khalifa, H.S.; Al-Salman, A.S. A proposed sentiment analysis tool for modern arabic using human-based computing. In Proceedings of the 13th International Conference on Information Integration and Web-Based Applications and Services, Ho Chi Minh City, Vietnam, 5–7 December 2011; pp. 543–546.
30. Farra, N.; Challita, E.; Assi, R.A.; Hajj, H. Sentence-level and document-level sentiment mining for arabic texts. In Proceedings of the Data Mining Workshops (ICDMW), Ho Chi Minh City, Vietnam, 5–7 December 2010; pp. 1114–1119.
31. Al-Harbi, S.; Almuhareb, A.; Al-Thubaity, A.; Khorsheed, M.; Al-Rajeh, A. Automatic Arabic text classification. In Proceedings of the 9th International Conference on the Statistical Analysis of Textual Data, Lyon, France, 12–14 March 2008.
32. Alanba News. 2022. Available online: https://www.alanba.com.kw (accessed on 26 July 2023).
33. Alazzam, H.; AbuAlghanam, O.; Alsmady, A.; Alhenawi, E. Arabic Documents Clustering using Bond Energy Algorithm and Genetic Algorithm. In Proceedings of the 2022 13th International Conference on Information and Communication Systems (ICICS), Irbid, Jordan, 21–23 June 2022; pp. 4–8.
34. Zeng, J.; Yin, Y.; Jiang, Y.; Wu, S.; Cao, Y. Contrastive Learning with Prompt-derived Virtual Semantic Prototypes for Unsupervised Sentence Embedding. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 7042–7053.