Clustering Students' Arabic Tweets using Different Schemes

Hamed Al-Rubaiee
Department of Computer Science and Technology,
University of Bedfordshire
Bedfordshire, United Kingdom

Khalid Alomar
Department of Information Systems
King Abdulaziz University
Jeddah, Kingdom of Saudi Arabia

***Abstract*—In this paper, Twitter has been chosen as a platform for clustering the topics that have been mentioned by King Abdulaziz University students to understand students' behaviours and answer their inquiries. The aim of the study is to propose a model for clustering analysis of Saudi Arabian (standard and Arabian Gulf dialect) tweets to segment topics included in the students' posts. A combination of the natural language processing (NLP) and the machine learning (ML) method to build models is used to cluster tweets according to their text similarity. K-mean algorithm is utilised with different vector representation schemes such as TF-IDF (term frequency-inverse document frequency) and BTO (binary-term occurrence). Distinct preprocessing is explored to obtain the N-grams term of tokens. The cluster distance performance task is applied to determine the average between the centroid clusters. Moreover, human evaluation clustering is performed by looking at the data source to make sure that the clusters are making sense to an educational domain. At this moment, each cluster has been identified, and students' accounts on Twitter have been known by their facilities or their educational system, such as e-learning. The results show that the best vector's representation was using BTO, and it will be useful to apply it to cluster students' text instead of the TF-IDF scheme.**

***Keywords*—Twitter; Arabic tweets; Saudi Arabia; King Abdulaziz University; data mining; data preparation**

I. Introduction

Today, students use university social media accounts on a daily basis to follow university events, to exchange thoughts, and to express their views and activities [1, 2]. In addition, some students consider social media such as Twitter as the first option for sending their inquiry and complaint. This is especially relevant for online distance education students, who depend mostly on the web. Therefore, it comes as no surprise that universities have started to study students' behaviours on social media and analyse their opinions for the purpose of improving university-provided services, teaching processes, and learning experiences [3, 4].

For Twitter accounts that have extraordinarily large numbers of followers (such as the Deanship of e-Learning and Distance Education at King Abdulaziz University), answering individual tweets is time consuming.

Therefore, the aim of this study is to group similar tweets into groups such that tweets within the same group bear similarity with each other, while tweets in different groups are dissimilar with each other. This will help the university understand students' behaviours, find out the most common problems, and contact students within the same group and answer their inquiry faster.

Data mining (DM) methods have been utilised to investigate the field of educational data mining, such as association rule mining, sequence pattern, visualisation, and clustering [5-7]. Unlike most studies, the authors believe that social media has a lot of 'informative information' mentioned by students in their Twitter accounts. Moreover, these pieces of information or inquiries are sent to the university's account. For this reason, this study will concentrate on text mining and text similarity to group students' accounts that contain similar tweets.

DM techniques were utilised to make recommendations directly to students with respect to their personalised activities [8]. The most popular methods are association rule [9], sequential pattern mining [9], and clustering [10]. For this study, the clustering method was chosen, as clustering organises similar objects into groups [11]. Clustering is utilised in different areas, such as web mining [12], document grouping [13], and bioinformatics [14]. The basic clustering technique depends on the following steps: data representation model, similarity attribute measure, clustering technique, and finally, validation [15].

The rest of this paper is organised as follows: In section 2, related work is presented. Section 3 shows the methodology used and the process of clustering analysis of Arabic tweets. Section 4 presents the experiment's results and evaluations. The final section is composed of conclusions, remarks, and future works.

II. Related Work

Perera et al. proposed an affective technique to categorise both teams and similar individual members. In their approach to mining and extracting of a sequence of frequent events in learning data, their method was a sequential pattern mining. Their conclusions indicate the consequence of leadership and group interaction and give a hopeful suggestion about whether events are taking place [16].

Tang and McCalla applied a clustering method based on large generalised sequences to examine a group of students, depending on their similar characteristics with web-based learning environments [17]. Chen et al. utilised a K-means clustering algorithm to aggregate students who showed identical behaviour in an e-learning environment [18].

Moreover, they identified the formative element in assessment rules based on web-based learning. In addition, teachers could conduct specific assessments per the learning portfolio of particular students in their learning environment. Vukícevíc et al. addressed their conception of research to utilise the clustering approach for class retrieval from microarray data [19]. They reported that there is no recommendation for the selection of the right parameters and distance measure, which can improve the clustering model.

Jovanovic et al built models that can help to identify students that require extra attention in their courses [20]. Their modules involved two important data mining tasks: first, a classification model to predict if students will perform well in different courses based on web utilization; second, a clustering model that groups students with similar cognitive styles but different levels of prosperity. They concluded that analysis of student behaviours on the educational system has true potential and is a worthy investment.

Perera et al stated that group work is encouraged by online implements; their goal was to empower departments and their facilitators to see applicable parts of a group's operation and give input as to whether these are more likely to be related to positive or negative outcomes [16]. They gathered their information from understudies who were working in groups and using an online tool in a software development project and utilized clustering and sequential pattern mining approaches: clustering was connected to discover comparative groups of similar teams and similar individual members, and sequential pattern mining was acclimated to concentrate sequences of continuous occasions. Their outcomes uncovered significant patterns characterizing the work of stronger and weaker students.

Nilsson and Liu collected and analysed web data and log files [21]. In their system, they utilized Canopy Generation with k-mean to devise a method for grouping college training materials in view of their text content features, integrating it into a spring web application. They represented their data vectors using TF-IDF schemes. Their developed framework is exceptionally adaptable and likely able to coordinate custom channels and supersede grouping calculations as required. They found that there is no basic method for assessing how well a clustering algorithms performs as a rule.

III. METHODOLOGY

Figure 1 outlines the grouping procedure for Arabic tweets. As one of the best-known social networks, Twitter has been chosen for this review as a subject for clustering. Different feature selection schemes are used, such as TF-IDF (term frequency-inverse document frequency) and BTO (binary-term occurrence). TF-IDF describes how important a word is for a document [22] [23] [24]. It consists of two parts: term frequency (TF) and invert document frequency (IDF). In addition, BTO is defined as the binary value. In other words, a word or term gets 1 if present in a document, 0 otherwise.

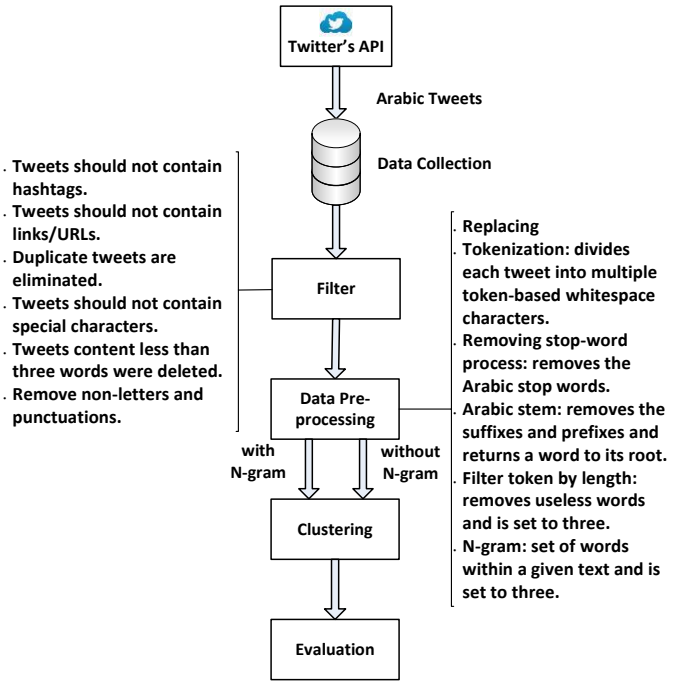

Fig. 1. The process of clustering analysis in Arabic tweets

A. Data Collection

To collect the corpus of data with Arabic tweets and save the relevant tweets and discard the irrelevant ones, a small desktop application (Twitter Data Grabber) is developed using C# with Twitter's official developer API. The tweets' source data were obtained from the Deanship of e-Learning and Distance Education at King Abdulaziz University in Saudi Arabia [25]. The tweets were collected in 50 days. The dataset included 1,121 tweets. Preliminary data were collected and contained many attributes, such as tweet ID, tweet original, tweet filtered, tweet time, and tweet user). This study has used only tweet filtered, data-type text. Only the tweet user was used to determine which cluster a tweet belongs to.

B. Data Preprocessing

Data preparation consists of different techniques that are concerned with the analysis of raw data to yield quality data, mainly including data collection, data integration, data transformation, data cleaning, data reduction, and data discretisation [26]. Data preparation is more time-consuming than data mining and equally, if not more challenging, than data mining. For instance, Arabic is a Semitic language with composite morphology; Arabic words are classified as particles, nouns, or verbs; Arabic is written from right to left; Arabic sentences do not begin with a capital letter as in many other languages such as English; and the 28 Arabic characters differ in shape depending on position in a word or sentence [15, 27-29].

The process begins with our system of filtering the original tweets and saving them in the tweet filtered field, with all the

collected tweets saved in the main dataset. Afterward, these tweets are filtered to decide which ones will move to the next stage. The filtering process is performed on specific criteria. For example, duplicate tweets are eliminated. Any tweet containing fewer than three words was deleted. Tweets could not contain hashtags, special characters, nonletters, punctuations, or links/URLs.

There are five steps for preprocessing documents using RapidMiner. The first is to replace English terms with Arabic equivalent terms; all hamza forms such as ئ, ؤ, ء are removed because of diacritics and replaced with initial أ, آ, إ, ى and with alif. ا. The second is the tokenisation step, in which each tweet is divided into multiple token-based whitespace characters. The third is removing the Arabic stop words. In the fourth step, the suffixes and prefixes are removes, and words are returned to their root in the Arabic stem. The fifth step is the filtering of token by length, which removes useless words and is set to three. Finally, the N-gram set of words within a given text is set to two.

Then the "process documents from the data" operator generate a word vector from the dataset after preprocessing and represent the text data as a matrix to show the frequent occurrence of each term. K-means is applied with weighting schemes such as TF-IDF and BTO. The evaluation is conducted using clustering distance methods.

C. Clustering methods

Clustering techniques are applied when there is no class to be predicted. Clustering is a type of unsupervised learning, which is used when you have unlabelled data. The classical clustering technique is called K-means. The following steps describe how k-means works [30]:

1) The user determines in advance how many clusters are needed by the set-up *k* parameter.

2) Cluster centres are randomly initialised using K points.

3) Clusters are assigned to their closest centre according to the Euclidean distance metric.

4) Centroids are moved by recalculating the positions of the centroid of the instances in each cluster.

5) Steps 3 and 4 are reiterated until the centroids no longer move.

IV. Experiment

This experiment examines Arabic text clustering in e-learning as an educational pattern with K-means. To quantify the homogeneous attribute among the dataset, distance metrics plays a very paramount role. Identifying the manner in which datasets are interrelated, how distinct data are dissimilar or identical with each other, and which quantifications are utilised for comparison is obligatory. For this purpose, the distance metrics function is calculated based on which data are clustered. This study was evaluated for effectiveness within a cluster distance. It is a task that takes this centroid cluster model and clustered set as input and estimates the leverage of the model based on the cluster centroids [31, 32].

V. Finding and Discussion

The experiments started by setting up k with 3 and then running the process. Then, the dataset includes 1,121 tweet clusters. Using TF-IDF, Table 1 shows that the average within centroid distance is –0.943.

TABLE. I. Performance within Cluster Distance using TF-IDF

Cluster	Total number of records	Centroid distance between clusters
0	804	–6.435
1	139	–0.899
2	178	–0.851
Avg. within centroid distance: –0.943		

The dataset tweets cluster distance using TF-IDF with N-gram = 2. Table 2 shows that the average within centroid distance is –0.972.

TABLE. II. Performance within Cluster Distance using TF-IDF with N-Gram = 2

Cluster	Total number of records	Centroid distance between clusters
0	823	–0.983
1	200	–0.967
2	98	0.891
Avg. within centroid distance: –0.972		

The dataset tweets cluster distance using binary-term occurrences. Table 3 shows that the average within centroid distance is –6.435.

TABLE. III. Performance within Cluster Distance using Binary-Term Occurrences

Cluster	Total number of records	Centroid distance between clusters
0	812	–5.868
1	245	–7.865
2	64	–8.157
Avg. within centroid distance: –6.435		

The dataset tweets cluster distance using binary-term occurrences with N-gram equal to 2. Table 4 shows that the average within centroid distance is –12.819.

TABLE. IV. Clustering Distance using Binary-Term Occurrences with N-Gram Equal to 2

Cluster	Total number of records	Centroid distance between clusters
0	876	–11.756
1	219	–16.278
2	26	–19.507
Avg. within centroid distance: –12.819		

Table 5 shows the difference between the average within centroid distances for the schemas TF-IDF and BTO without N-gram equal to –5.492. In addition, it shows the difference between the average within centroid distances for the schemas TF-IDF and BTO with N-gram equal to –11.847.

TABLE. V. AVERAGE WITHIN CENTROID DISTANCE

	Without N-gram	Within N-gram
TF-IDF	–0.943	–0.972
BTO	–6.435	–12.819
Total Avg. within centroid distance	–5.492	–11.847

In conclusion, the best clustering obtained were the smaller distances within centroids. The reason for multiplying by –1 is to calculate the average distance within centroids. Theoretically, the smaller the distances are, the better the clusters are. Performance within the cluster distance operator maximises the capacity of the k-means algorithm. On the other hand, if it is not multiplied by –1, the distance evaluation calculates with a higher average distance within centroids. So clear representation of our vectors using BTO will be useful to determine similarity with students' text than the TF-IDF scheme. In addition, using the N-gram feature provides the best and most reliable way to group tweets in this study.

Table 6 shows the subtraction between the recorders or the tweets that have been grouped or clustered by comparing and carefully reading only the disagreement tweets.

TABLE. VI. CLUSTERING DIFFERENTIATION WITHIN TWEETS

Cluster	BTO	TF-IDF	Different
0	876	823	53
1	219	200	19
2	26	98	72

Figure 2 shows that there were variance subjects or inquiries involved in the data; for example, cluster zero has three inquiries: exams, payment methods, and university study plan. Cluster 1 has only one subject, almost all of which is the Blackboard system and its problems and inquiries. In cluster 2, students talk about the next year's studies or their summer semesters. However, it was difficult even for the team to make a difference about the missed clustering that comes across those methods.

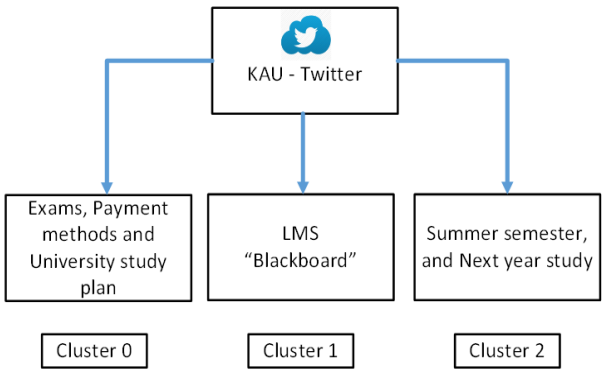

Fig. 2. Grouping tweets into three main subjects

Figure 3 shows the average within centroid distance between TF-IDF and BTO without the N-gram feature.

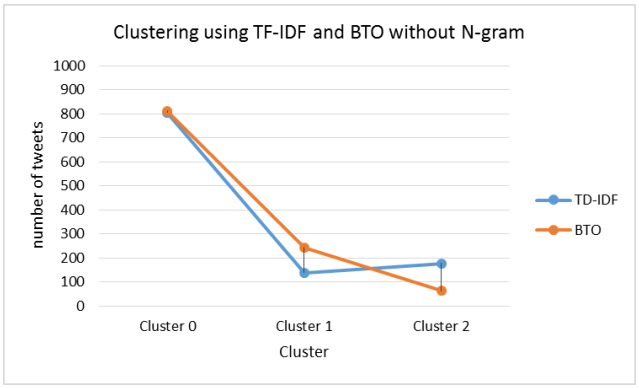

Fig. 3. Cluster Distance using TF-IDF and BTO without N-gram

Figure 4 shows the average within centroid distance between TF-IDF and BTO with the N-gram feature.

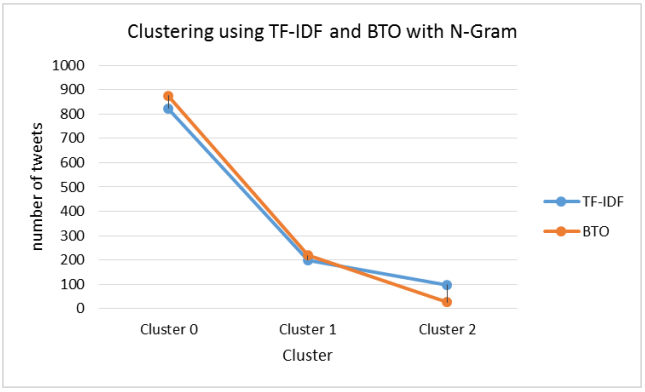

Fig. 4. Cluster distance using TF-IDF and BTO with N-gram

Finally, the university team observed our data scrupulously when the N-gram feature, which was engaged in both experiments because there was clear improvement within the average centroid distance.

VI. CONCLUSION AND FUTURE WORKS

In this paper, the design and implementation of Arabic text clustering over social media was presented to King Abdulaziz University students. The authors believe that this study demonstrates that text preprocessing is a consequential mission to prove that different generating vectors will affect the clustering of the targeted groups. The results of the two experiments lead us to conclude that the best clustering work were the smaller distances within centroids. It was achieved with BTO using the N-gram feature. On the other hand, humans interfered by looking carefully into our data when the N-gram feature was involved in both experiments. It was difficult to make a judgment of which schemes are better in representing this study corps. Nevertheless, the results of the two studies led us to conclude that Twitter can be used in education to provide such services as answering a group of students' inquiries.

For future research, we intend to integrate students' Twitter accounts through which they posted their tweets on the

Deanship of e-Learning and Distance Education account at King Abdulaziz University in the Tweeter platform to collect their data in the learning management system (Blackboard) to see the correlation between the data on social media and their real stations and grads.

ACKNOWLEDGMENTS

The first author would like to thank both the Saudi Arabian Cultural Bureau in London and the Deanship of e-Learning and Distance Education at King Abdulaziz University for their constructive comments and support in collecting the required data.

REFERENCES

[1] M. Cuesta, M. Eklund, I. Rydin, and A.-K. Witt, "Using Facebook as a co-learning community in higher education," Learning, Media and Technology, vol. 41, pp. 55-72, 2016.

[2] N. Drašković, A. K. Korper, and K. Kilian-Yasin, "Student attitudes toward use of social media in the learning process: A comparative study of Croatian and German students," International journal of management cases, vol. 19, pp. 53-64, 2017.

[3] S. Manca and M. Ranieri, "'Yes for sharing, no for teaching!': social media in academic practices," The Internet and Higher Education, vol. 29, pp. 63-74, 2016.

[4] J. Lim and J. C. Richardson, "Exploring the effects of students' social networking experience on social presence and perceptions of using SNSs for educational purposes," The Internet and Higher Education, vol. 29, pp. 31-39, 2016.

[5] P. Bansal, "EDUCATIONAL DATA MINING: A review of the state-of-the-art," 2014.

[6] C. Romero and S. Ventura, "Educational data mining: A survey from 1995 to 2005," Expert systems with applications, vol. 33, pp. 135-146, 2007.

[7] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," JEDM-Journal of Educational Data Mining, vol. 1, pp. 3-17, 2009.

[8] G. Dong and J. Pei, Sequence data mining vol. 33: Springer Science & Business Media, 2007.

[9] R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proc. 20th int. conf. very large data bases, VLDB, 1994, pp. 487-499.

[10] A. K. Jain and R. C. Dubes, Algorithms for clustering data: Prentice-Hall, Inc., 1988.

[11] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," IEEE transactions on pattern analysis and machine intelligence, vol. 24, pp. 881-892, 2002.

[12] R. Kosala and H. Blockeel, "Web mining research: A survey," ACM Sigkdd Explorations Newsletter, vol. 2, pp. 1-15, 2000.

[13] W. Ayadi, M. Elloumi, and J.-K. Hao, "BicFinder: a biclustering algorithm for microarray data analysis," Knowledge and Information Systems, vol. 30, pp. 341-358, 2012.

[14] E. Baralis, G. Bruno, and A. Fiori, "Measuring gene similarity by means of the classification distance," Knowledge and information systems, vol. 29, pp. 81-101, 2011.

[15] K. M. Hammouda and M. S. Kamel, "Efficient phrase-based document indexing for web document clustering," IEEE Transactions on knowledge and data engineering, vol. 16, pp. 1279-1296, 2004.

[16] D. Perera, J. Kay, I. Koprinska, K. Yacef, and O. R. Zaïane, "Clustering and sequential pattern mining of online collaborative learning data," IEEE Transactions on Knowledge and Data Engineering, vol. 21, pp. 759-772, 2009.

[17] T. Y. Tang and G. McCalla, "Student modeling for a web-based learning environment: a data mining approach," in AAAI/IAAI, 2002, pp. 967-968.

[18] C.-M. Chen, M.-C. Chen, and Y.-L. Li, "Mining key formative assessment rules based on learner profiles for web-based learning systems," in Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on, 2007, pp. 584-588.

[19] M. Vukićević, K. Kirchner, B. Delibašić, M. Jovanović, J. Ruhland, and M. Suknović, "Finding best algorithmic components for clustering microarray data," Knowledge and information systems, vol. 35, pp. 111-130, 2013.

[20] M. Jovanovic, M. Vukicevic, M. Milovanovic, and M. Minovic, "Using data mining on student behavior and cognitive style data for improving e-learning systems: a case study," International Journal of Computational Intelligence Systems, vol. 5, pp. 597-610, 2012.

[21] N. Nilsson and Y. Liu, "Adopting Text Clustering in web-based application to facilitate searching of education information," in Software Engineering and Service Science (ICSESS), 2014 5th IEEE International Conference on, 2014, pp. 393-396.

[22] G. Tripathi and S. Naganna, "Feature selection and classification approach for sentiment analysis," Machine Learning and Applications: An International Journal, vol. 2, pp. 1-16, 2015.

[23] J. Ramos, "Using tf-idf to determine word relevance in document queries," in Proceedings of the first instructional conference on machine learning, 2003.

[24] M. Lan, C. L. Tan, J. Su, and Y. Lu, "Supervised and traditional term weighting methods for automatic text categorization," IEEE transactions on pattern analysis and machine intelligence, vol. 31, pp. 721-735, 2009.

[25] K. A. University. (2016, 1 March). Deanship of e-Learning and Distance Education Twitter Account. Available: https://twitter.com/kau_elearning

[26] S. Zhang, C. Zhang, and Q. Yang, "Data preparation for data mining," Applied Artificial Intelligence, vol. 17, pp. 375-381, 2003.

[27] W. Cherif, A. Madani, and M. Kissi, "A new modeling approach for Arabic opinion mining recognition," in Intelligent Systems and Computer Vision (ISCV), 2015, 2015, pp. 1-6.

[28] A. S. Hussein, "Arabic document similarity analysis using n-grams and singular value decomposition," in Research Challenges in Information Science (RCIS), 2015 IEEE 9th International Conference on, 2015, pp. 445-455.

[29] H. Mubarak and K. Darwish, "Using Twitter to collect a multi-dialectal corpus of Arabic," in Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), 2014, pp. 1-7.

[30] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical machine learning tools and techniques: Morgan Kaufmann, 2016.

[31] Y. Thakare and S. Bagal, "Performance evaluation of K-means clustering algorithm with various distance metrics," International Journal of Computer Applications, vol. 110, 2015.

[32] S. Kapil and M. Chawla, "Performance evaluation of K-means clustering algorithm with various distance metrics," in Power Electronics, Intelligent Control and Energy Systems (ICPEICES), IEEE International Conference on, 2016, pp. 1-4.