



## Clustering Arabic Tweets for Saudi National Vision 2030

Ibtihal Ferwana<sup>1</sup>, Dana Alhenaki<sup>2</sup>, Aljohara Alfayez<sup>3</sup>, Souad Larabi Marie-Sainte<sup>4</sup>

College of Computer and Information Sciences  
 Prince Sultan University  
 214410226@psu.edu.sa  
 214410214@psu.edu.sa  
 214410418@psu.edu.sa  
 slarabi@psu.edu.sa

### ABSTRACT

Twitter provides a valuable resource for opinion mining, which many applications can take place in investigating peoples' interests and concerns. In this paper, we are interested to investigate users' main concerns regarding the national Saudi vision of 2030. As the vision has many tracks of planning, we investigated the education and health sectors. The aim of this study is to show people's focus and interest in implementing this vision. For this, some unsupervised Machine Learning techniques are implemented. Three clustering algorithms are experimented (K-means, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF)). LDA and NMF give promising results, as both clustered the data into two significant discriminative clusters with a silhouette coefficient equals to 0.639 and 0.538 respectively. While K-means provides overlapped clusters with a silhouette coefficient equals to 0.181. Therefore, our models are able to cluster Arabic tweets in the context of the Saudi national vision. Additionally, they showed that the national vision 2030 is implemented in education more than in healthcare.

**Key words :** Clustering, Text Mining, TF-IDF, K-means, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF)

### 1. INTRODUCTION

Twitter is being a platform for people to express their feelings, goals, decision, and attitudes. All internet users sum up to 24 million people and 18.3 million of them are Twitter users. Also, 40% of Arab tweets are written by Saudi Arabia people [1]. As Twitter is the most used social media platform, researchers find it a valuable research area to focus on [2]. Therefore, tweets are extracted from Twitter to be used for research purposes. Users' objectives can be grouped into several groups such as healthcare, education, transportation, etc, while, attitudes can be positive or negative.

Since the announcement of the Saudi Arabian National vision of 2030, there have been many opinions, discussions, arguments, judgments, and different talks about it. The national vision consisted of multiple tracks of improvements, such as the track of education, health, research, Pilgrimage, and transportation. Each institution has taken the responsibility to work on a special track. Many users express their work plans and achievements that go along with the vision in different fields.

The aim of this research is to analyze Saudi Arabia's users' attitudes, objectives, and aims toward the Saudi national vision 2030, through clustering the collected tweets of sentiments in the Arabic language. We only focused on the area of healthcare and education since they are the biggest sectors in Saudi Arabia and the most important topics according to the Saudi vision plans. Up to our knowledge, this study is the first of its kind. Moreover, many researchers used tweets for classification problems; however, for clustering, it is being brand-new aim [2]. Clustering algorithms are unsupervised Machine Learning techniques that consist of categorizing unlabeled data based on a specific similarity [3]. In this study, three clustering algorithms are investigated, K-means, LDA, NMF. K-means is widely used for clustering textual data, either Arabic or English. While LDA and NMF are used for clustering English textual data [4]. Many challenges make this study interesting. For instance, the tweets themselves have many misspelling, shortcuts, missing parts, and the variation of Arabic dialects [5]. These problems can make the standard algorithms of clustering and classification challenging. Also, the limited resources of Arabic corpus made it a challenge to create a model and cluster the data semantically [6]. This study is developed based on the following research questions:

1. Could the Arabic tweets be clustered into two significant clusters of education and health?
2. Could the biggest cluster be figured out to conclude the most important topic users are concerned about for the vision 2030?
3. Could the two clusters be sub-clustered into subtopics to investigate detailed topics?

In this study, the Twitter official API is utilized to extract tweets from Twitter automatically [4, 5]. This article is

organized as follows. Section 2 presents the related works. Section 3 addresses the methodology in details including the data collection, the preprocessing, the three clustering algorithms and the evaluation measures. Section 4 introduces a discussion about the obtained results. Section 5 concludes the study.

## 2. RELATED WORK

In the following, the existing works related to clustering tweets, either Arabic or English, are presented. Note that, clustering tweets is not mostly investigated in the Arabic language.

In [4], the authors showed that hashtags are good indicators of the tweet topic. They chose six predefined topics based on what was trending. Their total resulted data extraction was 1,107,007 tweets in the period of March 2011. The trending topics were 'News', 'Sports', 'Science and technology', 'Entertainment', 'Money', 'Just for fun'. For preprocessing, they created vocabularies, where each vocabulary contained specific features. Their vocabularies contain words in lowercase, whitespace tokenized, words which appeared less than 5 times are removed, and non-alphanumeric characters are removed (unless they are @ \_ - #). They clustered their data using K-means and LDA. K-means gave better purity and F-score measures by comparing the results with manually labeled data.

In [7], the authors were able to harvest 15,000 and 250,000 tweets of London and Brussels users respectively. They harvested only the tweets containing 5 keywords using python. Their resulted dataset is clustered into five clusters related to five keywords which are, 'Job', 'Time', 'Happy', 'Christmas' and 'Hire'. Therefore, they were able to identify the topics of interest in different regions. Their system was able to identify the main hottest topics from the tweets data.

In [3], the authors used Twitter data as a mean to improve education and its services provided by King Abdul Aziz University in Jeddah. They collected 1121 Arabic tweets related to the university and shared by students in 50 days. The authors used C# to develop their application and grab the Twitter data. They had three clusters to explain the data. They used TF-IDF to weight each word in tweets and then made the tweets ready to be clustered with k-means. They compared their resulted clusters with manually labeled data to calculate their clustering purity and then validate their study. They got three clusters to represent the concerns of their students, a cluster of exams and payments, another for their blackboard, and the third was about summer semesters.

However, the authors in [5] were interested to discover the dominant topics in Moroccan users' communication. The authors extracted and filtered Arabic tweets based on the geolocation information of Morocco. They collected the data using python and tweepy library to utilize the Twitter API for data streaming. They faced a challenge that Moroccan users write in different languages and dialects. They used TF-IDF

(term frequency-inverse document frequency) to measure the weight of each word and hence clustering the results. They used K-means for clustering with  $k=3$ . They figured out three clusters to represent main hot topics and mapped them on Morocco map to get an idea about where each topic is mostly communicated. Their first cluster contains words of Africa, group, Gabon. While their second cluster contains words of government and negotiation. Their third cluster contains words of politics and immigrants.

According to the previously presented works, several things are taken into consideration. First, data extraction is the first main step which was performed using different programming languages. In this study, Python and the tweepy library are used. Second, the research in Arabic tweets is still in its beginning; therefore, there is an area to contribute and improve the Arabic text clustering. However, it might be challenging to handle Arabic text, since there is a lack in Arabic corpus [6]. Third, researchers successfully cluster text using TF-IDF functions, then the k-means algorithm, which proved to be successful even in Arabic text. Consequently, in this paper, TF-IDF is used to represent the textual Arabic tweets. In addition, three clustering algorithms are employed including K-means which is extremely used, Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) which are not utilized in this context.

## 3. METHODOLOGY

In this study, three research questions are presented but only the first two questions are investigated due to the limited time.

### A. Clustering Arabic Tweets

Our first research question investigates the ability to cluster Arabic tweets into two discriminative clusters, the following stages helped us in achieving its answer.

#### Data Collection

Our main data source is the Twitter web server. Collecting data from Twitter has several phases. The first phase is to create a Twitter API, then to set the filtering criteria.

To be authorized to stream data from Twitter, we created an account in <https://apps.twitter.com>. Each created account is provided with four secret information to set up the Twitter API [5] including an access token, access token secret, consumer key, consumer secret. The tweepy library is used as recommended by several researchers [5].

```
auth = tweepy.OAuthHandler(consumer_key,
consumer_secret)
auth.set_access_token(access_token,access_token_secret)
api=tweepy.API(auth,wait_on_rate_limit=True)
```

The streaming needs to be filtered in order to search for tweets of our interest and to reduce the size of tweets dataset. Since we are interested in education and healthcare within the era of the 2030 national vision, the API filter is "2030 and Vision and Education" or "2030 and Vision and Health" in Arabic. For example, one query would be,  $q = u$  '2030 AND رؤية AND تعليم'.

The Twitter API is able to get the tweets existing for one week. Therefore, to get more data, the script was run for three times, once on November 30<sup>th</sup>, December 7<sup>th</sup>, and December 21<sup>st</sup> to get a total of 330 tweets.

After extraction, the tweets are written and saved in a text file in order to process them easily.

### Preprocessing

Before we process the tweets, we must eliminate the noise contained in the tweets for better results [7]. The purpose of preprocessing is to take the data (tweets) through multiple phases. In each phase, we deal with a specific technique until reaching the final phase which will be ready for analysis [6]. Performing preprocessing techniques is necessary to only keep relevant information in the tweet. Some of the preprocessing techniques that we applied are:

- Remove all non-Arabic tweets: remove any tweet that is not written in the Arabic language.
- Tweet Cleaning: remove irrelevant information like retweets, hyperlinks, usernames, punctuations, emails, and special characters (emojis, and smileys) [8].
- Stop words Removal: stop words are irrelevant words that do not hold relevant meaning and do not contribute to the process. Examples of stop words in English are “the, and, you...” and in Arabic “الذي، في...إن”
- Stemming: delete the suffix of a word until we find the root of the word [5]. For example, in English (stemming => stem). In Arabic, the root of words always consists of three letters only. ( تعليم => علم ) The library used is ISRI Arabic Stemmer, which is a rooting algorithm for Arabic text [9].

Table 1 shows an example of preprocessing a tweet using the techniques explained above.

**Table1:** An example of preprocessing a tweet

Original Tweet	تماشياً مع رؤية المملكة 2030 ودعماً للتحوّل الرقمي، والمساهمة في بناء جيل من الكفاءات الوطنية في مجال الحاسب، مشاركة نسبة عالية تقريباً من طالبات المدرسة في ساعة برمجة Umlj1450 @nawalsanyour @oihm1 @ @tabuk_edu
Tweet Cleaning	تماشياً مع رؤية المملكة ودعماً للتحوّل الرقمي والمساهمة في بناء جيل من الكفاءات الوطنية في مجال الحاسب مشاركة نسبة عالية تقريباً من طالبات المدرسة في ساعة برمجة
Stopwords Removal	تماشياً مع رؤية المملكة ودعماً للتحوّل الرقمي والمساهمة في بناء جيل الكفاءات الوطنية في مجال الحاسب مشاركة نسبة عالية تقريباً طالبات المدرسة ساعة برمجة
Stemming	مشي روة ملك دعم تحل رقم سههم بنء جيل كفاء وطن مجل حسب شرك نسب علي قرب طلب درس سعة رمج

### Feature Selection and TF-IDF

Before starting clustering, textual data must be converted into numerical values. To do this transformation, the TF-IDF is

calculated for each tweet to get the weight of the important words in each tweet and in the whole set of tweets. TF-IDF weight has two values. TF is the frequency of a word in a document, and IDF is the inverse of the word frequency in the whole document corpus [10]. This means two main things. First, the more the word occurs in a document the more representative it is. Second, the more text the word occurs in the less, it is a discriminative feature. In other words, it is not important [10].

Python library, scikit-learn, is used to implement TF-IDF calculations and vectorization of the text [5].

### K-Means Clustering Algorithm

According to [11], the k-means algorithm requires predefining the number of clusters, K. Therefore, we choose our clustering groups number to be k=2 since our interest is to discover two clusters including education and health. In each iteration, the algorithm chooses the most representative node (tweet) to make it the center of the cluster. Python *scikit-learn* library is used for the k-means, see the script below.

```
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import
TfidfVectorizer
vect = TfidfVectorizer(smooth_idf=False,
sublinear_tf=False, norm=None, analyzer='word')
text_fitted = vect.fit(tweets)
Xfitted = vect.fit_transform(tweets)
n_k = 2
model = KMeans(n_clusters=n_k, init='k-means++',
max_iter=100, n_init=1)
model.fit(Xfitted)
clusters = model.labels_.tolist()
print model.labels_
```

For each cluster, we print the most important words in the cluster, which represent the cluster centroid and its nearest neighbors as shown in the script above.

### LDA and NMF Clustering

LDA is an abbreviation of the Latent Dirichlet Allocation, which is a probabilistic graphical modeling algorithm. While NMF is an abbreviation of Non-negative Matrix Factorization that relies on linear algebra. Both algorithms are dedicated to being used for textual analysis; however, they were not tested or investigated for Arabic text.

Both algorithms take a bag of words as an input. *Scikit-learn* Python library is used to run LDA and NMF clustering algorithms. NMF processes the vectorized TF-IDF data, while LDA deals with raw counts, therefore CountVectorizer is used. CountVectorizer gives us the row count through calculating the term frequency only which LDA depends on intrinsically. The script used for both algorithms is below.

```
from sklearn.decomposition import NMF,
LatentDirichletAllocation
from sklearn.feature_extraction.text import
TfidfVectorizer, CountVectorizer
no_topics=2
# Start Clustering
lda=LatentDirichletAllocation(n_topics=no_topics,max_i
ter=15,learning_method='online',learning_offset=50.,ran
```

```
dom_state=0),fit(tf)
nmf=NMF(n_components=no_topics,
random_state=1,alpha=.1,l1_ratio=.5,init='nndsvd'),fit(tf
idf)
```

### Performance Evaluation

Since the ground truth labels (labeled data) are not used in our case, the evaluation is performed using the model itself. Therefore, the Silhouette Coefficient is selected where a higher Silhouette Coefficient score explains a better clustering discriminatory [12]. The silhouette coefficient depends on two measures, the mean of the intra-cluster distance and the mean of the nearest cluster distance for each node sample [12,13]. In other words, it calculates how much a node related to its assigned cluster and how much it relates to the nearest clusters. Eventually, it outputs the mean silhouette coefficient of all samples. This measure is imported from Python *scikit-learn* library. Sample code for calculating the silhouette coefficient for the k-means clustering is as follows.

```
from sklearn import metrics
clusters = model.labels_.tolist()
metrics.silhouette_score(Xfitted, clusters,
metric='euclidean')
```

### B. Larger Cluster

Our second research question tries to investigate the largest cluster among both clusters of education and health. Therefore, we needed to count tweets dedicated to each cluster according to the used algorithms. Each algorithm outputs a list of the assigned clusters (e.g. 0 or 1), therefore, these counts were calculated and graphed to visualize the larger cluster as shown in Table 2 (Percentage of each cluster).

## 4. DISCUSSION

Note that, the number of iterations and runs for each clustering algorithm are 120 and 10 respectively. These values are set after performing several experiments and noticing that the obtained results cannot be improved. Table 2 summarizes the comparison between the outputs of the three clustering algorithms. We outputted the first six most important words in each cluster. Qualitatively, the topmost words represent the topic of a cluster [7]. Therefore, LDA and NMF are better in outputting discriminative clusters from the first run. While the k-means does not stabilize in any number of iterations and any number of runs. For the first cluster, (علم) the root of “education” is at the top in LDA and NMF. In the second cluster, the highest TF-IDF term is (صحة), “health”. This outcome supported our aim in clustering the data into two main clusters of health and education. This also supports the accuracy of the clustering models, since the data is split into two significant clusters. This answers our first research question. The highest score of silhouette coefficient is awarded to LDA algorithm.

According to the graphs in Table 2, we represented the number of tweets in each cluster to figure out the majority cluster. We counted the tweets in each cluster for the best run only (See Table 2). The majority cluster is the “Education”. This answers our second research question about the largest

cluster. Therefore, this concludes that the national vision of 2030 is implemented in education more than in health.

## 5. CONCLUSION

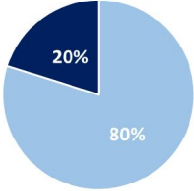
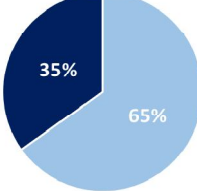
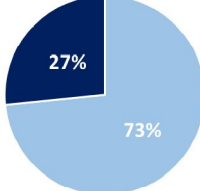
Due to its huge data variety, Twitter is an excellent option for data analytics and studies. In this article, we showed the interest of the Saudi national vision 2030 specifically in education and healthcare fields provided by tweets using the Arabic language. To extract tweets, Twitter API is acquired and the tweets are filtered to reduce dataset size. Then, the tweets are preprocessed by removing non-Arabic words, simple tweet cleaning, removing stop words, and stemming. We proposed a way to cluster Twitter tweets by k-means (with k=2), LDA and NMF clustering algorithms. We calculated the weights of each tweet using TF-IDF in order to transform the textual tweets into numerical data for clustering purpose. We found two clusters which are healthcare and education. The result of the experiments guided us to conclude that LDA and NMF algorithms are effective approaches to analyze Twitter Arabic data. Therefore we are able to answer our first question of the ability to clustering the data into two clusters. Also, we answered our second research question that the national vision 2030 is implemented in education more than in healthcare Our study initially aimed to discover all the topics related to the national vision 2030. However, the challenges we faced hindered us to extract and process all this huge amount of Twitter data. Instead, we focused on the field of education and healthcare.

Our third research question was not answered due to the limited time. This will be our future work to cluster each topic into subtopics to well explore each cluster in details.

## ACKNOWLEDGMENTS

This work was supported by the Artificial Intelligence & Data Analytics Lab (AIDA), Prince Sultan University, Riyadh, Saudi Arabia.

**Table 2:** Comparison results

Algorithm	K-means	LDA	NMF
<b>Top Words/Cluster (First run)</b>	<pre> K-means Cluster 0: علم عزم صحة نهر اصل ككل كبيرة  Cluster 1: علم صحة رؤية حقيق سعد ملك                     </pre>	<pre> LDA Cluster 0: علم رؤية حقيق طلب سعد ملك  Cluster 1: صحة رؤية وزر وطن ملك                     </pre>	<pre> NMF Cluster 0: علم رؤية طلب حقيق سعد ملك  Cluster 1: صحة ملك دين نور مشي                     </pre>
<b>Silhouette Coefficient</b>	0.1813949486598279	0.6391437308868502	0.5379590036618981
<b>Top Words/Cluster (Best run)</b>	<pre> K-means Cluster 0: علم رؤية طلب حقيق ملك سعد  Cluster 1: صحة رؤية وطن هدف حضر                     </pre>	<pre> LDA Cluster 0: علم رؤية حقيق طلب سعد ملك  Cluster 1: صحة رؤية وزر وطن ملك ال                     </pre>	<pre> NMF Cluster 0: علم رؤية طلب حقيق سعد ملك  Cluster 1: صحة ملك دين نور مشي                     </pre>
<b>Silhouette Coefficient</b>	0.04116794108572462	0.6391437308868502	0.5379590036618981
<b>Percentage of each cluster tweets (The best run)</b>	<p style="text-align: center;"><b>K-means Best Run</b></p>  <p style="text-align: center;">Cluster 0 - Education    Cluster 1 - Health</p>	<p style="text-align: center;"><b>LDA Best Run</b></p>  <p style="text-align: center;">Cluster 0 - Education    Cluster 1 - Health</p>	<p style="text-align: center;"><b>NMF Best Run</b></p>  <p style="text-align: center;">Cluster 0 - Education    Cluster 1 - Health</p>

**REFERENCES**

1. Addawood , A., AlShamarani, A., Alqhatani, A., Diesner, J., & Broniatowski, D. **“Women's Driving in Saudi Arabia –Analyzing the Discussion of a Controversial Topic on Twitte. sbp-brims”**, 2018. [http://sbp-brims.org/2018/proceedings/papers/latebreaking\\_papers/LB\\_16.pdf](http://sbp-brims.org/2018/proceedings/papers/latebreaking_papers/LB_16.pdf)
2. Abuaiadah, D., Rajendran, D., & Jarrar, M. **“Clustering Arabic Tweets for Sentiment**

**Analysis”**. 14<sup>th</sup>-IEEE International Conference on Computer Systems and Applications, pp. 449-456, 2017.

<https://ieeexplore.ieee.org/document/8308321/>

3. Al-Rubaiee, H., & Alomar, K. **“Clustering Students' Arabic Tweets using Different Schemes”**. International Journal of Advanced Computer Science and Applications, pp. 276 – 280, 2017.

<https://pdfs.semanticscholar.org/b924/57462d03f1d67ecdabf2e170d666d1bdbaf5.pdf>

4. Rosa, K. D., Shah, R., Lin, B., Gershman, A., & Frederkin, R. **“Topical Clustering of Tweets. CiteSeetx”**, 3Rd Workshop on Social Web Search and Mining, 2018. <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.207.4287>

5. Abdouli, A., Hassouni, L., & Anoun, H. **“Mining tweets of Moroccan users using the framework Hadoop, NLP, K-means and basemap”**. 2017 IEEE-Intelligent Systems and Computer Vision. Morocco, 2017. doi:10.1109/ISACV.2017.8054907
6. Al-Azzawy, D. S., & Al-Rufaye, F. M. **“Arabic Words Clustering by Using K-means Algorithm”**. 2017-IEEE Annual Conference on New Trends in Information and Communication Technology Applications, 2017. <https://ieeexplore.ieee.org/document/7976098/>
7. Sechelea , A., Huu, T. D., Zimos, E., & Deligiannis, N. **“Twitter Data Clustering and Visualization”**. IEEE 23rd International Conference on Telecommunications, 2016. <https://ieeexplore.ieee.org/document/7500379>
8. Aldayel, H., & Azmi, A. **“Arabic Tweets Sentiment Analysis - a hybrid scheme”**. Journal of Information Science, 6(42), 782-797, 2015. <https://journals.sagepub.com/doi/abs/10.1177/0165551515610513>
9. Tghva, K., Elkhory, R., & Coombs, J. **“Arabic Stemming Without A Root Dictionary”**. IEEE International Conference on Information Technology: Coding and Computing (ITCC'05), Vol. 1, pp. 152-157, 2005. doi:10.1109/ITCC.2005.90
10. Yun-tao, Z., Ling, G., & Yong-cheng, W. **“An improved TF-IDF approach for text classification”**. Journal of Zhejiang University SCIENCE, pp 49-55, 2005.
11. Jain, A. K. **“Data clustering: 50 years beyond K-means”**. Pattern Recognition Letters Journal, pp. 651-666, 2009.
12. Rousseeuw, P. **“Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”**. Journal of Computational and Applied Mathematics. Vol. 20, pp. 53-65, 1987. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
13. Saravana Balaji B , Krishna Kumar V , Ahmed Najat Ahmed. **“Semantically enriched Tag clustering and image feature based image retrieval”**. International Journal of Advanced Trends in Computer Science and Engineering. Volume 8, No.1.2, 2019