International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

Classification Arabic Twitter User's Insights Using Rough Set Theory

Ahmed Tamam[a]*, Hatem Abdelkader[b], Asmaa Haroun[c]

[a]*Department of Information System , Faculty of computers and Artificial intelligence, Cairo, Egypt*

[b,c]*Department of Information System , Faculty of computers and Artificial intelligence, Cairo, Egypt*

[a]*Email: Ahmedtamam415@yahoo.com,* [b]*Email: Hatem.AbdllKader@fci.cu.edu.eg,*

[c]*Email: asmaa.elsayed@fci.cu.edu.eg*

Abstract

Nowadays, people using social media from around the world to share their daily affairs. Arabic twitter for example is a platform where users read, reply, post which known 'tweets'. Users trading their opinions on different trends that are not equal in important and differed based on their power and interest. Tweets can provide rich information to make decision. The main objective of this paper is to present a framework for making a valuable decision through analyzing social users' insights based on their proximity to a particular trend with highlights their power in this trend. Tweets are exceedingly unstructured that makes it difficult to analyze. Nevertheless, our proposed model differs from previous research in this field it gathered the use of supervised and unsupervised machine learning algorithms. The process of performing this work as follows: classifying users based on the degree of their closeness/interest utilizing Mendelow's power/interest matrix, rough set theory to eliminate the features that may be found in user profiles to find minimal sets of data. The proposed model applied two attribute reduction algorithms on our dataset to determine the optimal number of reducts for improving decision making from the user replies. In addition to, unsupervised machine learning to group their replies into subcategories such as positive, negative, or neutral. The experimental evaluation shows that Johnson algorithm has reduced the user attributes by 71% than genetic algorithm that utilized in a classification model.

Keywords: Data mining techniques; sentiment analysis; Rough set theory; Classification social media.

---

* Corresponding author.

1. Introduction

Todays, Social media platform such as Twitter, Facebook, Google+ , instagram ,Keek, Myspace, LinkedIn has become an online community that allow users to share their affairs with the whole world. People can write their own opinions on trends or share their moments with differentiate degrees of satisfaction. The analysis of user's reviews can lead us for making useful decisions in particular field. Sentiment analysis also known opinion mining is one of natural language processing subfield, used for analyzing people's opinions and emotions (positive, negative and neutral) [1].

Data mining is an essential processes of extracting meaningful information from a massive quantity of data. Data mining have many methods utilized such as classification, clustering and visualization to investigate the field of social media data mining that have many of informative information referenced by user's twitter accounts. Additionally, this information is collected from captured tweets which can helps for important decisions to make efficient a particular field.

There are many users on the social media platform who differ in power/interest. To make their opinion influential regarding the certain trend based on their closeness/interest, Users of social networking sites must be identified based on their power/interest in a specific area [2].

The identity of the user on social media can be recognized through his profile, which contains a rich information that may divided into explicit and implied features [3]. Consequently, it's important decide on appropriate attributes and the suitable model to classify users based on the degree of their closeness/interest for the certain trend. Thus, it is possible to make people with high power/interest classification as experts in a particular field. In this instance, it is important to address their replies/ opinions when enhancing the valuable decisions for the field. In addition to, the collected user's opinions are gathered in the Arabic language which about 330 million speakers, it is a widely spoken language. While the Arabic language is a member of sematic language family. There are several obstacles related with the Arabic language, which has many equivalent dialects for distinct geographical places [4]. For example, some characters have many forms, which is an issue of Slang phrases that have a different meaning in Arabic. Moreover, numerous orthographic mistakes.

In this paper, we propose a model that collected tweets about certain trend for predicting a valuable decision of user's opinions based on data mining techniques as follows: first, classification of users using Mendelow's power-interest model and the rough set theory with the specified user's attributes. The rough set theory classifies users according to their power/interest. We applied and compare the features reduction given by Johnson algorithm and genetic algorithm. Then, we handled unsupervised technique (k means algorithm) which will subsequently be utilized to group the Arabic replies of the classified user into subcategories whether positive, negative, or neutral. The result that obtained from the classification model and the clustering process for the user's replies/opinions facilitates in making valuable decisions in a particular field. The sections of this paper are discussed in the following manner: Section 2, compare the work of different models that utilized data mining methods for making decisions from social media platforms. In section 3, the proposed model is covered in detail. In section 4, the presented model is examined with educational case study. Finally, in section 5 the

conclusion is introduced .in addition, future research.

2. Related Work

Numerous studies on data mining techniques in social media has a great deal information that led for making some decisions. In this section, some of these studies are shown as follows:

Salama hasan et L.M.R.J.Lobo [5], proposed a model that analyzed products reviews collected from twitter for making a decision about people reviews. This model classifies tweets in three categories (positive, negative or neutral) based on dynamic LM classifier for predicting sales performance .in addition, ranking the classified tweets according to their favourite count. The results compared with j48 classifier for many forms of data (textual, numeric, and tabular) the experimental work showed that dynamic classifier performed better than j48 classifier for numeric and tabular data. The model did not utilize natural language processing techniques on dataset that were essential for the classification process. Furthermore, Pradeep Vashisth et Kevin Meehan [6], chosen twitter as a platform for predicting the user's gender from tweets data. The system achieves the best solution to find the category of users, whether (male or female) via utilizing multiple natural language techniques (Ttf-idf, W2 vec, and Glove) with many traditional machine learning algorithms such as SVM, LR, NB, Random forest, and Boost. The final results showed the LR model gives the highest accuracy score. This model doesn't use the N-gram technique which can improve accuracy.

Juan Antonio et juandiego morzan [7], introduced a model that collecting the tweets of people who are sharing information about their health-care for improving their outcomes. The tweets are composed of two different subset (HPV) and (lynch syndrome) based on the topic modeling technique and document clustering application to extract the hidden topics from large text .The application set up six methods and k-means based on different representation (TF-IDF and Doc2vec). The experimental work shows that Gibbs LDA and Online Twitter LDA gives a better performance for extracting topics and grouping tweets. The proposed model neglected to utilize the elbow method to determine the optimal number of clusters.

Reference [8], presented a framework for collecting user's reviews from Arabic twitter for sentiment analysis by labelling the tweets in positive or negative. This framework utilized a discriminative multinomial naïve-Bayes algorithm with natural language processing techniques (TF-IDF, N-gram, Tokenize, and stemming). The experimental results is compared with more machine learning algorithms (SVM, KNN, NB, and D-Tree). The proposed framework improved the accuracy with 0.3%. The major problem for this model is not using feature selection in data pre-processing.

Reference [9], proposed a model that collected a sample of data about the final exam scores based on k-means algorithm which cluster the teachers in teaching and personalized coaching for improving decision making in teaching management .Additionally, Hamed Al-Rubaiee et Khalid Alamor [10], proposed a model for clustering Arabic tweets that mentioned by king Abdul-Aziz university students for understanding their behaviours and answering their inquiries about new semester. The model utilized unsupervised machine learning k-means algorithm with different vector's representation such as (TF-IDF and BTO).in addition, the N-gram feature is

involved and calculate the centroid of clusters. The experiment shows that the clustering using BTO with N-gram gives better results. The model has problem that did not depend on the elbow method for determining the number of clusters. The models discussed previously have some limitations that must be taken into consideration to improve a particular field/area. Furthermore, circulated user comments on social media varying in importance for making good decision. The comments can be not equal in terms according to the closeness /interest for the users. As for this paper, we proposed a model for assessing Arabic comments and classifying social media users as a step toward enhancing making a decision utilizing the rough set theory and Mendelow's power/interest model compared between two reduction algorithms according to data mining approaches. In order to conquer the constraints of prior models, our model has linked the different users and their opinions by utilizes Arabic natural language processing techniques and feature selection. Furthermore, the model encourages the elimination of some of the problems in relation to the content of the Arab social media sites.

3. The proposed model for decision-making

In this part, we listed the mechanisms that were utilized for the suggested model. The Mechanisms for providing a valuable decision that will take (categorise and identify Arabic twitter users, clustering their insights) are among the mechanisms. The concerned technologies are machine learning techniques such as clustering. Moreover, classification based on rough set theory and Mendelow's power-interest model. The framework of proposed modeling system is shown in figure 1. It is comprised of the following stages (Data Collection, Data Preparation, Classification and Clustering methods, and Making Decision). Finally, the proposed model result is presented as follows: the classification process for user's profile utilizing The Rough set theory according to Mendelow power-interest model and the clustering process for the replies of users according to k-means. The outcome can be combined into a single document that can contribute to make some valuable decision by introducing a prioritized twitter users and their list.

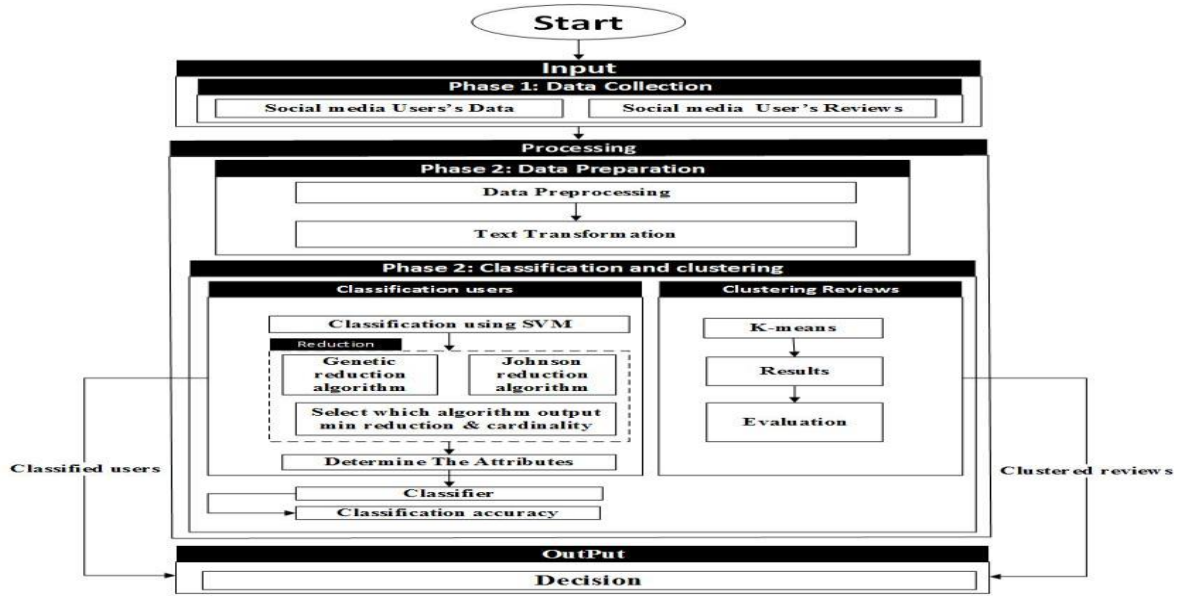

Figure 1: The Framework of the proposed model

3.1 Data Collection

Educational opinions/replies were chosen from Arabic Twitter as a dataset for constructing the proposed model. The dataset contained many features, such as (user-name, user profile, short-bio, and tweet replies). The replies of tweets utilized to determine which cluster belongs to and short-bio to determine the interest and job for the classification process.

3.2 Data Pre-processing

Data preparations is a process for transforming Arabic raw data that have many challenges (slang words, orthographic mistakes, unstructured language) that will help for making a good prediction [4]. The Arabic language included several steps for pre-processing as follows:

3.2.1 Text Pre-processing

This step has several basic procedures in the dataset to optimise the selected texts into subcategories:

3.2.1.1 Filtering stage

For reducing the complexity and the dimensions of dataset that pose numerous hurdles towards text mining [11]. For that, commas, diacritics/accents, punctuation marks are removed.

3.2.1.2 Tokenization stage

Is a common step to split the given text into tokens that might aid the model in processing step. As a consequence, the text data are posed by a single vector.

3.2.1.3 Removal stop words stage

These indicates to the words that do not affect text mining. Usually, they have been presented as conjunctions, and prepositions [12] such as "هُم/Hom", "إن/Eh", "إلى/Elah", which removed before any processing.

3.2.1.4 *Stemming stage*

The Arabic language has many words which written in different forms like "العب/Aleab", "نلعب/Nealb", "يلعبوا/Yealbu" …. etc. for that, we must use the stemming method which returns the word to its root [4].

3.2.1.5 *Normalization stage*

The Arabic language has some characters which have the same functions are written in different forms. For this reason, this stage is utilized for normalized the text according to the specific strategy [4]. For example, the character "ا/ Alef" might appear as character "أ/ Alef " (with Hamza above) in the text, the character" إ Alef" (with Hamza below), or "آ" (with made above), and these have to normalized as character "ا"(without any

Hamza).

3.2.2 Text Transformation

Is the process of converting raw data which can be text into numerical features (vectors). These process can be utilized in building machine learning model based on feature transformation techniques like term frequency and inverse term frequency (TF-IDF) which uses a combination of two metrics in its computation, namely:

3.2.2.1 Term Frequency (TF)

It is a measure of how frequently a term (t) occurs in a document (d). TF is outlined in equation (1):

$$\mathrm{TF(t)} = \frac{\mathrm{NAD}}{\mathrm{TAD}} \qquad (1)$$

(NAD) which considered the number of times the adjective term occurs in document and (TAD) considered as the total number of the adjective in the document.

3.2.2.2 Inverse Document Frequency (IDF)

It is a measure how the term is important. Whereas, the TF alone is not sufficient for understanding the importance of words. IDF is outlined in equation (2):

$$\mathrm{IDF(t)} = \log\left(\frac{\mathrm{ND}}{\mathrm{DF(t)}}\right) \qquad (2)$$

(ND) refers to the number of documents in the document collection where document frequency (DF (t)) refers to the number of document in which adjective term (t) occurs in the document collection.

3.3 Clustering Method

In this stage, we utilized unsupervised machine learning technique to cluster the selected features after pre-processing the text. As, collected replies or opinions of users are not specifically labelled, we utilized K-means technique that is the most effective grouping algorithms among the others [13]. k-means works without labeling to distinguish the undetermined groups in order to confirm the supposition regarding which groups are presented. Relative to our model, the algorithm predicts which replies of the users might be one of the following subgroups as (positive or negative, or neutral) this algorithm works in the manner mentioned below [14]:

a) Centres of cluster are randomly initialized utilizing K points.

b) The clusters are assigned to their nearest centre based on the Euclidean distance measure.

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

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

3.4 Elbow Method

Before performing K-mean algorithm that depends on initially to find out the K parameter which represents the number of clusters for our proposed model. At first, the elbow method is implement. This is essential step for determining the best number of clusters for the gathered dataset. In figure 2, describes the number of reached clusters for gathered data is =3 cluster.

Algorithm 1: Elbow Method to find K of K-means

---

1. Initialize $K=1$

2. Start

3. Increment the value of $K

4. Measure the cost of the optimal quality solution

5. If at some point the cost of solution drops dramatically

6. That's the true K

7. End

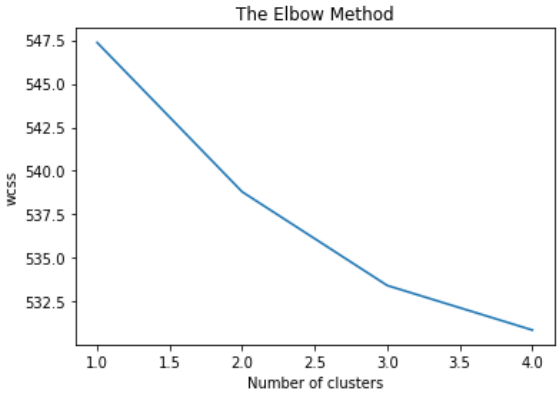

Figure 2: Elbow Method

In the case of the cluster evaluation process [15], we relied on utilizing a homogeneity score that describes the

closeness of the clustering algorithm to this perfection. The score for the clustering process of our model is = 0.15, the homogeneity score is used to determine whether the clustering method meets an important requirement which is presented in a cluster should only contain samples from a single class which defined as follows in equation (3):

$$H = 1\text{-}h(y\ true\ |y\ pred) \qquad (3)$$

h(y true) Refers to the range from 0 to 1, with low values indicating a low homogeneity.

3.5 Classification Method

After performing the Arabic natural language pre-processing techniques on the dataset amassed, the subsequent step is preparing for classification phase according to the power/interest for users. It considered a fundamental phase as it helps in fetching the attributes from the profiles of users by utilizing SVM technique (a supervised machine learning) to classify each determined attribute, whether it is a (power) or an (interest) class. SVM (support vector machine) [16] is regarded as an important classification method which plot each data item as a point in n-dimensional space (where n is number of selected features) with the value of each feature being the value of a particular coordinate. Then, finding the maximum margin, identifying the hyper-plane that developed during the training method to distinguish one class from another. The hyper-plane can be defining in two-dimensional through using the mentioned equation (4):

$$w.x + b = 0 \qquad (4)$$

If we define x = (x1, x2) and w = (a,-1) where w is the unit vector in two dimensional for x1 that refers to x coordinate, x2 refers to y. Once the hyper-plane is known, the predictions are completed for the classes by defining the hypothesis function (h) as follows in equation (5) :

$$h\theta(x) = \begin{cases} 1 & if\ \theta^{T}.x \geq o \\ 0 & otherwise \end{cases} \qquad (5)$$

class +1 will be assigned to point above the hyper-plane, while class -1 will be assigned to the point below the hyper-plane.

The classification table for (power-interest) Arabic Twitter users is the outcome of this phase that classified as position (power) class or interest class according to the features. Besides that, the clustered comments/replies table.

3.5.1 Mendelow's Classification Model

This model has a core objective to categories our social media users into a variety groups for identifying their importance to a certain field. We adopted stakeholder analysis and classification models for determining user's importance. Mendelow's power/interest model is most widely utilized in order to analysis and user

classification. our social media users calssificaion based on important questions have to find their answers [2]:

- What is the power of each stakeholder?
- What is the stakeholder's degree of interest?

Mendelow's matrix utilized in this classification for grouping social media users based on their (power-interest) in the proposed model and its resultaunt.

The steps that must be taken with stakeholders on this grid are depicted in figure 3:

1) High power/ High interested people: these people you must make the greatest efforts to satisfy and fully engage for them.
2) People who High power and less interested: you should keep them satisfied but not to the extent that make them bored.
3) People who Low power and High interested: these people must informed adequately with them, make sure that no major affairs are arising. AS, these people can often be very helpful.
4) Low power/less interested people: again, you must monitor these people and don't bored them by the excessive communication.

As a result, every social user will be classed as one of the following; high power, high interest or high power, low interest or low power, high interest or low power, low interest. These are used as one of the rough set theory's output decisions where the rough set theory's second decision attributes will be the actions that have to be taken with each classified user.

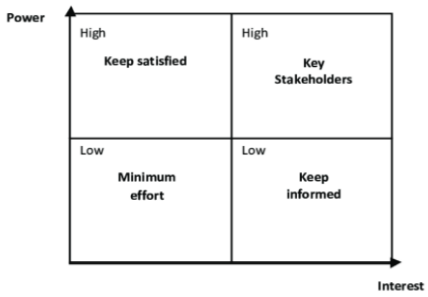

Figure 3: Mendelow's Matrix

3.5.2 Rough set theory and Attribute Reduction

Rough set theory (RST) was proposed by Z.Pawlak on 1982[17]. The essence of RST in the upper and lower approximation set. Rough set theory uses the indiscernibility relation and data pattern comparison to describe an information system with indiscernible data, where the data is ambiguous or inconsistent [18]. RST is used to categorize the stakeholders depending on the next phases:

3.5.2.1 Pre-processing phase

It necessitates the use of the decision table to create a Rough set analysis. Many data preparation tasks have been applied .Also, Data is split into two generated subsets at random: one of the subsets is utilized for analysis, which containing some of the objects in the dataset (Training set) and the other subsets are utilized , which containing the remainder of objects in the dataset (Test set) [2]. As, mentioned prior, the decision attributes are excluded from Mendelow's model.

3.5.2.2 the reduction of Attribute & generating the Rule Phase

It comprises the development of preparatory knowledge, such as decision table analysis to extract and reduce duplicate attributes form applying two different Rough set algorithms (Genetic reduction and Johnson reduction) to choose the optimum reduction according to reduction, and rule cardinality numbers.

3.5.2.2.1 Johnson Reduction

this is an heuristic algorithm that has special kind of methods. The main objective of this algorithm is to find a single attribute(reduct) that occurs in clause. This algorithm begins by setting S,that considers acandidate for the present of reduction for the empty set.in the next step, the algorithm counting the appearance of each attribute into clause.the attribute with the highest count has been added to S, and the entire clauses within F have been removed from the discernibiliy function(fD).Then, S is returned as a reduction[19].

3.5.2.2.2 Genetic Reduction

This algoritm is a develpoing method which simulates the biological evolution. It has ability of handling the complex optimization problems that are nonlinear and even involve space.Genetic algoritm has straightforward main idea,AS well as a standard operation mode and implementation steps[20].As aresult, utilizing both GA and RST is appropriate for accomplishing attribute reduction. Such a hybrid model has the potential to produce optimal or semi-optimal attribute reduction results [20].

The result of the comparison has shown through our model which of Genetic or Johnson algorithms provides optimal Reduction Attribute for the proposed model.

3.5.2.3 Classification and Prediction phase

From the previous phase, the generated rules are utilized to predict a class for each new social user. To convert a reduct into a rule, the condition feature values of the object class from that the reduct originated must be bound to the reduct's corresponding features. Thereafter, to complete the rule, a decision part includes the resulting part of the added rule. This is completed in the same way as for the condition features. For classifying the objects that have never been seen before the rules are generated from a training set that will be used. The actual classifier is represented using these rules. This classifier is used for predicting to which class assigned for the new object.

4. Case study: Educational Tweets of social users

The motivation of our study was applied to some issues on the social media platform, especially educational tweets for demonstrating the effectiveness of our model. Between these tweets, there are some tweets might be related to prominent/influential users where the power of (user's position) in education or the interest of users in education is determined by their proximity to the trend. If we can recognize these prominent people, we can benefit their insights/comments as a good opinion for improving the render of the educational sector. our study into exploring and analyzing the tweets for surveying people's insights/comments on whether public education or online education is the best educational direction for students to complete the current semester, we found out that majority of people support or opponent online the education more than public education, and there were also people's insights/opinions that were neutral. Consequently, we proposed a model that can classify and characterize the important users as well as, analyze the users' insights for making a worthy decision. The Arabic (student, teacher, pharmacist, and physiotherapist) tweeter's user accounts, tweets, and answers were analyzed. For gathering user profiles and tweets, the manual data collection approach is utilized. Data collection manually are more significant for further in-depth data analysis, while granularity is critical because data can give worthy insights.

4.1 Data-preprocessing&ClassificationModel

As previously stated in step 1, the collected Arabic dataset is fully processed. The result has been presented in Table 2 and Table 3. Subsequently, supervised machine learning (SVM) is implemented to the features which associated with the profiles of users for identifying the classes whether for role/ (power) and interest that important for applying to Mendelow's model. Table 1 presents some of the user's profile data and its classification into power and interest classes.

Table 1: SVM (POWER & INTEREST CLASSIFICATION)

Short Bio(Twitter user's profile)	Power	Interest
لسه طالب __ مهتم بكره القدم I am still student, interested in football	student	Sport
انا مُدرس (التدريس و التعلم هدفى) I am a teacher (teaching is my goal)	Teacher	pedagogical
صيدلى متابع اخبار الفنانين artist news Pharmacist following	Pharmacist	Art
دكتور صحة بيئة مدارس _ بحب اتعلم School environment health doctor _ I love to learn	Physiotherapist	Education

Table 2: Example of processing short-bio

Short-Bio	Processed Short-Bio
انا صيدلى ~ هاو الغناء I am Pharmacist , singing lover	صيدل هاو غنء
انا مدرس تاريخ & تدريس تاريخ الوطن مصر I am a history teacher & teach the history of the homeland Egypt	درس ارخ درس ارخ وطن مصر
طالب ف مدرسة شبرا الثانوية__/ مهتم بالرياضة A student at Shubra Secondary School __/ interested in sports	طلب درس شبر ثني هتم ريض

Table 3: Example of processing replies

User's Replies	Processed User's Replies
التعليم عن بعد احسن وسيله المحافظه علي صحة اولادنا Distance education is the best way to maintain the health of our children	علم احس وسل حفظ صحة ولد
انا موید لفکره التعلیم الحضوری I support the idea of attendance education	وید فکر علم حضر
لولا ان ممكن يبقى فى امتحانات انا مش هتعلم اصلا If there wasn't exams, I wouldn't learn at all	مكن يبق محن علم اصل

4.2 Social Users Classification based on Rough Set

The user profiles based on their role and interest were classified, a rough set and Mendelow's classification model are applied to the classified dataset. The rough set and Mendelow's power/interest models are considered fundamental to allow the presented model for determining the majority of significant users. As a result of previous, every user has been classified into subcategories whether; high or low for their power/interest [21]. The prediction model according to rough set classification have several stages:

4.2.1 Pre-processing stage

This stage has divided into several sub-phases

4.2.2 Preparing the information system table

As mentioned in Table 4, the attributes used as input dataset are shown in the information table ,which consists of conditional features (power, interest) identified from the user profiles in the prior stage and decision features (power/interest, action).

Table 4: The information Table

Short-Bio	Power(Position)	Interest	D1:Power/Interest	D2:Action
انا مُدرس .. التدريس هوا حياتى I am a teacher..teaching is my life	Teacher	pedagogical	High/High	Manage Closely
طالب اول اهتمامته الرياضة I am student whose first interest is sports	Student	Sport	High/Low	Keep Informed
صيدلى متابع اخبار الفنانين i am Pharmacist following the artists news	Pharmacist	Art	Low/Low	Minimum Effort
دكتور امراض معدية مدارس , مهتم باخبار التعليم Infectious disease doctor schools, interested in education news	Physiotherapist	Education	Low/High	Keep Satisfied

4.2.3 Completion of data, Discretization, Conversion , and Splitting

Arabic dataset have been collected and stored manually, that could comprise interference factors such as vacancy data, noise data, and inconsistency in data which making it unsuitable in knowledge discovery. As a consequence of, preprocessing of user's data is required; the missing values for objects are removed whether one or more, Data conversion is conducted for generating an appropriate form for processing utilized Rosetta tool [22]. In Table 5, Discretization result was presented. There are two subsets were generated at random, a training set for the analysis that contained 80% of the objects in the data set and a testing set that contained the remaining objects.

Table 5: Discretization Results

Meaning	0	1	2	3
Power	**Pharmacist**	**Physiotherapist**	**Student**	**Teacher**
(Position)				
Interest	**Art**	**Education**	**Sport**	**pedagogical**
Decision 1	**Low/Low**	**Low/High**	**High/Low**	**High/High**
Decision 2	**Minimum Effort**	**Keep Satisfied**	**Keep Informed**	**Manage Closely**

4.2.4 Attribute Reduction & Rule generation phase

In our experiment. Firstly, we utilized a rough set with Johnson reduction, then in the second experiment, we utilized Genetic reduction. Table 6 provides the results of the implementation of GA and Johnson algorithms to present the attribute data reduce. According to this table, Johnson produces a better result by having less number of reducts, fewer rules, and fewer cardinality. All social users are classified as follows: "a class with high power and interest, a class with low power and interest, a class with high power but low interest, and a class with low power but high interest". Johnson's algorithm is applied as figure 4, showed the numbers of reduct sets that produced through the application.

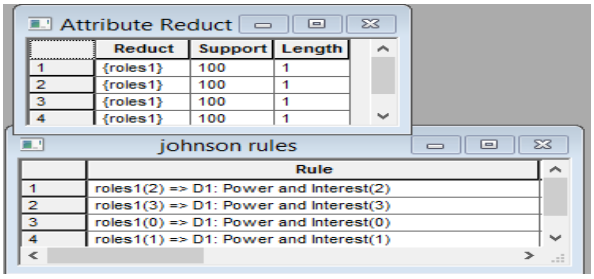

Figure 4: The generated Reduction Rule

Figure 5, presents the percentage of reduction achieved by implementing of Johnson Algorithm on the dataset features that gives four reduct with two attributes. This means that the reduction percentage is 71%.

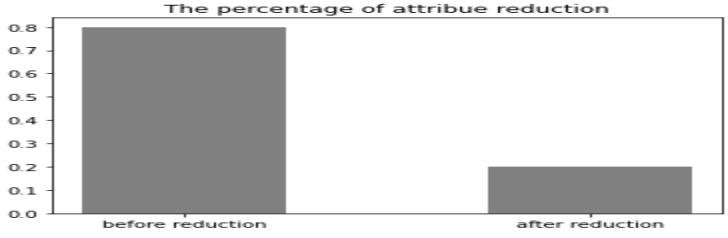

Figure 5: The percentage of attributes through reduction

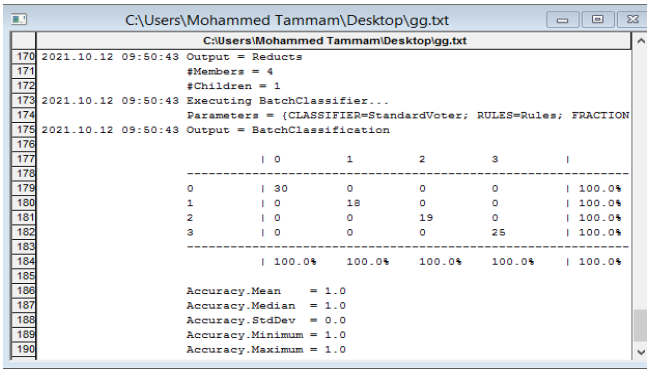

Figure 6: The confusion matrix

The confusion matrix is represented in figure 6 that described the visualization of the performance of the classification process. The cross-validation has been applied through dividing the data into training and testing sets (50-50%).

Table 6: The results of reduction

Algorithms	# Reducts	# Rules	# cardinality
Johnson	4	4	2
Genetic	8	8	3

4.2.5 Classification and prediction phase

To measure the effectiveness of rules obtained from consequently of applied the Johnson algorithm that utilized for assessing the applicable rules in classifying new cases performed. The rules are applied from the training set data to the test set data. “Fig. 7”, Shows the actual decisions that found in the test set are correct and not required to be re-entered again by the proper decision provided in the predicted line. For example, object 0 (the first in the test set) is classified in actual decision 2 (2; high power/Low interest) and the right predicted one is (2; High power/Low interest). Finally, the action that will take in accordance with each classified decision for each object figure7: Example for predicted user’s power and interest is (keep informed or keep satisfied or minimum effort or managed closely) for each decided user.

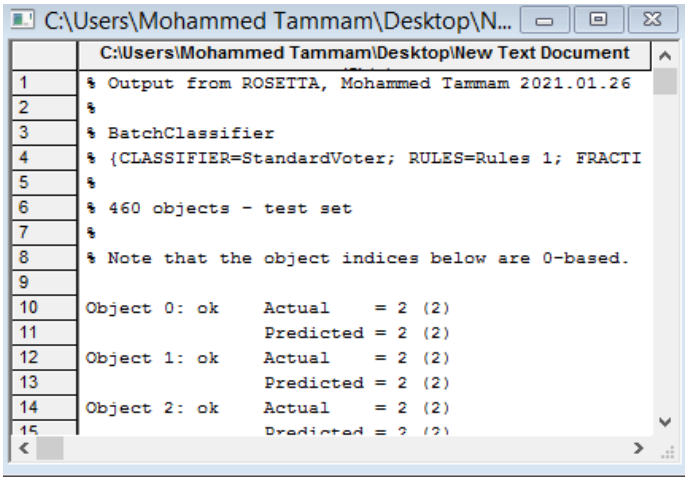

Figure 7: Example for predicted user's power and interest

4.3 Clustering Model

The study examines an aspect of Arabic tweets as an educational pattern by K-means. To understand the similarity and relationships of insights between different social users for grouping their replies in many categories. For this purpose, the replies of users have been processed using preprocessing Arabic text. Besides, utilizing an unsupervised learning for determining groups for these replies are graphically represented. The replies are related to the role (position) for social users that obtained from classification model to demonstrate the significance of users with their clusters that may be negative, positive or neutral mentioned in figure 8 that displays a scatter plot for data points that relates to each cluster. Three clusters were defined as (c0, c1, and c2) where gathered outcomes around an 'X' can represent as positive comment, neutral comment, or negative comment. Each cluster's centroid is represented by the symbol 'X'.

Figure 8: The clustered Replies with classified Users

Table 7, shows a summary of the information obtained from our study about user insights on Arabic twitter. This table comprises of required attributes. We provide a high and a low hierarchy for every users' power, as well as high and low users' interest. Moreover, presenting users' replies in many categories based on the trend. Thus, the outcomes of our study revealed that 45 percent of high-powered people recommend online education. Furthermore, 27 percent had neutral comments and the same percent for negative comments which helps a decision for the educational direction of the upcoming semester based on supervised/unsupervised learning

techniques and rough theory depend on mendelow model.

Table 7: User classification & replies

Power	Interest	D1:Power /Interest	D2: Action	Positive Reply	Negative Reply	Neutral Reply
Teacher	pedagogical	High/High	Manage Closely	التعليم عن بعد جيد distance learning is good		
Student	Sport	High/Low	Keep Informed	التعليم الاونلاين روعه Online Education is awesome		
Physiotherapist	Education	Low/High	Keep Satisfied		نظام التعليم اونلاين فاشل The online education is failure	
Pharmacist	Art	Low/Low	Minimum Effort			نفسى اكون زى الناس اللى بتفتخر بنظام تعليميها I want to be like the people who are proud of their education system

5. Conclusion and future work

People have come to rely on Twitter as a community for presenting their daily social life, with many users can post or share trends on popular social media networks. As a result, we suggested model based on a supervised learning technique for identifying and classifying social users according to their proximity to the trends on a specific field. Besides, an unsupervised machine learning technique for clustering their Arabic language responses/comments to a trend if it was positive or negative or neutral. Further, the importance of Social users was determined by utilizing Rough Set based on the power and interest matrix which depends on some attribute reduction algorithms such as (Johnson reduction and Genetic reduction). In addition, we have reached that we can improve the education system as a result of the appropriate decision that obtained from analysis for the social user's insights from Arabic twitter based on Data mining techniques depending on the rough set theory. For further study, we suggested utilizing lots of data to discover additional jobs which may be relevant to the same trend.

References

[1] AL-Rubaiee H, Qiu R, Alomar K and Li D," Sentiment Analysis of Arabic Tweets in e-Learning," JOURNAL OF COMPUTER SCIENCE, 12 (11), pp.553-563, 2016.

[2] A. H. Elsaid, R. K. Salem, H. M. Abdul-kader,"Automatic Framework for Requirement Analysis

Phase," 2016.

[3] G. BAI, L. LIU, BO SUN, J. FANG," A survey of user classification in social networks," IEEE INTERNATIONAL CONFERENCE ON SOFTWARE ENGINEERING AND SERVICE SCIENCES (ICSESS),SEPT. 2015.

[4] T. Hughes, T. Osman. ,G. Alwakid, "Challenges in Sentiment Analysis for Arabic Social Networks," pp. 89-100, November 2017.

[5] S. H. SHAIKH, L. M. R. J. LOBO," Revealing insights for sales based on analysis of Twitter product reviews," INTERNATIONAL CONFERENCE ON GLOBAL TRENDS IN SIGNAL PROCESSING, INFORMATION COMPUTING AND COMMUNICATION (ICGTSPICC),DEC.2016.

[6] P. VASHISTH, K. MEEHAN," Gender Classification using Twitter Text Data," 31ST IRISH SIGNALS AND SYSTEMS CONFERENCE (ISSC), JUNE.2020.

[7] J. ANTONIO,J. MORZAN,H. ALATRISTA,T. HERNANDAZ, AND J. BIAN, "Clustering and topic modeling over tweets: A comparison over a health dataset," IEEE INTERNATIONAL CONFERENCE ON BIOINFORMATICS AND BIOMEDICINE (BIBM), NOV.2019.

[8] H. AlSalman, "An Improved Approach for Sentiment Analysis of Arabic Tweets in Twitter Social Media," 3rd International Conference on Computer Applications & Information Security (ICCAIS),MARCH, 2020.

[9] Dainwei Chi, "Research on the Application of K-Means Clustering Algorithm in Student Achievement, "IEEE International Conference on Consumer Electronics and Computer Engineering(ICCECE),2021.

[10] H. Al-Rubaiee, K. Alomor," Clustering Students‘ Arabic Tweets using Different Schemes," In 2017 International Journal of Advanced Computer Science and Applications(IJACSA), Vol. 8, No. 4, 2017

[11] M. Saad, W. Ashour, "Arabic Text Classification Using Decision Trees," pp. 75-79, 2010.

[12] S. Larabi,N. Alalyani,S. Alotaibi, S. Ghouzali, and I. Abunadi, "Arabic Natural Language Processing and Machine Learning-Based Systems," vol. 7, pp. 7011-7020, 2019.

[13] E. Umargono, J. Endro, V. Gunawan, "K-Means Clustering Optimization using the Elbow Method and Early Centroid Determination Based-on Mean And Median," the International Conferences on Information System and Technology (CONRIST 2019), pages 234-240, Sept.2020.

[14] I. H. Witten, E. Frank, M. Hall, and C. J. Pal," Data Mining Practical machine learning tools and techniques," ,2016.

[15] M. Al-Mhairat,R. Alabbadi, R. Shaban, and A. AlQudab, "Performance Evaluation of Clustering Algorithms,",May.2019.

[16] P. NAGAMMA, H. R. PRUTHVI, K. K. NISHA, AND N H SHWETHA," An improved sentiment analysis

of online movie reviews based on clustering for box-office prediction," INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION & AUTOMATION,MAY.2015.

[17] Z. Pawlak, " Rough Sets," International Journal of computer and Information Sciences, Vol.11, No. 5, 1982.

[18] M. Bekkali, I. Sahmoudi, and A. Lachkar," Enriching Arabic tweets representation based on web search engine and the rough set theory," International Conference on Advances in Social Networks Analysis and Mining (ASONAM),Aug.2015.

[19] WANG, J. YANG, R. JENSEN, X. LIU, "Rough set feature selection and rule induction for prediction of malignancy degree in brain glioma", COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE, 83, PP.147-156, 2006

[20] C. LIAN, H. LIU, AND Z. WAN, "An attribute Reduction Algorithm Based on Rough Set Theory and An Improved Genetic Algorithm", JOURNAL OF SOFTWARE, 9(9), PP. 2276-2282, 2014.

[21] A. Elsaid,R. Salem, "A dynamic stakeholder Classification and Prioritization Based on Hybrid Rough-Fuzzy Method," Journal of Software Engineering, no. 11, p. 143–159, 2017.

[22] A Rough Set Rosetta Toolkit for Data Analysis and its functionality. available in: http://www.lcb.uu.se/tools /rosette/accessed 26 October 2013.