Research Article

Classification of Arabic Twitter Users: A Study Based on User Behaviour and Interests

Abdullatif Alabdullatif, Basit Shahzad, and Esam Alwagait

College of Computer & Information Sciences, King Saud University, Riyadh, Saudi Arabia

Correspondence should be addressed to Basit Shahzad; basit.shahzad@gmail.com

Received 11 November 2015; Accepted 18 January 2016

Academic Editor: Miltiadis D. Lytras

Copyright © 2016 Abdullatif Alabdullatif et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Social networks are among the most popular interactive media today due to their simplicity and their ability to break down the barriers of community rules and their speed and because of the increasing pressures of work environments that make it more difficult for people to visit or call friends. There are many social networking products available and they are widely used for social interaction. As the amount of threading data is growing, producing analysis from this large volume of communications is becoming increasingly difficult for public and private organisations. One of the important applications of this work is to determine the trends in social networks that depend on identifying relationships between members of a community. This is not a trivial task as it has numerous challenges. Information shared between social members does not have a formal data structure but is transmitted in the form of texts, emoticons, and multimedia. The inspiration for addressing this area is that if a company is advertising a sports product, for example, it has a difficulty in identifying targeted samples of Arab people on social networks who are interested in sports. In order to accomplish this, an experiment oriented approach is adopted in this study. A goal for this company is to discover users who have been interacting with other users who have the same interests, so they can receive the same type of message or advertisement. This information will help a company to determine how to develop advertisements based on Arab people's interests. Examples of such work include the timely advertisement of the utilities that can be effectively marketed to increase the audience; for example, on the weekend days, the effective market approaches can yield considerable results in terms of increasing the sales and profits. In addition, finding an efficient way to recommend friends to a user based on interest similarity, celebrity degree, and online behaviour is of interest to social networks themselves. This problem is explored to establish and apply an efficient and easy way to classify a social network of Arab users based on their interests using available types of information, whether textual or nontextual, and to try to increase the accuracy of interest classification. Since most of the social networking is done from the mobiles nowadays, the efficient and reliable algorithm can help in developing a robust app that can perform the tweet classification on mobile phones.

1. Introduction

The impetus for this project stems from the need for an effective method of classifying users on social networks, identifying each user's interests based on the similarity of these interests and finding relationships between users. The users of a social network like Twitter find it difficult to be sure about suggested friends without seeing if they have the same interests. The same applies at the organisational level: social networks need to be able to identify user groups to interact with them effectively, such as for targeted advertising of products. The main data that can be analysed for classification

are users' texts and posts, but performing content analysis on short texts is more difficult compared to long texts.

Twitter is one of the largest social networks globally, and it has excellent resources for sharing information and marketing, and it is also increasingly used for real-time interactions like discussions, news, and suggestions [1–4]. In addition to the other usages, the Arabic language is very well represented in Twitter; there were about 4 million active Arabic users on Twitter as of the end of 2012 [5]. There are about 22 Arab countries and millions of people who understand Arabic, since it is the language of the holy Quran. Despite extensive research, we could not find considerable

FIGURE 1: Classification of data-mining activities.

work published for classifying Arabic users based on their interests. Within a social network environment, especially Twitter, we encountered some challenges when we tried to classify users [6]. Profiles are generally ignored since most users are not concerned about their profiles or people insert inaccurate information [7–9]. Thelwall et al. [10, 11] have stated that people’s vocabularies change on social networks, since they may write different words in different ways. Different languages are used in tweets, and for each language, there are different ways of writing [10]. The text length limitation is one of the main challenges, because only 140 characters are allowed for a tweet [12, 13]. Attached links are a challenge because most tweets today include HTML links; the same applies to hashtags and symbols. The informal language is also a challenge, as it may include abbreviations and emoticons [14–16]. Finally, because we did not find related work on the classification of Arabic users on social networks, there is a knowledge challenge at the outset of this research.

The problem being addressed in this study considers that the amount of threading data is growing and producing analysis from this large volume of communications is becoming increasingly difficult for public and private organisations. One of the important applications of this work is to determine the trends in social networks that depend on identifying relationships between members of a community. This is not a trivial task as it has numerous challenges. Information shared between social members does not have a formal data structure but is transmitted in the form of texts, emoticons, and multimedia. The inspiration for addressing this area is that if a company is advertising a sports product, for example, it has a difficulty in identifying targeted samples of Arab people on social networks who are interested in sports. In order to accomplish this, an experiment oriented approach is adopted in this study. A goal for this company is to discover users who have been interacting with other users who have the same interests, so they can receive the same

type of message or advertisement. This information will help a company to determine how to develop advertisements based on Arab people’s interests. In addition, finding an efficient way to recommend friends to a user based on interest similarity, celebrity degree, and online behaviour is of interest to social networks themselves. This problem is explored to establish and apply an efficient and easy way to classify a social network of Arab users based on their interests using available types of information, whether textual or nontextual, and to try to increase the accuracy of interest classification.

This research provides potential benefits to advertising companies by giving a good guideline for targeting samples of people as well as for studying people’s preferences and trends. Companies can use this guideline to review their strategic plans, as well as encouraging potential users to follow them based on interests. Finally, the main contribution here is the novelty of this work for the Arabic language, which has not been considered before. Furthermore, we attempt in this work to establish a primary reference for work in other languages. Section 2 of the paper discusses the social mining process and Section 3 addresses the Arabic language process methods. The existing literature is presented in Section 4 of the paper that provides the logical grounds for carrying out this research. The discussion on the classification algorithms is carried out in Section 5 while the experiments and evaluation are carried out in Section 6. Section 7 discusses the results and findings of this study.

2. Classification of Twitter Users

Social mining is a subset of data-mining, which is studied under computer science disciplines (database, data analysis, statistics, data structure, and artificial intelligence or machine learning). The goal of data-mining is processing knowledge from data as mentioned in Figure 1.

From Figure 1, we can classify data-mining activities as follows:

- (i) Text-mining: here, the data is text (structured or unstructured).
- (ii) Web mining: the raw data include web content, links, and log files.
- (iii) Media mining: the raw data are images, video, and speech.
- (iv) Social mining: this is the focus of this research. It includes extracting trend patterns from streams of tweets or posts on a social network such as Twitter or Facebook. Social media data are vast, noisy, unstructured, and dynamic in nature.
- (v) Time series or “bioinformatics”: this includes identifying DNA sequences.

Structured and unstructured forms of data and multimedia need suitable algorithms to analyse and extract useful information from them through data-mining or knowledge discovery [17, 18]. Text-mining is a simple process of extracting knowledge from text. In this research, we need text-mining algorithms as part of our suggested solution at the level of user interest classification [19, 20]. This classification is based on the text in the tweets of users. The content of tweets is important in defining user’s interests. The classification process includes the following sentiment analysis activities:

- (1) Search for information access.
- (2) Monitor social media.
- (3) Group documents and web pages.
- (4) Classify news, stories, and web pages based on content.
- (5) Categorise emails and news.
- (6) Arrange databases of document-related meta-information for queries.
- (7) Get information about behavioural interactions between people, locations, and/or companies.
- (8) Check associations between the database entities.

When there are multiple documents to classify into four classes, for example, economics, sports, science, and lifestyle, there are two text-mining approaches to do that. In general, we can say classification algorithms can be divided into two main types as follows.

2.1. Rule-Based Approach. This is based on some rules applied to data entities inside data, like association rules, and it is suitable for structured data in databases or data warehouse-based classification. A very well-known example of this is the problem of an “item set” at a supermarket when some products are bought with others frequently; there is discovered knowledge from this relation, so it is important to arrange the placement of these two products.

2.2. Machine Learning-Based Approach. The machine learning-based approach uses the history from a set of example records that are categorised into sessions (training data), to keep an algorithm learning from previous knowledge, for

example, if there is an old database for customers. From this database, we can teach our classifier to detect the usual sample of ages for customers to predict whether a given person is a likely customer. However, the process of text-mining is not easy, as it has many challenges such as the following:

- (i) Information usually is not in a structured text form.
 - (ii) Database engines need more processing power to deal with large amounts of textual data.
 - (iii) A method must be chosen to determine all possible types of word senses in the language.
 - (iv) In text, there are complex relationships between concepts.
 - (v) Word ambiguity and context sensitivity create challenges.
 - (vi) There are multiple words for the same meaning: automobile = car = vehicle = Toyota.
 - (vii) It is difficult to determine a brand name from nouns like orange (the company) or orange (the fruit).
- (i) Noisy data, for example, spelling mistakes, make data more difficult to interpret.

Text files in general are semistructured; they require a lot of effort to remove stop words and less meaningful text. There are mainly three main types of text-mining classification: document classification, document clustering, and keyword based association rules. There are many techniques of text classification, but the best known include the following:

- (i) Support vector machines algorithm.
- (ii) *K*-nearest neighbors algorithm.
- (iii) Neural networks algorithm.
- (iv) Decision trees algorithm.
- (v) Association rule-based algorithm.
- (vi) Boosting algorithm.
- (vii) Naïve Bayes classifier algorithm.

As an example, in the Bayesian classifiers algorithm, building a text classifier is based on a probabilistic model and underlying word features in different classes. This concept includes making text classifications on probabilities for documents related to different classes by word presence classes and similarity in the texts [1].

3. Arabic Language Processing

The Arabic language is characterised by not being duplicable based on roots like English is, and this is something that increases the challenges. It is necessary to return to the root word to complete the automated process in Arabic. In addition, there is a multiplicity of dialects and multiples per word. There are a large number of letters in the Arabic language, and all 36 characters affect the meaning. The most prominent characteristics of the Arabic language in contrast with English are as follows:

- (i) Arabic letter forms depend on the letters before, after, or both or can be isolated.
- (ii) Saving and coding the data are complex.
- (iii) Dealing with the line when writing is different from western languages.
- (iv) The writing direction is from right to left.
- (v) The relationship between the operative and written language is different.
- (vi) Arabic characters are detailed in templates.
- (vii) Words may consist of more than one syllable.
- (viii) Letters in Arabic words touch each other.
- (ix) Arabic writing must be on the line to be one reference.
- (x) Derivations in Arabic are from the root (the origin of the word), and the root consists of a series of three letters or a quad.
- (xi) The presence of vowels in Arabic is key.
- (xii) The existence of private substitutes in Arabic is unique.

The Arabic language is different from English in many respects, so in the processing, we need to take into consideration that the stemmer results differ from rooter results regarding the same word in Arabic, and sometime this may change the meaning. For example, the word “wrote” in Arabic “يكتبون” gives different results in the Arabic rooter but not in the stemmer.

4. Related Work

This research aims to discover a way to classify Arabic users in social networks by studying Twitter users' properties and how they interact with each other and by determining accurate factors for classification. Kumar et al. [21, 22] have discussed some of the recent research in the Twitter domain and give a Twitter data analysis technique while Bollen et al. [21] have presented a contemporary analysis technique. Boyd et al. [23] show how to differentiate users by focusing just on their activity and ignoring the content of exchanged messages to give a user profile. A case study [24] focused on the UK 2010 General Election to determine “who you supported” from the content of tweets. Wu et al. [25] have looked at classifying trending topics on Twitter into 18 general categories. The earthquake in Japan was also one of the good applications of Twitter data analysis [26]; this study considered each Twitter user as a sensor and applied Kalman filtering and particle filtering, which are widely used for location estimation in ubiquitous/pervasive computing. An empirical study performed by Benhardus and Kalita [27] determined participants in a conversation by analysis of retweeting activity, mapping out retweeting as a conversational practice.

In addition, gender can be classified by identifying the text patterns, which can be observed by the work of Thelwall et al. [28–30]. This paper investigates statistical models for determining the gender of uncharacterised Twitter users. The

work of [31], “Who Says What to Whom on Twitter,” found that 50% of URLs consumed are generated by just 20,000 elite users and also found significant homophily within categories: celebrities listen to celebrities, while bloggers [32, 33] listen to bloggers, and so forth. This study noted the attention paid by different user categories to different news topics. The work found five distinct categories of retweeting activity in Twitter: automatic/robotic activity, newsworthy information dissemination, advertising and promotion, campaigns, and parasitic advertisement. Conover et al. have argued that a trend can be detected from streaming tweets from Twitter by accessing the Twitter API [34]. Tinati et al. have identified that, by using texts and tweeting behaviour, the locational source of tweets and the home locations of Twitter users can be found [35].

Lim and Datta [36] are arguers of the fact that automatic classifiers for Twitter users were built based on three different types of user, organisation, journalists/bloggers, and individuals, while Collier et al. [10, 37, 38] described a robust machine-learning framework for large-scale classification of users according to dimensions of interest, including Democrats, Republicans, and Starbucks aficionados. Rao et al. [39] investigate the political polarisation on Twitter in the USA in 2010. Althubaity et al. [40] analyse conversations around specific topics and identify key players in a conversation to get communicator roles in Twitter. The work of [41] classifies the celebrity of Twitter users by using Wikipedia in real time.

A comparison was made of a support vector machine (SVM) and Naive Bayes (NB) classification in making a syndromic classification of Twitter messages [42]; this study found that SVM is better than NB in four out of six syndromic classifications. The classification schemes where NB is found better are those that are significant to this study. In this context, the detection of Twitter user attributes is addressed in the work of Zubi [43], which was an exploration study of the attributes of user detection in Twitter using simple features such as n -gram models. Simple sociolinguistic features like the presence of emoticons, statistics about a user's immediate network like the number of followers and friends, and communication behaviour like retweeting frequency are also presented in the model.

In Arabic text classification, there are some works on document classification. A Saudi Arabian example called KACST introduces an overview and preliminary results for Arabic text classification [44]. There is also the automatic categorisation of Arabic documents based on the NB algorithm [45] which introduces a Naive Bayesian method, based on Chi-squares to categorise Arabic data. The work of [32, 33] uses web content mining techniques for Arabic text classification. Some of this work is similar to the current study, but not in Arabic and not concentrating on interest classification, such as the work of [32, 33], “Classification of Twitter Users Based on Following Relations.” The work of [45] is most relevant to our work but is not about Arabic.

The work of [32, 33] introduces a natural language processing- (NLP-) based approach to the classification of Twitter users to address how to discover new Twitter accounts to follow. Various approaches have been used to tackle this

FIGURE 2: Preliminary algorithm flowchart.

issue, including NB, language models, decision trees, and the MaxEnt model. The work of [45] provides Twitter relevance filtering via a joint Bayes classifier from user clustering. The overall accuracy of the collated classifier was around 75–85% based on the average results of all 25 users. A simple NB classifier from the NLTK natural language processing package reached around 70% accuracy. The advantage of this work over the toolkit implementation lies in the collated nature of the classifier, which strengthens the classification by bringing in extra information for each user's base Bayesian classifier. The work of [32, 33] uses a machine-learning approach to Twitter user classification by leveraging observable information such as user behaviour, network structure, and the linguistic content of the user's Twitter feed. This shows that rich linguistic features prove to be consistently valuable across three tasks and shows great promise for further user classification.

5. Classification Algorithm

The flowchart in Figure 2 shows the proposed algorithm. The algorithm first asks about which users are to be classified. Then, the algorithm requests a one-time access to the Twitter API and downloads the timeline for all the requested users.

It downloads a tweet for each user and checks whether the tweet is the last one, and if not, it will ask for another tweet.

The algorithm cleans each tweet by removing symbols, hashtags, stop words, and streaming. Because of the difficulty of classifying short texts, it collects all the cleaned tweets for each user in one document to make it easy to classify the latest and most efficient text classification algorithm, like a NB classifier algorithm or SVM. The algorithm is trained by a ready data set in the language, and the results for each user are stored by the algorithm in a suitable table in a database. Since tweets are unstructured data, it is necessary to convert the important results into a normalised database, which can be used as a data warehouse.

Since 22% of tweets include a URL [11], the efficiency of the classifier is increased by adding more data other than the tweets. The algorithm checks the type of each tweet—pure text tweet or including HTML links. Most tweets use external links but the problem is that these tweets use services like tiny URL because of the limitations on the length of a tweet. If a tweet includes HTML links, then the links processing will get the long URL, take its metadata, and add it to the document grouping all tweets. The algorithm uses these interests: politics, economy, sport, lifestyle, and religion. From any newspaper on the Internet, you can

see that the main categories of news match our five main interests.

The algorithm also has a profile classifier and a behaviour classifier. These are used for information about users to give a nontextual classification. From these two classifiers, we can get important classes of users and increase the efficiency of our classifier. The bio, if activated for a user, gives good knowledge of the user's character. Another factor for classification is profile fields, like the number of followers, which indicate if the user is a celebrity. This can be derived from the following equation [3]:

$$\begin{aligned} &\text{Celebrity degree for a user} \\ &= \frac{\text{number of followers}}{\text{number of people following}}. \end{aligned} \quad (1)$$

The algorithm checks if the user has a page in Wikipedia. This feature is explained elsewhere in detail [21]. The algorithm groups similar users together by calculating the similarity for each user from the results. From the above, we can categorise the classification based on three types of criteria:

- (i) *Textual Classification*. Collect all the tweets of each user and clean them from extra tags and links, so they can be considered as pure textual tweets.
- (ii) *Profile Classification*. Classify them based on profile attributes like age, location, and biography.
- (iii) *Behaviour Classification*. Classify them based on the hits behaviour of the user. For example, tweeting at midnight may indicate a younger user. If the user has not been active since they created the account, this may not be a personal account.

Empirical evidence is needed to find the percentage for each approach and determine the most important approach. A NB

algorithm is used with the Arabic language as the main classifier in this research work as mentioned in works [24–27]. We will explain the utilisation and application of this algorithm in our work. This classifier is based on statistical models, and the equation used is

$$P(C_i | D) = P(W_0 | C_i) \times P(W_1 | C_i) \times \dots \times P(W_{m-1} | C_i) \times P(C_i), \quad (2)$$

where C_i is the class, for example, sport (رياضة). D is a collected text tweet. $P(C_i | D)$ is the probability of classifying a tweet D in the class C_i . $W_0, W_1, \dots, W_{m-1}$ are the words in the tweet after cleaning and stemming and all text processing steps. $P(W_i | C_i)$ is the probability of finding the word W_i in class C_i . $P(C_i)$ is the probability of class C_i .

To see in detail how this classifier works for the Arabic language, here is an example.

$D = \text{”الدوري السعودي لكرة القدم اليوم”}$, which means “the Saudi football league today”:

$$\begin{aligned} P(\text{رياضة} | D) &= P(\text{الدوري} | \text{رياضة}) \\ &+ P(\text{السعودي} | \text{رياضة}) \\ &+ P(\text{الكرة} | \text{رياضة}) \\ &+ P(\text{القدم} | \text{رياضة}) \\ &+ P(\text{اليوم} | \text{رياضة}) + P(\text{رياضة}). \end{aligned} \quad (3)$$

The calculation proceeds in this way:

$$P(\text{الدوري} | \text{رياضة}) = \frac{[\text{Number of occurrences of the word (الدوري) in the class (رياضة)}]}{[\text{Number of words in the class (رياضة)}]}. \quad (4)$$

It continues in this way until we get the probabilities of classifying the tweet to each class, from which we get the maximum. Thus, the main steps in training the Naive Bayesian model include the following:

- (i) Collecting a set of texts for each class.
- (ii) Preprocessing the text by cleaning it, deleting stop words, and stemming.
- (iii) Calculating the basic probabilities of the frequency of keywords by using the above equations.
- (iv) Saving the results in the database as training sets.

By adopting this approach, we are able to classify any text and get the class probabilities immediately. After those steps,

we need the nearest class using the similarity calculation methods with each class to determine the main class by using this equation:

$$\begin{aligned} \text{Similarity} &= \cos(\theta) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \\ &= \frac{\sum_{i=1}^n A_i \times B_i}{\sqrt{\sum_{i=1}^n (A_i)^2} \times \sqrt{\sum_{i=1}^n (B_i)^2}}. \end{aligned} \quad (5)$$

This is the similarity between user A and user B . We applied the above classification method to classify all tweets. Then, the calculation of the percentage of each tweet in each class is done. For example,

- (i) 27% for sport “رياضة”
- (ii) 25% for politics “سياسة”
- (iii) 0% for others.

Thus, each user is represented by a vector of interests; each item in the vector represents the percentage of interest of the user in a certain class; for example,

- (i) $A = (75\%, 25\%, 0\%, 0\%, 0\%)$,
- (ii) $B = (10\%, 10\%, 10\%, 60\%, 10\%)$,

where A is the interests of the first user and B is the interests of the second user. The item in each vector is the interest in a certain class, so we calculate the similarity using

$$\text{sim}(A, B) = \frac{[0.75 \times 0.10 + 0.25 \times 0.10 + 0 \times 0.1 + 0 \times 0.6 + 0 \times 0.1]}{[\sqrt{0.75^2 + 0.25^2}] + [\sqrt{0.1^2 + 0.1^2 + 0.1^2 + 0.6^2 + 0.1^2}]} \quad (6)$$

The final result is a number between 0 and 1, where 0 means there is no similarity at all and 1 means exactly the same interest. The results have been presented as a percentage for clarity.

6. Evaluation and Experiments

We collected many texts related to each topic from news websites like <http://www.kooora.com/> for sport, <http://skynewsarabia.com/> for politics, <https://www.aliqtisadi.com/> for economy, <http://www.ahadith.net/> for religion, and so forth. We used 1500 articles for testing, and the results are given in Table 1.

The results for the text classification of documents for tweets in the classifier must be different: the method of experimentation is to collect a corpus of testing data based on very well-known Twitter users in Arab countries with

TABLE 1: Sample results for document classifier.

	Number of items	Correct	Wrong	Accuracy (%)
Sport	300	296	4	98.7%
Technology	300	274	26	91.3%
Religion	300	277	23	92.3%
Economy	300	273	27	91.0%
Politics	300	296	4	98.7%
All	1500	1416	84	94.4%

different interests, pass those users into the classifier, and compare the results with what can be known about those influencer users. To evaluate the system, we collected some data from a real-world influencer in the Twitter social network, and after that we used the classifier to check the accuracy. For each interest, we had ten users, and the corpus of data is shown in Table 2.

For each class, we collected 20 active Twitter users with an average number of tweets of 4000 per user and accessed the Twitter API using our own application to get the stream of tweets and profile information for each user. These were stored in our system as a text file. By running the following classifier equation, the results are obtained:

$$\text{Average (class } x) = \frac{\text{average of class results}}{\text{number of samples}} \quad (7)$$

Usually, when we talk of individuals, 100% interest match, and as we have five interests, class I is used to normalise the results. So the main class values will be multiplied by 2. This is because if class x has more than 50% then the other class will not have the highest percentage, so we need to multiply the result by two, and for each result that is more than 50% we will set it to 50%:

$$\text{Accuracy (class } x) = \text{average of users results} \times 2 \quad (8)$$

The calculation is as follows:

$$\begin{aligned} & \text{Average (Religion)} \times 2 \\ &= \left(\frac{(50 + 50 + 50 + 50 + 37 + 49 + 50 + 50 + 37 + 50 + 50 + 34 + 50 + 50 + 50 + 50 + 50 + 50 + 48)}{20} \right) \times 2 \\ &= 95.5\%, \\ & \text{Average (Politics)} \times 2 \\ &= \left(\frac{(50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 50)}{20} \right) \times 2 \\ &= 100\%, \\ & \text{Average (Sport)} \times 2 \\ &= \left(\frac{(50 + 50 + 22 + 25 + 30 + 36 + 50 + 50 + 50 + 35 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 23 + 50)}{20} \right) \times 2 \\ &= 84.6\%, \end{aligned}$$

TABLE 2: Sample of Twitter user collections.

Number	Politics user	Religion users	Economy users	Sport users	Technology user
1	kasimf	Abdulaziztarefe	Alwaleed.Talal	faisalbinturkil	MeetTechnology
2	anwarmalek	NabilAlawadhy	AbAmri	alnassr_news	Applewd
3	Yzaatreh	al_rasekhoon	cnnarabic	Altemyat	Technya
4	Recept_Erdogan	SalehAlmaghamisi	Reuters_Busines	mustafa_gha	SafaTeqnia
5	ElBaradei	mishari_alafasy	Hamzaalsalem	SamiAlJaber	Arabapps
6	Adeeb_Emad	mohamadalarife	Alhayat_Bus	nawaffinfaisal	NokiaKSA
7	IsmailHaniyyeh	afaaa73	essamz	battalalgoos	Android_arab
8	AzmiBishara	Saudalshureem	aleqt_fb	Alhila_FC	alwagait
9	SafaNews	Shugairi	dubaiFinancials	k_alshenaif	COEIA_KSU
10	alnilanyq84ever	Abuabdelelah	SkyNewsArabiaBs	waleedalfarraj	3bdulla
11	Politic_affairs	Shaikh_alQattan	aleqtisadiah	AlArabiya_spt	techwd
12	amremoussa	Asowayan	qunaibet	ActionYaDawry	saudigamer
13	alhayatdaily	BenJebreen	Agary4u	Almoj_alazra8	RayzCo
14	LebPolitician	MaherAlMueaqly	CNBCArabia	realmadridarab	GoogleArabia
15	AJArabic	Hwsh1434	RashidALFowzan	BarcelonaAR	estidafaty
16	Elssisy	Khalid_aljulyel	tfrabiah	ryadda	mSaudiCommunity
17	SkyNewsArabia_B	ala7adeth	SaudiMCI	ESN_EgySports	IntelGet
18	iranianaaffairs	NfaeesAlelm	Riy_Econ	sadaalmalaeb	Tiqaniat
19	JKhashoggi	BINTIMIAH	MubasherSA	ReutersSport	iPhoneIslam
20	Ahmadmuaffaq	islamdor	BorsahNews	CityArabia	akhbar_tech

Average (Technology) $\times 2$

$$= \left(\frac{(50 + 50 + 50 + 50 + 50 + 50 + 50 + 50 + 48 + 39 + 50 + 50 + 47 + 50 + 50 + 50 + 50 + 50 + 50 + 50)}{20} \right) \times 2$$

$$= 98.4\%,$$

Average (Economy) $\times 2$

$$= \left(\frac{(37 + 50 + 7 + 2 + 26 + 50 + 17 + 50 + 50 + 50 + 44 + 19 + 50 + 50 + 39 + 50 + 40 + 50 + 50 + 50)}{20} \right) \times 2$$

$$= 78.1\%.$$

(9)

So the accuracy of the classifier will be the average of the accuracy of the five classes:

$$\text{Accuracy} = \frac{95.5 + 100 + 79.6 + 84.6 + 98.4 + 78.1}{5} \quad (10)$$

$$= 90.32\%.$$

7. Findings and Results

The main problem addressed in this work is classifying Arab users in social networks. This research proposes a new model of an automatic suggestion mechanism for social network users based on three criteria: posts (tweets), celebrity degree, and tweeting behaviour (number of tweets). This model depends on these three aspects and may help social network

companies or users themselves to determine suitable friends from millions of users in social networks.

Figure 3 was compiled using the data in Table 3. In Figure 3, note that the trend line function that separates the favourite users to follow is based on the three factors. The size of a ball is the interest degree (percentage) for a class, religion, for example, and the other axes show the celebrity degree and tweeting behaviour. Finally, it can be determined that there are two main factors we need to consider to improve the classification of text in social networks: performance and accuracy. These are discussed as follows.

7.1. Performance. The speed-up due to parallelisation is very important because most works in this field use sequential versions of algorithms, but some sequential algorithms cannot be adapted as a parallel version. Some algorithms work better in

FIGURE 3: Suggestion for users to follow.

the parallel version but other algorithms perform better in a sequential version. So, this fact is required to be considered as well. To get a virus signature (structure and stream of text files), the task is to access the contents of text files, get all the words, stem them, and use an NLP process, which can take a long time. It has been discovered that it is sufficient for finding the signature of a file to use a simple and available function, like a hash, that gives only a numerical value. We just need to run a similarity check between these values to detect the class.

7.2. Accuracy. The contents of documents may be related to other classes not determined in our classifier class, which reduces the accuracy. For example, say we determine that A is related to sports by 90% and 10% for politics. This identification is only for the words that are known by our classifier. The solution is to use a fuzzy logic algorithm on the unknown tokens and calculate whether the tweet should be moved to others. Alternatively, a cluster algorithm with the following equation can be applied:

$$\text{Class} = \text{percentage of others} - \text{percentage of discovered classes.} \quad (11)$$

In addition, there is the problem of negative prefixes, as in “not sport,” so there is need for semantic and sentiment analysis since suffix and prefix tokens may change the meaning of any token.

8. Conclusion and Future Work

This research work benefits from the integration of statistical science, artificial intelligence, and data-mining and tries to provide accurate algorithms. The focus of this work is designing and building a highly accurate classification of Arabic Twitter users. The proposed user classifier can help social scientists, teachers, companies, and governments to classify users of social networks or in learning by experiment. A supervised approach for texts using profile properties to classify users is presented. It is applicable for the social network Twitter but also may be useful for other social networks.

TABLE 3: Sample of religious users.

z (interest)	x (celebrity degree)	y (behaviour)
24	500,000	62
22	600,000	81
10	150,000	22
20	680,000	25
20	95,000	30
24	250,000	36
5	550,000	52
24	122,000	78
22	200,000	62
20	250,000	35
20	500,000	90
6	600,000	70
5	550,000	80
10	90,000	86
4	400,000	97
3	165,000	76
20	90,000	73
24	880,000	62
2	900,000	23
9	100,000	83
24	500,000	62
22	600,000	81
10	150,000	22

Through this application, we accessed the streaming posts of Arabic Twitter users. After normalisation and stemming of the text of a tweet, it was ready for further processing. We extracted the features of the users and added them to a database and text files. The classifier was then applied on the stored data of user contents. The NB classifier is used as multinomial classifier to detect five classes (sport, religion, economy, politics, and technology) in Arabic with 90% accuracy. There are many applications for this classifier, like recommending users to follow on Twitter based on textual content, tweeting behaviour, and celebrity degree and studying trends on social networks. Furthermore, measuring retweeting activity is important for influencing weights. The application of the algorithm has significantly improved the accuracy and the performance of the classification. The speed-up due to parallelisation is very important because most works in this field use sequential versions of algorithms, but some sequential algorithms cannot be adapted as a parallel version. Some algorithms work better in the parallel version but other algorithms perform better in a sequential version. By applying the proposed algorithm, the accuracy of the system has also increased. The efficient mobile app for this algorithm can help the effective tweet classification on the go since most of social networking is done on mobiles nowadays.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

This work is supported by the Research Centre of College of Computer and Information Sciences in King Saud University. The authors are grateful for this support.

References

- [1] A. M. Kaplan and M. Haenlein, "Users of the world, unite! The challenges and opportunities of Social Media," *Business Horizons*, vol. 53, no. 1, pp. 59–68, 2010.
- [2] D. Scanzfeld, V. Scanzfeld, and E. L. Larson, "Dissemination of health information through social networks: Twitter and antibiotics," *American Journal of Infection Control*, vol. 38, no. 3, pp. 182–188, 2010.
- [3] N. Eltantawy and J. B. Wiest, "Social media in the Egyptian revolution: reconsidering resource mobilization theory," *International Journal of Communication*, vol. 5, 18 pages, 2011.
- [4] R. M. Merchant, S. Elmer, and N. Lurie, "Integrating social media into emergency-preparedness efforts," *The New England Journal of Medicine*, vol. 365, no. 4, pp. 289–291, 2011.
- [5] A. S. M. Report, *Arab Social Media Report*, 2014, <http://www.arabsocialmediareport.com/home/index.aspx>.
- [6] D. Zhang and G. Guo, "A comparison of online social networks and real-life social networks: a study of Sina Microblogging," *Mathematical Problems in Engineering*, vol. 2014, Article ID 578713, 6 pages, 2014.
- [7] E. Alwagait, B. Shahzad, and S. Alim, "Impact of social media usage on students academic performance in Saudi Arabia," *Computers in Human Behavior*, vol. 51, pp. 1092–1097, 2015.
- [8] B. Shahzad and E. Alwagait, "Does a change in weekend days have an impact on social networking activity?" *Journal of Universal Computer Science*, vol. 20, no. 15, pp. 2068–2079, 2014.
- [9] B. Shahzad, E. Alwagait, and S. Alim, "Impact of change in weekend days on social networking culture in Saudi Arabia," in *Proceedings of the 2nd International Conference on Future Internet of Things and Cloud (FiCloud '14)*, pp. 553–558, IEEE, Barcelona, Spain, August 2014.
- [10] M. Thelwall, K. Buckley, and G. Paltoglou, "Sentiment strength detection for the social web," *Journal of the American Society for Information Science and Technology*, vol. 63, no. 1, pp. 163–173, 2012.
- [11] I. Himelboim, S. McCreery, and M. Smith, "Birds of a feather tweet together: integrating network and content analyses to examine cross-ideology exposure on Twitter," *Journal of Computer-Mediated Communication*, vol. 18, no. 2, pp. 40–60, 2013.
- [12] E. Fischer and A. R. Reuber, "Social interaction via new social media: (How) can interactions on Twitter affect effectual thinking and behavior?" *Journal of Business Venturing*, vol. 26, no. 1, pp. 1–18, 2011.
- [13] X. Hu and H. Liu, "Text analytics in social media," in *Mining Text Data*, pp. 385–414, Springer, New York, NY, USA, 2012.
- [14] M. H. Davis and I. S. Johnsrude, "Hierarchical processing in spoken language comprehension," *The Journal of Neuroscience*, vol. 23, no. 8, pp. 3423–3431, 2003.
- [15] O. M. Foong, A. Oxley, and S. Sulaiman, "Challenges and trends of automatic text summarization," *International Journal of Information and Telecommunication Technology*, vol. 1, no. 1, pp. 10–20, 2010.
- [16] A. Munigal, "Use of microblogs in India: a study of twitter usage by librarians and in libraries," *Journal of Library Administration*, vol. 54, no. 7, pp. 590–608, 2014.
- [17] H. Woo, E. Kang, S. Wang, and K. H. Lee, "A new segmentation method for point cloud data," *International Journal of Machine Tools and Manufacture*, vol. 42, no. 2, pp. 167–178, 2002.
- [18] W. He, S. Zha, and L. Li, "Social media competitive analysis and text mining: a case study in the pizza industry," *International Journal of Information Management*, vol. 33, no. 3, pp. 464–472, 2013.
- [19] A. Haidar, S. Naoum, R. Howes, and J. Tah, "Genetic algorithms application and testing for equipment selection," *Journal of Construction Engineering and Management*, vol. 125, no. 1, pp. 32–38, 1999.
- [20] R. Malik, L. Franke, and A. Siebes, "Combination of text-mining algorithms increases the performance," *Bioinformatics*, vol. 22, no. 17, pp. 2151–2157, 2006.
- [21] J. Bollen, H. Mao, and X. Zeng, "Twitter mood predicts the stock market," *Journal of Computational Science*, vol. 2, no. 1, pp. 1–8, 2011.
- [22] S. Kumar, F. Morstatter, and H. Liu, *Twitter Data Analytics*, Springer, New York, NY, USA, 2014.
- [23] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: conversational aspects of retweeting on twitter," in *Proceedings of the 43rd Annual Hawaii International Conference on System Sciences (HICSS '13)*, pp. 1530–1605, Honolulu, Hawaii, USA, January 2010.
- [24] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 1301–1309, Association for Computational Linguistics, Edinburgh, UK, July 2011.
- [25] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts, "Who says what to whom on twitter," in *Proceedings of the 20th International Conference on World Wide Web (WWW '11)*, pp. 705–714, ACM, April 2011.
- [26] R. Ghosh, T. Surachawala, and K. Lerman, "Entropy-based classification of 'retweeting' activity on Twitter," <http://arxiv.org/abs/1106.0346>.
- [27] J. Benhardus and J. Kalita, "Streaming trend detection in Twitter," *International Journal of Web Based Communities*, vol. 9, no. 1, pp. 122–139, 2013.
- [28] M. Thelwall, D. Wilkinson, and S. Uppal, "Data mining emotion in social network communication: gender differences in MySpace," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 1, pp. 190–199, 2010.
- [29] J. Mahmud, J. Nichols, and C. Drews, "Where is this tweet from?: Inferring home locations of Twitter users," in *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM '12)*, pp. 511–514, Dublin, Ireland, June 2012.
- [30] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," *Journal of Sociolinguistics*, vol. 18, no. 2, pp. 135–160, 2014.
- [31] M. De Choudhury, N. Diakopoulos, and M. Naaman, "Unfolding the event landscape on twitter: classification and exploration of user categories," in *Proceedings of the ACM Conference on Computer Supported Cooperative Work (CSCW '12)*, pp. 241–244, ACM, Seattle, Wash, USA, February 2012.
- [32] M. Pennacchiotti and A.-M. Popescu, "Democrats, republicans and starbucks aficionados: user classification in twitter," in *Proceedings of the 17th International Conference on Knowledge*

- Discovery and Data Mining (ACM SIGKDD '11)*, pp. 430–438, San Diego, Calif, USA, August 2011.
- [33] M. Pennacchiotti and A.-M. Popescu, “A machine learning approach to twitter user classification,” in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, pp. 281–288, Barcelona, Spain, July 2011.
 - [34] M. Conover, J. Ratkiewicz, M. Francisco, B. Goncalves, F. Menczer, and A. Flammini, “Political polarization on twitter,” in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM '11)*, Barcelona, Spain, July 2011.
 - [35] R. Tinati, L. Carr, W. Hall, and J. Bentwood, “Identifying communicator roles in twitter,” in *Proceedings of the 21st ACM International Conference Companion on World Wide Web*, pp. 1161–1168, 2012.
 - [36] K. H. Lim and A. Datta, “Interest classification of Twitter users using Wikipedia,” in *Proceedings of the 9th International Symposium on Open Collaboration*, ACM, Hong Kong, August 2013.
 - [37] N. Collier and S. Doan, “Syndromic classification of Twitter messages,” in *Electronic Healthcare*, pp. 186–195, Springer, Berlin, Germany, 2012.
 - [38] G. Paltoglou and M. Thelwall, “Twitter, MySpace, Digg: unsupervised sentiment analysis in social media,” *ACM Transactions on Intelligent Systems and Technology*, vol. 3, no. 4, article 66, 2012.
 - [39] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, “Classifying latent user attributes in twitter,” in *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC '10)*, pp. 37–44, ACM, October 2010.
 - [40] A. Althubaity, A. Almuhareb, S. Alharbi, A. Al-Rajeh, and M. Khorsheed, “KACST Arabic text classification project: overview and preliminary results,” in *Proceedings of the 9th IBIMA Conference on Information Management in Modern Organizations*, Marrakech, Morocco, January 2008.
 - [41] M. El Kourdi, A. Bensaid, and T. Rachidi, “Automatic Arabic document categorization based on the Naïve Bayes algorithm,” in *Proceedings of the 20th Workshop on Computational Approaches to Arabic Script-Based Languages (COLING '04)*, pp. 51–58, Association for Computational Linguistics, August 2004.
 - [42] F. Thabtah, M. Eljinini, M. Zamzeer, and W. Hadi, “Naïve Bayesian based on Chi Square to categorize Arabic data,” in *Proceedings of the 11th International Business Information Management Association (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies*, Citeseer, Cairo, Egypt, January 2009.
 - [43] Z. S. Zubi, “Using some web content mining techniques for Arabic text classification,” in *Recent Advances on Data Networks, Communications, Computers*, pp. 73–84, WSEAS, 2009.
 - [44] T. Yamashita, H. Sato, S. Oyama, and M. Kurihara, “Classification of twitter users based on following relations,” in *Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS '13)*, Hong Kong, March 2013.
 - [45] A. L. Churchill, E. G. Liodakis, and S. H. Ye, “Twitter relevance filtering via joint bayes classifiers from userclustering,” *Journal of University of Stanford*, 2010.