Cognitive Computation (2019) 11:71–86
https://doi.org/10.1007/s12559-018-9592-7

# A Study of Arabic Social Media Users—Posting Behavior and Author's Gender Prediction

**Abdulrahman I. Al-Ghadir**[1] · **Aqil M. Azmi**[1]

Received: 16 August 2016 / Accepted: 4 September 2018 / Published online: 24 September 20 18


**Abstract**

Social media opens up numerous possibilities to study human interaction and collective behavior in an unprecedented scale. It opened a whole new venue for research under the name "social computing". Researchers are interested in profiling individuals (e.g., gender, age group), groups, community, and networking. We are interested in studying the collective behavior of Arabic social media users. Most studies covering Arabic social media has focused on the sentiment analysis of, say tweets. This study, however, looks into *who* and *when* users interact with the Arabic social media. Specifically, there are two objectives of this work. First, studying the demographic posting behavior of social media users from two different perspectives: gender and educational level. Second, author profiling. Identifying author's gender of a social media post. We use Saudi Arabia, a very prolific country when it comes to social media in general, as a backdrop for this study. The results in this study are based on mining huge amount of metadata of a popular local social media forum covering the period 2011–14 inclusive. The extracted features (normalized list of $k$ highest scoring words, and likewise for stems) from the posts were used to train classifiers to identify the author's gender. We used two different classifiers, Support Vector Machine (SVM) with linear kernel and 1-NN (1-nearest neighbor), and experimented with different sizes for the list of features. When the number of features (size of the features vector) is small ($\leq 50$) both classifiers perform equally well in identifying the author's gender, but we risk overfitting the data. The classifiers achieved their best result when using 100 features. The 1-NN classifier delivered a better performance, achieving a balanced accuracy of 93.16% vs 87.33% for the SVM in predicting the author's gender. And for a larger set of features, SVM delivered a better performance and more stable behavior than 1-NN, but still nowhere close to its best performance. We used $t$ test to confirm our assessment that the difference between the performance of both classifiers is statistically significant. Based on that, we recommend using 100 features, and we get our best result using 1-NN with a balanced accuracy of 93.16%.

**Keywords** Behavior study · Arabic social media · Social media mining · Informal Arabic · Gender identification · Author profiling

## Introduction

We live today in an "information society," in which individuals, governments, and organizations alike generate and accumulate enormous amounts of data that can no longer be analyzed manually. Data mining, with its various algorithms and applications, is therefore a natural response to this information barrage. According to the McKinsey Global Institute, the amount of generated data will see an annual increase of 40% worldwide, and already companies with over 1000 employees have stored at least 200 TB of data [26]. Social media plays a significant role in the daily lives of many people and organizations. According to [2], social media refers to a combination of three different elements. These are content, user communities, and Web 2.0 technologies. Quoting Jue et al., "Social media has been broadly defined to refer to 'the many relatively inexpensive and widely accessible electronic tools that enable anyone to publish and access information, collaborate on a common effort, or build relationships'" (see [24], p. 4). Social media operates under the many–to–many (multiple source to multiple receivers)

✉ Aqil M. Azmi
aqil@ksu.edu.sa

Abdulrahman I. Al-Ghadir
abdulrahman.alghadir@gmail.com

[1] Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

model which makes it more appealing and popular than the traditional media which operated under the 1–to–many model. Interestingly, just a few years back, a debate lingered over whether or not the social media is a fad.

Social media grew and evolved largely due to free platforms, e.g., WordPress, LinkedIn, Facebook, ... etc.; however, many wrongly associate the social media with Facebook and Twitter. We may attribute this perception due to penetration of active users. In Q1/2013, Facebook topped the social media sites with 51% penetration, and Google+ a distant second with penetration of 26% [22]. In the same report, LinkedIn was ranked 12th with only 8%. In his book, Standage [37] argues that social media is an old phenomenon. He goes on citing ancient graffiti as one such example. In the digital era, forums were one of the earliest incarnation of the modern social media, which was later morphed into blogs. Though the forums have been around for many years, it was the affordable smart mobile phones and devices that created a shift in the way we communicate. Some businesses recognized this new media as a new venue for their traditional marketing, and embraced the new media early on. This further helped the new media survive the early years. Facebook has nearly 1.5 billion users who log in at least once a month, and recently Mark Zuckerberg boasted over a billion people used Facebook on a single day [16]. A study by PEW's research Center on Internet & American Life Project showed an 800% rise of social media usage among US adults in just 8 years. Social media has been very popular amidst the Arab Spring, with 1.4 million mention of hashtag `#Egypt` during the first 3 months of 2011. And this not counting the tweets which were in Arabic. A report by Dubai School of Government gives empirical credence that Facebook and Twitter abetted if not enabled this region-wide 2011 uprising [20].

One of the biggest area-wise countries in the Middle East, Saudi Arabia sits atop one of the largest known oil reserves in the world and is a vast country that is sparsely populated (32 million in an area over 2 million square km). Besides, Mekkah and Madina, the two most sacred cities for Muslims, are in Saudi Arabia. The population in the country is quite young, with roughly half the population under the age of 25 [40]. For the outside world, little is known about the country, and what makes our study important, is that there are no organizations (official or private) that have access to Saudi public in the form of survey/polling. We were able to mine the metadata of a large local social media site, and compile the results in this work.

Worldwide, the youth are the prime users of the so many social media sites, and Saudi Arabia is no exception. Twitter was ranked seventh most visited sites in Saudi Arabia, and with a 42% increase in the number of account holders between the period Q1/2013 and Q2/2012, the country ranked second among the world's fastest growing countries with Twitter [22, 23]. The lack of public expression in the country is what made Twitter appeal to the Saudis. As one social business strategist said, "Twitter appeals to the Saudi user. He just wants to say what is on his mind, float an idea, debate and discuss it and jump onto a new subject" [23]. Another interesting fact, Arabic is the fastest growing language on Twitter and Saudi Arabic accounts for 30% of global Arabic tweets [12], with 90% of the tweets in Saudi Arabia are in Arabic [23]. Moreover, Saudi Arabia ranked 7th worldwide in the number of individual accounts in social media, with an average of seven accounts per person [9]. The same news mentioned that 26% of the Saudi adolescents use Snapchat, which places the country 8th worldwide in the number of Snapchat users. And with an average of seven videos per day per Internet user, Saudis watch over 90 million YouTube videos daily [12], which places Saudi Arabia as the country with the highest YouTube videos viewership in the world.

The worldwide growth in the number of users of social media has led to a massive increase in the size of social data. In turn, this has led to increased interest in researching this field. There exists much research covering different aspects of social media. These aspects include age group estimation, demographic interests, gender, ... etc. Most of the studies in the field of social media have focused on Twitter [10, 11]. Some of the works that focused on gender classification for the English language are [27, 29, 33]. Few works have addressed the Arabic language in the social media [1, 5]. In future, we expect more research into social media directed towards addressing specific questions, e.g., age or gender of the user, fueled by interest in social media advertising which is set to explode. In just 2014, Youtubers spent about one billion minutes watching its top 10 ads [25], that is an equivalent of approximately 8000 man-years. In the USA alone, 60% of the time spent accessing digital media is through smart phones and tablets. So, the next big wave in advertising is reaching the users while they are on their mobile devices, and social media advertising is the best native option.

This paper is a side work of an ongoing project, detecting the dynamics of community behavior through mining big data. Our main interest is the Arabian Gulf countries (GCC), and Saudi Arabia in particular. Saudi Arabia has long been regarded as a conservative country; however, there is a changing trend among part of the society. Disturbing trend of extreme thinking and abnormal behavior (e.g., child pornography) was unheard of less than two decades ago. We, as computer scientists, are interested in ways to forecast upcoming behaviors. We believe that mining the local social media contents will help in detecting the dynamics of community behavior, and possibly predict the trend of a behavior over time. It is possible to spot some new unexpected behavior. Vinciarelli et al. [39] discussed

modeling, analysis, and synthesis of human behavior for technologies that deal with human-human and human-machine interactions. We are interested in community behavior, which may be regarded as interaction in a large social group. It is possible that we may adapt some of the phases used for modeling and analyzing human behavior for studying community behavior. The five phases are [39] (1) data design and collection, (2) behavior analysis, specially modeling interpersonal influence, (3) multimodal embodiment, (4) computational models of interaction, and (5) applications. Each phase has its own challenges which need to be addressed and solved.

For our study of community dynamics, we start by building a dataset which targets the community we wish to analyze. In this paper, we focus on studying the collective usage of Arabic social media based on two factors: gender and education level. We closely look at how different genders interact with the social media for the 4-year period, 2011–2014. We also look at the educational level of users and their interaction behavior for the same period. The interaction is limited to the posting activity (using the time stamp of the post), i.e., time spent browsing is not looked at. This has more to do with the data provided by the social media site we used in this study. One interesting observation, many posting activities were centered around the daily prayer time.

Cai and Chen [18] improved the quality of the retrieved documents by inferring the purpose behind the search operation which is based on tracking user's past behavior. We have constructed a model which is used to recognize the gender of social media users based on their behavior. Author profiling is a problem of growing importance in applications in forensics, security, and marketing. Author profiling distinguishes between classes of authors studying their sociolect aspect, that is, how language is shared by people. This helps in identifying profiling aspects such as gender, age, etc. The particular profile dimensions we consider here are author's gender. That is, given a post, we are interested in identifying the author's gender. Our approach takes into account the psycholinguistic features of text, and for this we use top-$k$. The top-$k$ makes use of $k$ most relevant words based on their occurrences and the frequency of the stems. Experimentally, we achieved an accuracy of 87.33% using linear SVM.

The rest of this paper is divided as follows. In "Related Work," we examine some related works focusing on feature extraction and classification methods. "Dataset Used in this Study" explains the dataset for the Arabic language, which is generated from a widely used local social forum. In "Analyzing the Posting Behavior," we analyze the behavior of social media users. In "Author Profile: Gender Identification," we introduce the feature vector representation (top-$k$) for our gender classification problem. Results and discussion are presented in "Performance Evaluation of Our Gender Identifier." Finally, we conclude in "Conclusion."

## Related Work

We will review the different approaches used for feature extraction which is needed to identify the gender in social media sites, mainly for Twitter. The gender prediction model relies on two different categories of features: behavior-related and statistical-related features. In the former, the focus is on the lexical and behavior of the social users based on the text, images, metadata, . . . etc., while in the latter, the focus is on some statistical parameters for the words and characters in the text.

We start with two recent works that deal with Arabic Twitter users. In [4], the authors were set on determining the trends in social networks, Twitter in particular, which depends on identifying the relationships between members of a community. Their approach was used to discover the interests of users to help in selecting the best fitting advertising. Also, finding an efficient way to recommend friends to a user based on interest similarity, celebrity degree, and online behavior is of interest to social networks themselves. Their classification algorithm used Twitter API to retrieve tweets of requested users as their dataset. These tweets were passed to their system which they applied preprocessing and classification of behavioral and profile. The authors used Naïve Bayesian model for the classification. The reported balanced accuracy was 94.4%. Aldayel and Azmi [5] proposed a hybrid approach to analyze the sentiment of Arabic tweets. Their approach combined semantic orientation and machine learning to identify the polarity of Arabic tweets. To save time manually labeling the training data, the authors used the lexical-based classifier to do this task. The output of this classifier was then used as a training data for the SVM classifier. Experiments by the authors showed that neither of the individual classifiers, lexical-based or SVM, fall short of the performance of their hybrid system. Their system accuracy was 84.01%.

Reviewing works that deal with behavior-related features, Pennacchiotti and Popescu [33] introduced a system to infer user attributes, e.g., ethnicity, using the user behavior information network structure, and the linguistic content of the tweet. They relied on four feature classes: user profile, user tweeting behavior, linguistic content of user messages, and the user social network features. The authors introduced profile-based features to identify the users' gender. Based on data from 15000 random users, they concluded that profile fields are insufficient, as it does not contain enough good-quality information, for user classification purposes. In [29], the authors introduced a feature set which maximized

the accuracy of gender and age prediction. The feature set had two different categories, content-based features and stylistic features. The content-based features include sentiment among others, while the stylistic features include readability, spelling and Grammatical Errors, ... etc. And for gender classification, the authors also employed a system of heuristics-based adjustments for gender prediction using a customized lexicon of phrases. The system was evaluated on four online media genres: blogs, tweets, hotel reviews, and an unspecified social media. Training SVM on their feature set, they reported 71.15% accuracy for English-based content vs 50% for the baseline. Alowibdi et al. [6] used a large dataset of Twitter profiles to identify the features that classified the users based on gender. They used the phoneme sequences of first names and usernames as features. The system utilized a set of KNIME (Konstanz Information Miner, an open-source data analytics, reporting, and integration platform) classifiers. The training process used a "ground truth" of a user's gender by following the links from the profiles to other social sites. They reported an accuracy of around 74.78%. The same group of authors [7, 8] proposed a language-independent scheme to identify the gender. For this, they defined a novel approach for predicting the gender by using color-based features, e.g., background color, text color, link color, ... etc. Training KNIME classifiers using "ground truth" reported an accuracy of around 71%. To identify a Twitter user's gender, Liu and Ruths [27] tried to link between the gender and the user's self-reported first name in English tweets. The authors claimed that using the first name into a gender classifier increased the accuracy over a classifier where name information was omitted (baseline method). They built a large gender-labeled dataset of Twitter users which was used to train their SVM classifier. The system's highest balanced accuracy was 87.1% compared to the baseline's balanced accuracy of 83.3%. The accuracy was averaged over the genders, with a slightly better accuracy for the males.

Next, we look into works dealing with statistical-related features. Argamon et al. [13] considered the problem of authorship profiling trying to determine information such as gender, age, native language, and personality. They proposed using Bayesian Multinomial Regression learning algorithm with two basic features: content-based and style-based features. For gender, the best reported accuracy was 76.1% using both features. Santosh et al. [35] proposed a machine learning approach using a decision tree of classifiers (SVM, MaxEnt, and Decision Tree) to determine an unknown author's age and gender. They considered three different types of features: content-based features (frequencies of different $n$-grams in the documents written by a particular gender), style-based features ($n$-grams of POS tags in the documents, punctuation symbols, ... etc.), and topic-based features ($n$-gram based approach modeling the top words used by both genders). The $n$-grams is a collection of contiguous sequence of $n$ items from a given text, where the items can be letters, words, DNA base pair, ... etc. according to the application. The authors who trained their system on PAN 2013 provided blog corpus and reported an accuracy of 64.08% for age and 56.53% for gender for English contents, and 64.30% and 64.73% respectively for Spanish contents. In [36], the authors created predictive lexica (words and weight) for determining gender and age. They created weighted lexicon by using the 1-gram coefficients from linear multivariate regression and classification models. For gender prediction, the authors trained SVM with a linear kernel using dataset of Facebook messages from the users of MyPersonality application. They reported an accuracy of 91.9% for gender and 83.1% for age. Burger et al. [17] presented a language-independent classifier for predicting the gender of tweets. They used a large volume of data drawn from Twitter users with completed profile. The classifiers were trained using gender-tagged tweets. They applied $n$-gram for characters ($n = 1$–5) and for words ($n = 1$–2), with and without case-folding on the tweets. The authors trained four different classifiers on 4.1 million tweets. The accuracy, as measured on the development set, ranged from 67% for Naïve Bayes to 74% for Balanced Winnow2. In [30], the authors proposed a system to classify the gender of Twitter users using Perceptron and Naïve Bayes classifiers covering the tweets in English only. The authors roughly collected tweets of 3000 users totaling 36,238 tweets. For each user, they manually labeled all but one tweet for the gender. The data was split into two equal sets for training and testing. For features, they employed $n$-grams ($n = 1$–5) of what was observed during training. The balanced accuracy for the Perceptron was 94%, while that of Naïve Bayes classifier was slightly better at 96.8%.

In [32], a corpus of over 1.5 million Flemish Dutch posts from Netlog, a Belgian social networking platform, was studied. With the objective of pedophile detection system, the authors investigated predicting age and gender of short chat messages. For age, they were interested in classifying it as adult or adolescents. The authors presented a feature extraction algorithm using the chi-square feature selection on word $n$-grams ($n = 1$–3) and character $n$-grams ($n = 2$–4). They built the feature sets by selecting 1000, 5000, 10,000, and 50,000 features with the highest chi-squared values from the training data. For the evaluation, SVM was used and achieved an accuracy of 71.1% for the age-based classification. In [38], the authors worked on gender identification of Sina Weibo, one of the most popular social media sites in China. They used three different feature selection algorithms (*tf-idf*, CHI, and COS-CHI) and retrained multi-attribute to improve the accuracy of two classifiers, Naïve Bayes and SVM. CHI uses the

chi-squared test for classification assigning more weight to low frequency words, while COS-CHI is CHI with Cosine similarity. The authors crawled the information of about 200,000 users randomly from hundreds of different professions. Preprocessing rules were applied to cleanse the data. The authors reported achieving an accuracy of 89.01%.

## Dataset Used in this Study

Arabic is a Semitic language that is native to over 300 million speakers, and over 1500 million Muslims worldwide use it in their regular daily prayers. The Arabic language can be classified as classical, modern, and colloquial. The Modern Standard Arabic (MSA) is an evolving language with constant borrowing and innovation to meet the modern challenges [21]. The vernacular or colloquial Arabic, which until recently was confined to daily informal verbal communication, has gained popularity, thanks to the social media. It now trickled down into written form [14]. The Arabic alphabet consists of 28 basic letters including three long vowels. In addition, the Arabic orthographic system uses small diacritical markings to represent short vowels, and there are a total of 13 different diacritics which are placed either above or below the letter. Besides the phonetic information associated with each letter, they serve to clarify the sense and meaning of the word. For example, the plain word (علم) may mean one of (عَلَم: *Ealam*) *flag*, (عِلْم: *Eilom*) *knowledge*, (عُلِمْ: *Eulimo*) *became known*, (عَلَّمَ: *Eal~ama*) *taught*, or (عَلِّمْ: *Eal~imo*) *teach*.[1] In MSA, the diacritical marking is absent from the written text. So, the reader must disambiguate the word sense based on the context in which it appears, a task the natives excel in. However, there are cases in which a large world knowledge is necessary, for a good example see [15, p. 479].

The lack of datasets covering the popular Arabic dialects of Saudi Arabia and the Gulf region compelled us to compile our own. The lack of a vocabulary dictionary of the dialectical Arabic complicates the matter as the individuals will spell the words/sentences phonetically, giving rise to the possibility of multiple spellings of a word within the same dialectical group [5]. There are six major dialectal groups within Saudi Arabia alone, the most widespread being the Hejazi and the Nejdi dialect. We compiled a dataset from a popular local social forum in Saudi Arabia (www.eqla3.com). The bulk of users of this site are from the capital, Riyadh (pop. 5.2+ million), and its vicinity. This is a huge forum with a vast amount of content, with just under a million user base (987,000 registered users) and more than 46 million posts covering slightly over 1.3 million topics. There is a broad set of predefined categories under which users post a topic, initiating a series of other posts in response. Only registered users can post. All the posts are written in informal Arabic, mainly in Saudi Arabian and neighboring Gulf countries colloquies. Each post shows the username, gender, age, educational level, nationality, and the membership date. All the above information is public, and this is why we picked this site as a backdrop for our study. We followed a certain criteria when selecting the material, these are (1) a topic must be generated by a user, i.e., quoted, copied, or auto-generated topics were avoided; (2) the number of characters in each post must not go below our defined threshold of 10 words; (3) the topic should cover different genres that is of interest to both genders; (4) we filter-out all non-Arabic contents; and (5) the context of the topic must be related to gender. The threshold of 10 words was meant to rid of posts having dubious messages such as (الف شكر عالموضوع: *Alf $kr EAlmwDwE*) *thousand thanks on the subject*, (يعطيك العافيه: *yETyk AlEAfyh*) *good job*, etc. Table 1 shows some statistics regarding the compiled dataset.

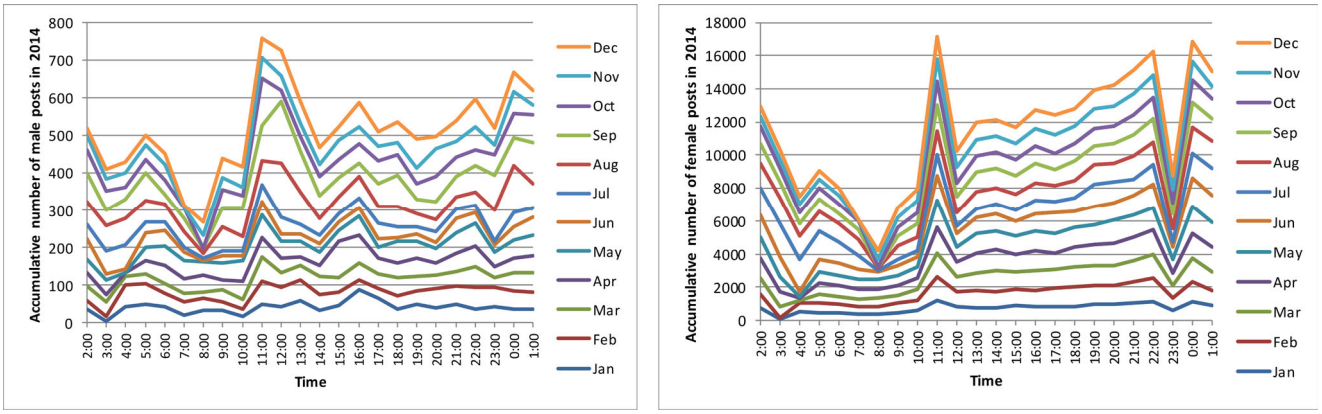
## Analyzing the Posting Behavior

The huge amount of data of the social media at our disposal made it possible to analyze the behavior of users regarding their posting patterns. Our focus is to study the post from two different perspectives: the author's gender and his/her educational level. We are also interested in finding unique features that would be useful in author profiling, predicting the gender. In future, we may extend the profiling to include educational level and age group. In this analysis, we focus on four aspects: (a) gender-wise daily activities of users in social media, (b) analysis of patterns in data posting for each gender, (c) daily activities of users in social media with respect to their educational background, and (d) posting data and the educational background. We have defined a subset of the dataset that contains 100,000 randomly selected topics, containing 433,199 total posts covering the years 2011–2014 inclusive.

**Table 1** Some statistical parameters of our compiled dataset

| | Number of words | | | Number of stems | | |
|---|---|---|---|---|---|---|
| | Min | Max | Avg | Min | Max | Avg |
| Male | 131 | 24,279 | 16,673 | 71 | 6542 | 4739 |
| Female | 77 | 41,101 | 25,554 | 43 | 9980 | 6851 |

[1] We will be following Buckwalter transliteration scheme http://www.qamus.org/transliteration.htm [last accessed Sep 28, 2017], for those having difficulty reading the Arabic script.




(a) Male (b) Female

**Fig. 1** Accumulative monthly posting (for **a** male, and **b** female) on an hourly basis in a 24-h cycle for the full year 2014. Female posting pattern is more consistent compared to the males. The posting activity peaks at 11a.m. for both genders

In our study, we tried to answer the following questions: what are the daily activities of users in social media? What are the posting rate for males and females? What are the peak hours/months for social media? Do holidays impact daily activities? Can educational background affect posting rate? We do have the exact posting time, expressed as hh:mm. However, for reporting the activity, we use an hour resolution. That is, any activity in an hour period is treated as taking place at the top of the hour. In the ensuing discussion, if we claim that a certain activity peaks at 11a.m., it means the activity took place anywhere between 11a.m. and just before noon.

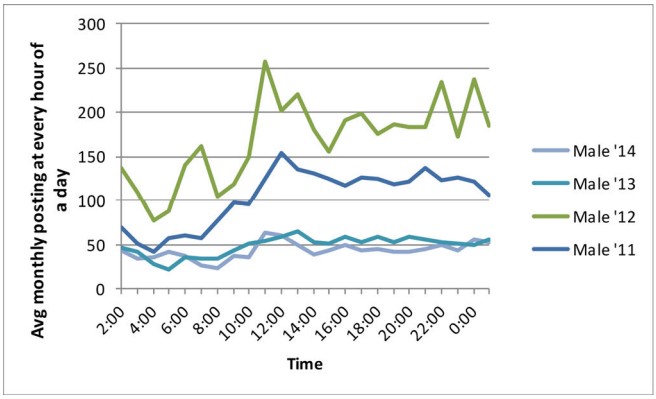
Figures 1, 2, 5, and 7 show the activity graphs of the accumulative number of posts at a specific time in a 24-h cycle. Taking 2014 as an example, we note that females posted around 13,000 messages at 2a.m. in the entire year. Figure 1b shows the break down of the number of posts for each month at every hour during the 24-h cycle. This translates to an average of 1080 messages per month at 2a.m. in 2014 (see Fig. 2b). Looking into females' posting pattern in a 24-h cycle (Fig. 1b), we see nicely stacked lines that are very similar, while for males (Fig. 1a), the stacked lines are not much consistent. It goes on to say that females are more consistent in their posting pattern than their male counterpart, and moreover, they are more prolific and avid posters. Indeed, for the years 2011–2014, females are consistently more actively posting, about ten times more than males (Fig. 2).

In Fig. 1, we see that the posting activity peaks at 11a.m. for both genders, which is during the break time for the Noon prayer (Dhuhr).[2] Since we are plotting activity time using 1-h resolution, so 11a.m. covers all the posting activities taking place between 11a.m. and noon. In Riyadh, for example, the actual call for Dhuhr swings between 11:36a.m. and 12:07p.m., depending on the month of the year. After all, for most of the year, the Dhuhr break is before noon, and that explains the surge of posting activity at 11a.m. in Fig. 1. There is a few minute gap between the call to prayer (Azan) and the physical prayer, and many people make use of this time to catch up with their posting activity.

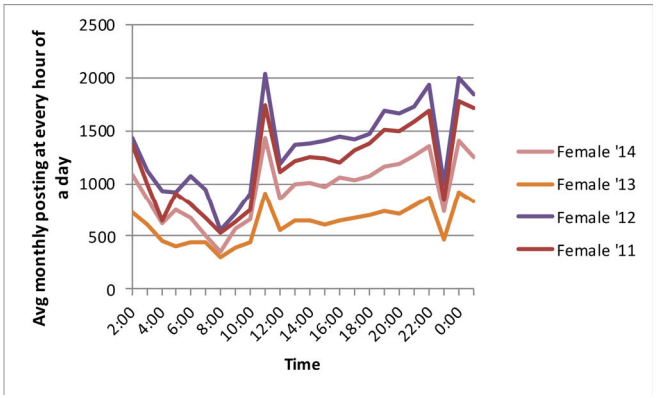
There is also a noticeable surge of activity at 5a.m., an odd time for many. This time is associated with another prayer activity, the Dawn prayer (Fajr). Again taking Riyadh for example, the call for prayer swings between 3:32a.m. (in June) and 5:17a.m. (in January). Those who miss praying Fajr in congregation can still pray it any time until before the sunrise which occurs at 5:03a.m. (in June) and 6:39a.m. (in January). If we look at Fig. 1, we see that there is an increase of activity at 4a.m., and more so at 5a.m., followed by a decline at 6a.m. This can only be explained by people waking up for Fajr, some may post before the prayer, while most postpone till after the prayer and before resuming their sleep.

There are two dips in the posting activity for which we could not find a solid explanation. A no posting activity at 8a.m. for the second half of 2014 (actually starting with the month of August). This is true for both genders. Typically, 8a.m. is the time when people commute to their work place; however, what is not clear is why this pattern is not showing up in the years 2011–2013. Then, there is a sudden decline of posting activity between 11p.m. and midnight for the females, before it peaks out again at midnight. This latter phenomena has been consistently observed for females in all the 4 years of this study (Fig. 2b). We are thankful for the following explanation given by R.K. Akbar (MA Psychology). This is the time the mothers are busy putting their children to bed. Staying up late in Saudi culture is so

[2]Muslims pray five times a day. These are Fajr (dawn); Dhuhr (noon); Asr (afternoon); Maghrib (sunset); and Isha (night). The prayer times are dynamic and are impacted by the Sun movement, and accordingly they vary between seasons.




(a) Male




(b) Female

**Fig. 2** Average monthly posts for **a** male, and **b** female, on an hourly basis in a 24-h cycle, for the years 2011–2014. For example, for the full year 2012, the average number of male posts at 11a.m. was 256.7 per month while it was 2031 for the females. Females have always been more prolific when it comes to posting, and in our case about 10 times more than their male counterpart

common that it has become the norm. A study conducted by Boston College in 2013 found that 73% of 9- and 10-year-old school children in Saudi Arabia were sleep deprived [3]. With kids going to bed around 11p.m. and waking up at 5a.m., so to arrive at school by 6:30a.m.; it leaves them with little time to get a full night's sleep.

From Fig. 3, we see an increase in females posting activity during the months of June, July, and August, where these months correspond to the annual summer break in Saudi Arabia. For females, the highest posting rate was during the month of August for the years 2012–2014. In 2011, August was the third highest posting month, after January and February. In fact, these first two months of 2011 saw the climax of the Arab Spring, and we expect a warm social media interaction during this period.

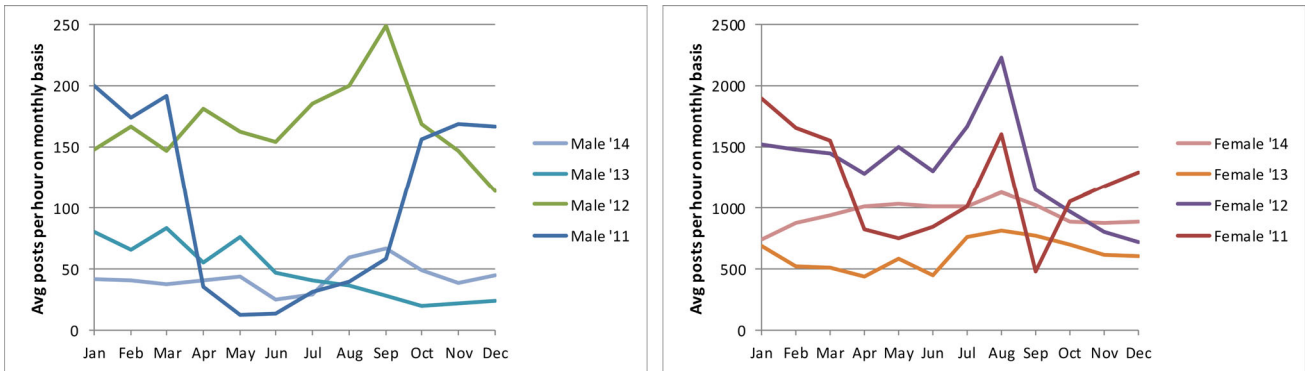
The fasting month of Ramadan[3] fell during the summer break in this 4-year period. Due to the differences between lunar and solar calender, Ramadan shifts by 10 days every year. In 2014, Ramadan coincided with the entire month of July. Falling during July, a summer month where temperatures hover around 120°F, it would be interesting to have a closer look into the posting activity for Ramadan 2014. First, we combined the postings of both genders and calculated the average number of posts per hour and the standard deviation (SD) for each individual month in 2014. The SD ranged between 300 and 370, except for July where it jumps to over 560. That is, the activity during Ramadan is far more dispersed than the other months. Figure 4 plots the difference between the actual number of posts at each hour and the yearly mean for the number of posts per hour. The $x = 0$ coordinate represents the yearly mean, bars above (below) this line means activity is respectively above (below) the mean. The dark solid bar is for July (Ramadan), while the light bar is the average for the remaining 11 months. For instance, at 3a.m., we have more than 1000 posts over the average in Ramadan, whereas it is mildly below average for the remaining months. In Ramadan, the bulk of activity takes place between midnight and dawn, the time fasting starts; and the next stream of activity is after breaking the fast at sunset (around 6:45p.m.) and continues until night (Isha) prayer time 2 h later. This tells that people stay up all night during Ramadan and sleep after dawn, and they remain inactive most of the day until the time to break the fast. The pattern during Ramadan is drastically different than the regular non-fasting months, where the major activity occurs late afternoon till midnight.

In 2011, we note an abnormal and a lasting decline in the posting activity for the 5-month period starting April of the same year, see Fig. 3. A trend is repeated for both genders, though it is a little lengthier for males. Initially, we suspected that this decline was due to the introduction of Twitter in Saudi Arabia, but our investigation confirmed that we were wrong.[4] The first quarter of 2011 witnessed the Arab Spring and this behavior is related to this Spring and the fallout which kept many people at bay. Normal activity resumed after the start of the academic year in September 2011.

In Saudi Arabia, the 12 years of schooling is compulsory. The social media's metadata also contained the educational background of individual users. The users get to declare their highest educational qualification from among seven categories: *Primary* (covering grades 1–6), *Intermediate*

[3]The ninth month in the Islamic lunar calendar. The fasting typically starts at dawn and ends by sunset.

[4]We failed to find the exact date when Twitter was launched in the kingdom; however, using the service https://discover.twitter.com/first-tweet, and going through the first tweet of some of the popular local celebrities, we were able to confirm that Twitter was introduced around Q3/2010.




(a) Male (b) Female

**Fig. 3** Average number of posts per hour for different months for **a** male, and **b** female covering the years 2011–2014. For example, males posted on average 200 messages per hour during January '11. We note increased activity during the 3-month summer break (June–August), especially the females

(grades 7–9), *Secondary* (grades 10–12),[5] *Diploma* (2-year colleges), *Bachelors*, *Post graduates* (masters and PhDs), and *Other*, a choice for those who would rather remain anonymous. For convenience, we collate them into two distinct groups while leaving out those declaring themselves *Other*: high school, which covers users having at most Secondary diploma; and college, those with higher education succeeding the Secondary schooling. We respectively designate these groups as HS- and College+. Figure 5 covers the daily posting activity for both groups for 2014. Interestingly, the posting pattern appears almost identical for both groups except for the volume, where the College+ users are more actively using the social media than the HS- users by a factor of four. We correlated the 1-year average of the number of posts at each hour of the day: 2a.m., 3a.m., …, 1a.m. for the year 2014 (Fig. 5a, b) for each group, and the correlation coefficient is 0.98. Given that, we may safely conclude that the daily posting pattern is independent of the educational level for the two major groups (HS- and College+) in 2014.

We collated three school levels as one group, namely HS-. What about posting activity at individual schooling levels. In 2014, for example, we see that kids in Primary schools were slightly more active in posting than those in Intermediate school. This trend reverses for Secondary schooling level, see Fig. 6. For the 4-year period 2011–2014, Fig. 7 shows the average, over 1-year period, of the number of posts per hour in a 24-h cycle for the two groups: HS-, and College+. We see that the posting pattern for both groups is very much the same. The 4-year posting activity are positively correlated with a correlation coefficient of 0.91. This confirms what we noted earlier, the educational level has little or no impact on the posting pattern. However, College+ users are more actively posting than HS-.

There is a steady gradual decline in the usage of the social media starting third quarter 2012 for both groups, though it reversed action in later part of 2014 for College+ users, see Fig. 7. We attribute this to an increase in alternative new forms of social media. From Fig. 8, we note an increase in the usage of social media during the summer break. Though we did not investigate the educational level posting pattern from the gender perspective, we however can spot an interesting pattern. The posting pattern for 2014 (Fig. 5) closely resembles the posting pattern for the females (Fig. 1b). We can deduce that females, regardless of their educational background, are the ones who are doing most of the posting. It also tells another story. There is a large pool of highly educated females, i.e., those with bachelors and post graduate degrees. If, for example, there were fewer females with higher education, then Fig. 5b would not have so closely resembled Fig. 1b. According to one report, women account for 58% of all Saudi university students, and about 60% of all university graduates [31], and our results tend to corroborate that.

We looked into gender-based posts for the word usage. It was observed that females tend to use more descriptive words (more words to explain their ideas) whereas males use fewer words. The top 3 words as used by females are (من: *mn*) *from*, (في: *fy*) *in*, and (إلى: *AlY*) *to*, with their respected frequencies: 2543, 1439, and 1405; while the top 3 words for males are (من: *mn*) *from*, (في: *fy*) *in*, and (على: *ElY*) *on*, with their respected frequencies: 1247, 1128, and 662. In both cases, the top 3 words are stop-words. If we exclude the stop-words, the highest occurring word as used by the females is (الموضوع: *AlmwDwE*) *the subject*, and for males it is (راح: *rAH*) *went* (vernacular), and their respective occurrences are 236 and 87. For females, there

[5] In US educational system, these will be roughly known as Elementary, Middle, and High schools (respectively).

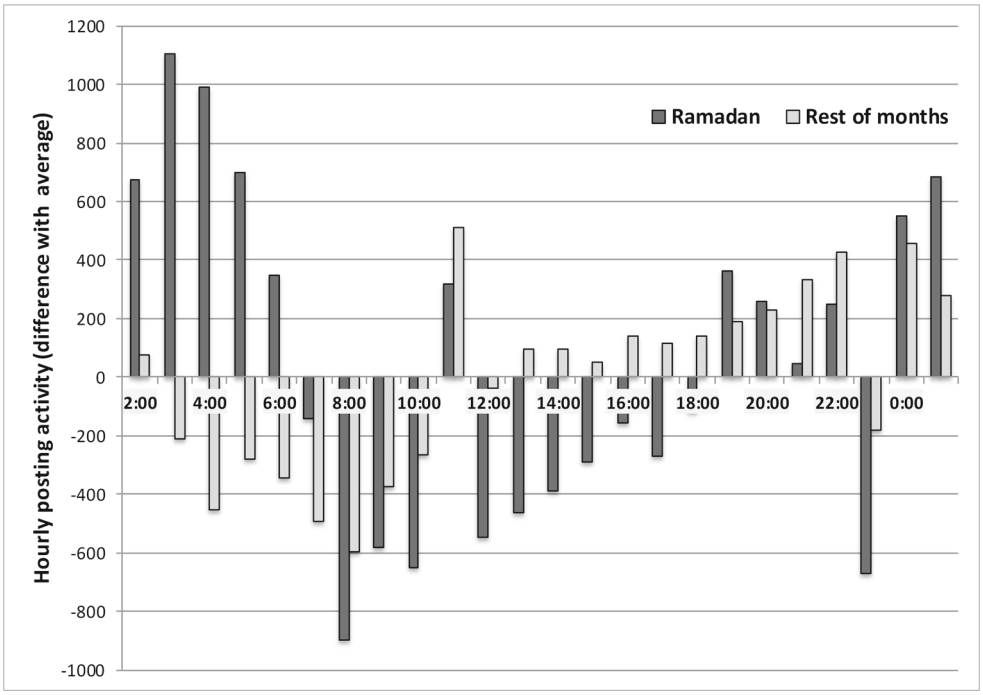
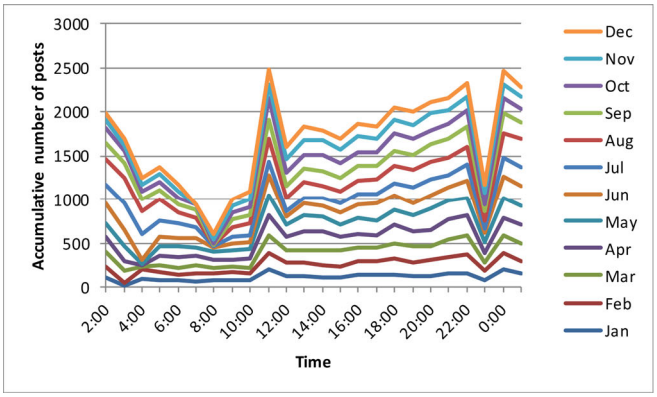



**Fig. 4** Hourly activity (genders combined) in a typical 24-h cycle in the fasting month of Ramadan 2014 (dark solid bar) and the average for the remaining months (light bar). The $y = 0$ coordinate represents the hourly mean for 2014. The height of the bar is the number of posts over/below the mean. In Ramadan, most of the posting activity takes place between midnight and dawn

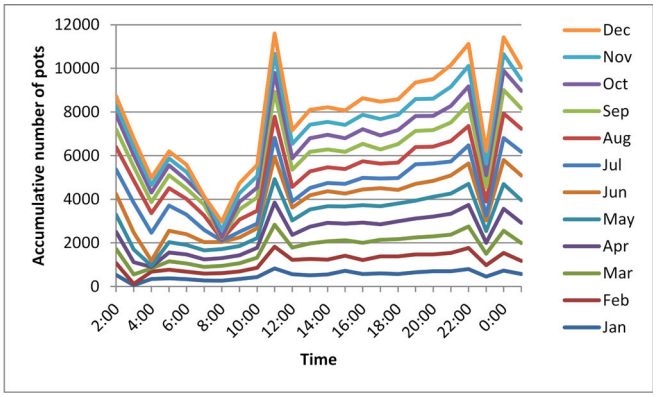
are two words that occur 87 times, and they are ranked 50 and 51. The full list of top 100 words (stop-words are excluded) for each gender is in Table 2. For convenience, we translated the individual words in the list. The list is a combination of MSA and vernacular words. Some MSA words have a different meaning when appearing in a vernacular text. Though it is rare, but when translating the words in Table 2, we opted for the meaning of the word in the local vernacular, since it is more common than MSA in social media. For instance, (واحد: *wAHd*) means *one* in MSA, while in vernacular it means *someone*.

## Author Profile: Gender Identification

Author profiling is a problem of growing importance, e.g., forensics and security. The profile includes among others the identification of the gender. Given a post, can we predict the author's gender? Based on the state-of-art schemes, we will use SVM classifier which follows the linear model. We will also examine the applicability of variants of $K$-NN. The existing behavior-related feature methods have proven to be robust in binary classifications. We follow the lead in [27], and use top-$k$ words. Additionally, we may incorporate new




(a) At most high school diploma (HS-)




(b) College and beyond degree (College+)

**Fig. 5** Comparison of the number of posts based on the highest educational level attained by the user for 2014, grouped into two categories: **a** those with at most high school diploma and **b** those with post high school (college and beyond) degree. From (**a**) and (**b**), we note that the educational level has no impact on the posting pattern, other than the volume

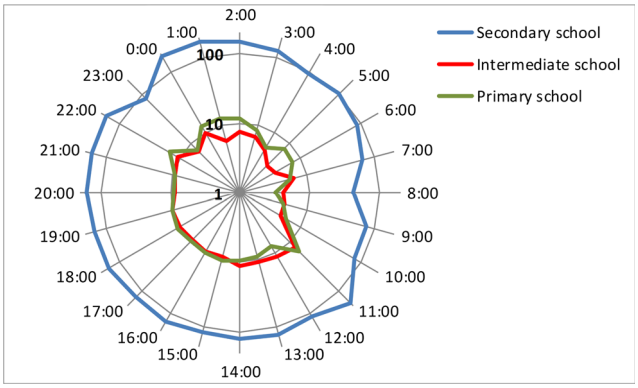



**Fig. 6** Spider chart (log scale) showing the average number of posts, on an hourly basis in a 24-h cycle for the year 2014, for all three levels of school

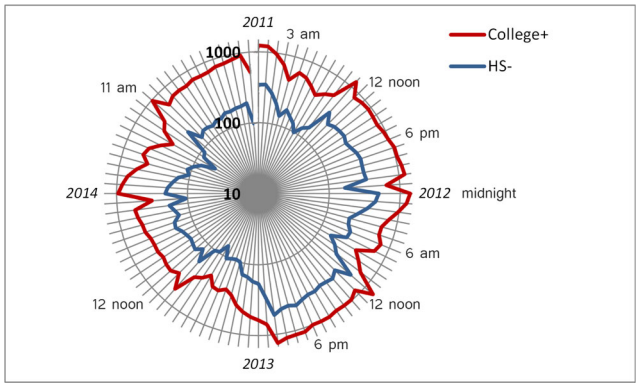



**Fig. 7** Spider chart (log scale) showing the average number of posts based on the educational background, on an hourly basis in a 24-h cycle, for the years 2011–2014. Each radial line represents an hour (total of $24 \times 4 = 96$ radials). The values at 12, 3, 6, and 9 o'clock represents the average over 365 days for the number of posts at midnight in their respected year. We see that both group's chart looks similar, and this means posting activity is mostly independent of the educational level

features to counter some of the peculiarities of the Arabic language. Algorithm 1 lists our proposed system which consists of three phases: preprocessing, feature extraction, and classification.

**Algorithm 1** Proposed system for predicting the gender

**Input**: Raw text

```
begin
    Remove digits and special symbols, e.g. !, @, #, $, %, ... etc.
    Remove all the diacritical marking.
    Normalize the words.
    Tokenize the words of each topic, and generate a stem version of the topic.
    Remove speech effects.
    // feature extraction
    Calculate the score of each word using tf-idf.
    Generate the feature vectors, top-k words list, top-k stems list, and frequent characters.
    Do the classification.
```

The raw data requires some cleaning and normalization before features are extracted. There are non-Arabic words, digits, codes (HTML and others), and some special characters, e.g., emojis. All these will be dealt with in the preprocessing stage, where we (a) remove any character not in the alphabet range of the Arabic language (including digits and the diacritical marks); (b) normalize, where variants of a letter is represented by a single form, e.g., unify the different forms of the character Alif (آ، إ، أ، ا) → (ا); (c) tokenize all the words in a topic, generating a stem version of the topic; and (d) normalize the sequence of repeated characters, such as in "Hellooooo," which we call speech effects. These are common in social media posts and are mainly intended to emphasize. We would like to normalize the variants to a single form, so "Hellooo," "Hellooooo," etc. will all be mapped to "Hello." Untreated speech effect will cause the different variants to be treated as different words and will likely be ignored by *tf-idf* as being a rare word. This normalization was used to preprocess Arabic tweets in [5]. Emojis are another common phenomenon in social media posts. What makes them appealing is that an emoji summarizes a whole set of emotions which otherwise would have taken many words to express. While studying Android devices, [19] noted that female users were more likely to use emojis in text messages than their male counterpart. The authors also noted certain patterns of emoji usage that varied according to the gender. Thus, emojis could be used as one of the factors to predict the gender. However, in the forum we used for this study, no emoticons were used and emojis were scarcely used. The only emojis we encountered were

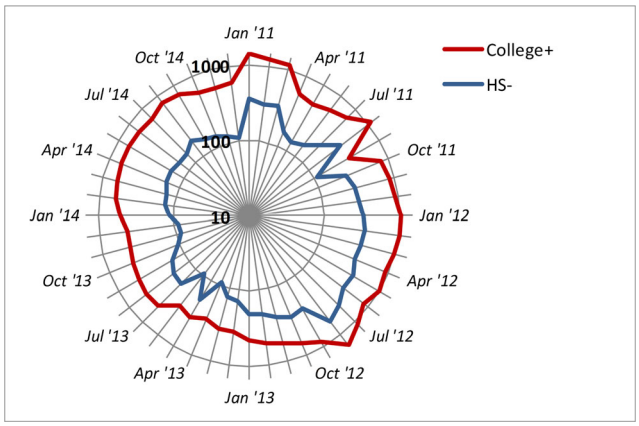



**Fig. 8** Spider chart (log scale) showing the average number of posts per hour for each month based on the educational background covering the years 2011–2014. Each radial line represents 1 month

**Table 2** List of top 100 words (stop-words excluded) along with their English translation as used by each gender in their posting along with the number of occurrences. Entries are ordered top-to-bottom and left-to-right

| Male | | | Female | | | Male | | | Female | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Word | | Occ | Word | | Occ | Word | | Occ | Word | | Occ |
| راح | *went* | 87 | الموضوع | *the subject* | 236 | تم | *done* | 42 | البنت | *the girl* | 87 |
| وش | *what* | 83 | الشعر | *the hair* | 159 | شيء | *thing* | 40 | عندي | *I have* | 85 |
| الموضوع | *the subject* | 82 | الناس | *the people* | 157 | بداية | *beginning* | 40 | ربي | *my Lord* | 83 |
| عام | *public* | 82 | لها | *for her* | 157 | المباراة | *the match* | 40 | فترة | *(time) period* | 81 |
| هي | *she* | 79 | لكم | *for you* (pl.) | 155 | وانا | *and I* | 40 | لنا | *our* | 81 |
| لي | *mine* | 78 | هذي | *this one* | 152 | يمكن | *can* | 39 | اليوم | *today* | 80 |
| لك | *for you* | 74 | جدا | *very much* | 148 | الثاني | *the second* | 39 | الكلام | *talk* | 79 |
| أول | *first* | 73 | كنت | *was* | 148 | خلال | *during* | 38 | أكيد | *surely* | 79 |
| ولكن | *but* | 70 | وحده | *someone* (f.) | 144 | الموسم | *the season* | 38 | عادي | *simple* | 79 |
| احد | *anyone* | 70 | كذا | *like this* | 144 | مبروك | *congrats* | 37 | مبروك | *congrats* | 77 |
| كنت | *I was* | 66 | جميل | *lovely* | 143 | فريق | *team* | 37 | يعطيك | *gives you* | 76 |
| عشان | *because* | 64 | عليكم | *on you* | 142 | لهم | *for them* | 37 | بالنسبة | *regarding* | 76 |
| موضوع | *subject* | 64 | كيف | *how* | 138 | الرسالة | *the message* | 37 | الوقت | *the time* | 75 |
| عليها | *on her* | 64 | عند | *at* | 137 | لكل | *for every* | 36 | إيش | *what* | 75 |
| واحد | *someone* | 63 | كانت | *she was* | 137 | قال | *said* | 36 | خير | *good* | 74 |
| انت | *you* | 62 | لازم | *must* | 128 | جميع | *all* | 36 | كم | *how many* | 74 |
| فقط | *only* | 61 | عليها | *on her* | 128 | برشلونة | *Barcelona* | 36 | أنت | *you* | 73 |
| بشكل | *in a way* | 59 | بشكل | *in a way* | 126 | وهذا | *and this* | 36 | قلت | *I said* | 72 |
| يقول | *says* | 57 | وانا | *and I* | 126 | الأصلية | *the original* | 36 | جميلة | *beautiful* (f.) | 72 |
| أبو | *father of*[a] | 55 | السلام | *peace* | 126 | كتبت | *I wrote* | 35 | شخص | *person* | 72 |
| شاء | *he will* | 53 | الجسم | *the body* | 125 | كيف | *how* | 35 | تقول | *you say* | 71 |
| السلام | *peace* | 52 | موضوع | *subject* | 121 | الجديدة | *the new* | 35 | الشك | *doubt* | 71 |
| يوم | *day* | 51 | فقط | *only* | 121 | السيارة | *the car* | 35 | الحين | *now* | 71 |
| شخص | *person* | 51 | زي | *like* | 119 | العافية | *wellness* | 34 | نفسي | *I wish* | 70 |
| اخر | *another* | 51 | منها | *of* | 118 | أبي | *my father* | 34 | يارب | *O Lord* | 70 |
| مباراة | *match* | 50 | الف | *thousand* | 114 | ليس | *not* | 33 | لهم | *for them* | 70 |
| هذي | *this one* | 49 | شوي | *little* | 110 | حصان | *horse* | 33 | اخرى | *other* | 69 |
| عليكم | *on you* | 49 | بدون | *without* | 109 | رد | *reply* | 33 | ماشاء | *whatever He wants* | 69 |
| وهو | *and he* | 49 | منه | *from him* | 108 | نفس | *the same* | 33 | المكان | *the place* | 69 |
| مرة | *once* | 48 | طبعا | *of course* | 106 | هدف | *goal* | 33 | وهذا | *and this is* | 69 |
| اليوم | *today* | 48 | المرأة | *the woman* | 104 | اخرى | *other* | 32 | بنت | *girl* | 69 |
| الناس | *the people* | 47 | نفس | *the same* | 104 | بدون | *without* | 32 | صار | *he became* | 68 |
| الف | *thousand* | 47 | احد | *anyone* | 101 | بالنسبة | *regarding* | 32 | الطفل | *the child* | 67 |
| انها | *it* | 47 | العافية | *wellness* | 99 | صار | *he became* | 32 | ورحمه | *and His mercy* | 67 |
| الآن | *now* | 45 | أهم | *most important* | 98 | مكان | *place* | 31 | ناس | *people* | 67 |
| أفضل | *best* | 45 | واحد | *someone* | 98 | وهي | *and she* | 31 | أحب | *I love* | 66 |
| لها | *for her* | 45 | البنات | *the girls* | 97 | منه | *from him* | 31 | ملكة | *queen* | 65 |
| لاعب | *player* | 45 | الحمل | *pregnancy* | 95 | ليفه | *scrubber* | 31 | صح | *correct* | 64 |
| السيارات | *the cars* | 45 | الزواج | *marriage* | 95 | الأسرة | *the family* | 31 | الون | *color* | 64 |
| لنا | *our* | 45 | أفضل | *best* | 94 | جديد | *new* | 31 | جديد | *new* | 64 |
| الفريق | *the team* | 45 | طريقة | *way* | 93 | المهم | *the important* | 31 | لون | *color* | 63 |
| تلك | *that* | 45 | كتبت | *I/she wrote* | 92 | العالم | *the world* | 30 | ولكن | *but* | 62 |
| كذا | *like this* | 44 | حلو | *nice* | 92 | تغير | *change* | 30 | ساعة | *hour* | 62 |
| المكان | *the place* | 44 | شاء | *he will* | 91 | المواضيع | *the subjects* | 30 | هم | *they/worry* | 62 |
| لازم | *must* | 44 | وقت | *time* | 91 | عندما | *when* | 30 | أم | *mother of* | 61 |
| موقع | *location* | 44 | الأصلية | *the original* | 90 | جميل | *beautiful* (m.) | 30 | الشخص | *the person* | 61 |
| سيارة | *car* | 42 | أخر | *postpone* | 89 | يعطيك | *gives you* | 30 | شكرآ | *thanks* | 61 |
| زي | *like* | 42 | الرسالة | *the message* | 89 | هناك | *there* | 30 | عندنا | *we have* | 61 |
| جدا | *very* | 42 | بنات | *girls* | 88 | أيام | *days* | 30 | قال | *he said* | 61 |
| الأول | *the first* | 42 | المهم | *the important* | 87 | قاسم | *divides* | 30 | وبركاته | *and His blessing* | 61 |

[a]In the Arab culture, there is a tradition of calling people politely by an epithet particularly when they have children. Usually, the first son's first name is added to Abu (father of). The same goes for Umm (mother of)

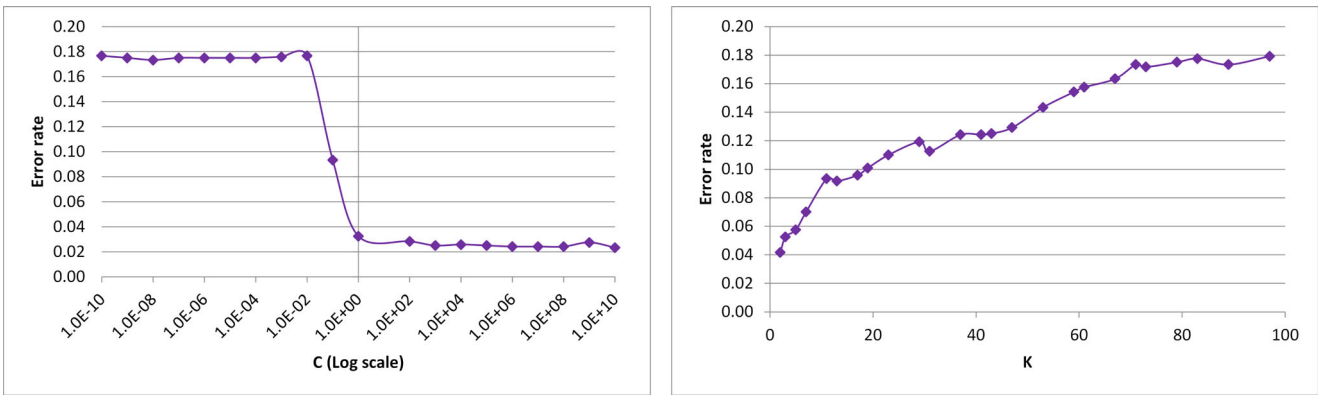
"heart," "flower," "smily face," and "angry face." These were equally used by both genders, making them unusable for classifying the post based on gender.

Top-$k$ is a feature vector, where it holds the top $k$-ranked entries. In the discussion, $k$ will designate the size (number of entries) of the feature vector, top-$k$. In [27], the author used top-$k$ words as their feature vector. We, on the other hand, use two separate lists. The top-$k$ words and the top-$k$ stems.

The top-$k$ words are the $k$ highest scoring list. This list is a subset of a general list of words on each topic. We normalize the words prior to generating top-$k$ words list. The list represents the *tf-idf* scoring for each word in the topic, and it is sorted in descending order. The *tf-idf* (short for term frequency and inverse document frequency) is a common method used to calculate the weight of a word in a document. The main idea of *tf-idf* is, if the term frequency of a word in a document is high and this word rarely appears in other documents, then we believe that the word has a good ability in distinguishing categories. The *tf-idf* is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus [34]. We also define another feature vector, the top-$k$ stems, which is used to remove words that have no root. It is similar in concept to top-$k$ words; the difference is that top-$k$ stems vector is based on the stem version of the dataset. This process is used to adjust the values of the data to capture words with high occurrences used by both genders. For example, we have two forms for *teachers*: (المعلمون) (m.) and (المعلمات) (f.), both are derived from the stem (معلم). Frequent characters is another feature that we use. In this feature, we simply obtain the histogram of the number of characters in each word within the topic, where the histogram is generated after the preprocessing stage. Each pin in the histogram represents a set of the range of characters of words. For classification, we chose SVM. Due to the nature of our problem, where the data follows linear model, we evaluate the feature set using SVM, and for the sake of comparison, we also use 1-NN.

## Performance Evaluation of Our Gender Identifier

For the performance evaluation, we have generated a subset from our dataset which consists of 1200 samples, equally split between genders. For each gender, we evaluate the system using 10-fold cross-validation. For males, we split at 60 samples per fold. Nine folds will be used for training and the tenth for testing. This process is repeated 10 times picking a different fold each time for the testing, while the other nine is used for training. The whole process is repeated for the females. For each test, we use different values of $k$ in top-$k$ (size of the features vector), with $10 \leq k \leq 5000$. We trained both the SVM and the $K$-NN classifiers for these tests. Before training and testing cycle we need a model selection; the cost $C$ for SVM, and $K$ for the $K$-NN. For selecting these two parameters, one sixth of the dataset (randomly selected) was used to evaluate the parameters. After plotting the error rate, the selected parameters were chosen based on two factors: best accuracy and best performance. For SVM, we use the linear model and cost $C = 1$. Figure 9a plots the error rate for different values of cost $C$, where $10^{-10} \leq C \leq 10^{10}$. We see that the error rate stabilizes around 0.02 for $C \geq 1$. The reason for choosing $C = 1$, rather than, say $C = 1000$, is that the difference in error rate is insignificant, moreover, it takes longer to train and validate for larger $C$. For $K$-NN, we evaluated the error rate for values of $K = 2, 3, 5, 7, 11, \ldots, 97$ (all prime integers less than 100), see Fig. 9b. We note that the error rate is the least when $K = 1$. This means that 1-NN yields the best error rate. For the distance function, we evaluated 1-NN using




(a) SVM (b) $K$-NN

**Fig. 9** Error rate for **a** different values of cost $C$ for SVM and **b** different values of $K$ in $K$-NN

**Table 3** Summary of the balanced accuracy for three different distance functions: Euclidean, Mahalanobis, and Manhattan using the 1-NN classifier. Numbers following ± are the standard deviation (SD)

| | $k$ (= number of words in top-$k$ list) | | | | | |
|---|---|---|---|---|---|---|
| | 10 | 50 | 100 | 500 | 1000 | 5000 |
| Euclidean | 95.92±1.69 | 99.25±0.73 | 93.17±1.79 | 49.92±3.48 | 51.33±4.53 | 57.42±3.69 |
| Mahalanobis | 71.25±4.16 | 59.33±5.19 | 60.83±4.95 | 54.42±4.65 | 48.83±6.29 | 54.58±6.08 |
| Manhattan | 95.50±1.81 | 99.17±0.56 | 93.50±1.75 | 50.00±3.64 | 49.92±5.95 | 56.75±2.9 |

three different distance functions: Euclidean, Mahalanobis [28], and Minkowski. The Minkowski distance defines a distance $d$ between two points in a normed vector space, $d(\mathbf{x}, \mathbf{y}) = (\sum_{i<n} |x_i - y_i|^p)^{1/p}$, where $\mathbf{x}$ and $\mathbf{y}$ are points in $n$-dimensional space. When $p = 1$, then Minkowski yields Manhattan distance; and for $p = 2$, it yields Euclidean distance. In our case, we used $p = 1$. Table 3 summarizes the balanced accuracy for 1-NN classifier using the three distance functions. The Euclidean distance function is a little better than the other two distance functions when dimension increases. We note from Table 3, there is no dramatic change when using different distance functions. This is a good indicator that the set of features we devised for this problem are invariant. As to why the balanced accuracy drops for larger $k$, we hypothesize that it is due to the curse of dimensionality, and feature padding. The curse occurs because when the dimension increases, the volume of the space increases exponentially fast that the available data becomes sparse. Padding is a technique to balance feature vectors to be able to use them in classification, and we adopted this mechanism. In feature padding, we pad zeros so all the features vector reach the desired size. For example, if $k = 100$, and the actual size of a features vector is 89, so we pad it with 11 zeros. Larger $k$ means possibly more padding.

To see if dimension reduction techniques would help improve the performance, we repeated the experiment after reducing the dimension of the features using principal component analysis (PCA). Table 4 summarizes the results using 10-fold validation which yields an average accuracy per gender ± standard deviation. Figure 10 plots the balanced accuracy for both genders using SVM and 1-NN classifiers. It is clear that SVM delivered a better performance and a more stable behavior compared to 1-NN classifier, especially for $k \geq 500$. The performance of 1-NN classifier dropped sharply as $k$ increased. Though the accuracy for $k = 10$ and $k = 50$ were excellent, we however prefer a larger $k$. To avoid over-fitting during the classification process, we prefer $k = 100$ over smaller values of $k$. For $k = 100$, both classifiers achieved good accuracy, though 1-NN did better; and for larger $k$, the accuracy of both classifiers suffered. This is due to padding the feature vector with non-related data. By applying PCA on the features, it helped improve the accuracy of 1-NN over the same classifier without PCA when feature size $k$ is over 500. For SVM, there was no significant increase in the accuracy, except that the training process took more longer for the larger set of features (case $k \geq 1000$).

We need to include appropriate significance testing to confirm our assessment that the difference between the

**Table 4** Summary of gender wise and balanced accuracy for both tests using SVM and 1-NN classifiers, with and without principal component analysis (PCA). The values in the table are the average accuracy ± SD computed using 10-fold cross-validation

| | | $k$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | 10 | 50 | 100 | 500 | 1000 | 5000 |
| SVM | Male | 99.66 ± 0.7 | 99.16 ± 1.17 | 87.16 ± 4.97 | 68.66 ± 3.83 | 74.33 ± 6.24 | 70.00 ± 5.03 |
| | Female | 97.83 ± 1.93 | 97.16 ± 2.61 | 87.50 ± 3.86 | 69.83 ± 9 | 72.33 ± 3.06 | 62.83 ± 7.07 |
| | Balanced accuracy | 98.75 ± 1.05 | 98.16 ± 1.51 | 87.33 ± 2.9 | 69.25 ± 4.85 | 73.33 ± 2.81 | 66.41 ± 4.08 |
| 1-NN | Male | 96.17 ± 2.23 | 99.67 ± 0.7 | 92.17 ± 1.58 | 22.67 ± 13.2 | 42.00 ± 7.65 | 92.17 ± 5.16 |
| | Female | 95.67 ± 2.38 | 98.83 ± 1.12 | 94.17 ± 3.36 | 77.17 ± 10.09 | 60.67 ± 7.38 | 22.67 ± 8.96 |
| | Balanced accuracy | 95.92 ± 1.69 | 99.25 ± 0.73 | 93.17 ± 1.79 | 49.92 ± 3.48 | 51.33 ± 4.53 | 57.42 ± 3.69 |
| SVM PCA | Male | 99.17 ± 1.18 | 99.5 ± 0.81 | 58.17 ± 7.55 | 60.67 ± 8.83 | 63.67 ± 5.92 | 67.67 ± 6.77 |
| | Female | 96.50 ± 2 | 98.33 ± 1.57 | 80.17 ± 11.12 | 90.50 ± 6.34 | 83.00 ± 4.43 | 51.67 ± 7.16 |
| | Balanced accuracy | 97.83 ± 1.19 | 98.92 ± 0.88 | 69.17 ± 6.41 | 75.58 ± 4.25 | 73.33 ± 3.26 | 59.67 ± 3.5 |
| 1-NN PCA | Male | 96.17 ± 3.05 | 99.33 ± 1.41 | 88.33 ± 2.48 | 84.00 ± 6.05 | 82.83 ± 4.58 | 77.17 ± 6.09 |
| | Female | 95.00 ± 1.76 | 97.83 ± 1.93 | 88.50 ± 5.12 | 77.33 ± 5.28 | 75.83 ± 6.77 | 65.17 ± 4.61 |
| | Balanced accuracy | 95.58 ± 2.29 | 98.58 ± 1.47 | 88.42 ± 2.62 | 80.67 ± 3.14 | 79.33 ± 4.68 | 71.17 ± 4.5 |

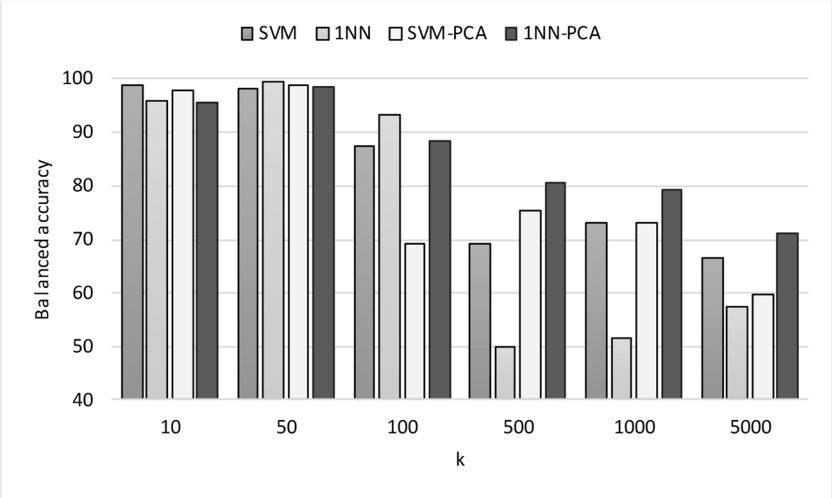



**Fig. 10** Comparison of balanced accuracy using two different classifiers, SVM and 1-NN (with and without PCA). Parameter $k$ is number of words in top-$k$ list

performance of both classifiers (SVM and 1-NN) is not due to some random error. Student's $t$ test is one type of inferential statistics. It assesses whether the means of two groups are statistically different from each other. In other words, could those differences have happened by chance. The results of the $t$ test are reported in terms of $p$ value or calculated probability. A low $p$ value, say $p = 0.01$, means that there is only a 1% probability that the results from an experiment happened by chance. In the literature, the significance level ($\alpha$) is usually set at 5%, and the value $p < \alpha$ symbolizes that it is not likely to be a result of chance. Most authors refer to statistically significant as $p < 0.05$ and statistically highly significant as $p < 0.001$.

**Table 5** Summary of Student's $t$ test between our two proposed methods: SVM and 1-NN classifiers, with and without PCA for different values of $k$. The $p$ values are expressed in scientific E-notation

| $k$ | $t$ test | SVM vs 1-NN | |
|---|---|---|---|
| | | No PCA | PCA |
| 10 | $p$ value | 0.3653 | 0.4482 |
| | Result significant | No | No |
| 50 | $p$ value | 0.1346 | 0.1149 |
| | Result significant | No | No |
| 100 | $p$ value | 2.93E−05 | 5.85E−04 |
| | Result significant | Yes | Yes |
| | Outperforming classifier | 1-NN | 1-NN PCA |
| 500 | $p$ value | 4.65E−05 | 2.48E−05 |
| | Result significant | Yes | Yes |
| | Outperforming classifier | SVM | SVM PCA |
| 1000 | $p$ value | 3.29E−05 | 4.07E−07 |
| | Result significant | Yes | Yes |
| | Outperforming classifier | SVM | SVM PCA |
| 5000 | $p$ value | 2.39E−03 | 9.51E−05 |
| | Result significant | Yes | Yes |
| | Outperforming classifier | SVM | SVM PCA |

Table 5 shows the result of using a $t$ test on different values of $k$ (the size of features vector) in Table 4. For small $k$ ($\leq$ 50), the $p$ value is big, meaning there is no clear winner. As we said earlier, we prefer to avoid such small $k$ because of the possibility of over-fitting during the classification. For the cases $k \geq 100$, the results are more clear. We note that all $p$ values are low (see Table 5). Both classifiers achieved their best result when $k = 100$. The 1-NN classifier delivered a better performance, achieving a balanced accuracy of 93.16%, compared to 87.33% for SVM in inferring the author's gender. The low $p$ value confirms that the obtained performance for 1-NN is indeed statistically significant when compared to SVM, with and without PCA (this was done on cross-validation results on both classifiers). For a larger set of features, SVM delivered a better performance and more stable behavior than 1-NN, but not as good as its best performance.

It is difficult to compare our results with others as there is no work on gender identification in Arabic social media. However, our scheme is better than the one in [7] where the authors proposed a language-independent scheme with an accuracy of 71%.

## Conclusion

Social media is part of everyday life of a sizable Arab population. There were two objectives of this study. First, an in-depth individuals behavior regarding their interaction with the social media and second, profiling the author of a post. There are many statistics about the popularity of individual social applications, but no in-depth study of demographic posting pattern. Our first part of the study was meant to cover the missing picture. Using a popular social

media forum in Saudi Arabia, we mined for the gender-wise activity with the site through the years 2011–2014 on an hourly basis and repeated the same experiment with the backdrop of user's educational background, a first of its kind for Arabic social media. The wealth of insight gained allowed us to do author profiling (gender only) of a given post. For our features representation, we used two lists both of size $k$, calling it top-$k$. A list of $k$ topmost occurring words and the other list of $k$ topmost occurring stems. Both lists are ordered by the entries *tf-idf* value. To test our scheme, we trained two different classifiers, linear SVM and 1-NN. Using different sizes for the list, we noticed that the 1-NN classifier performance degraded significantly for $k \geq 500$, while the SVM classifier showed a more stable performance. We marked the best performance when $k = 100$, with 1-NN yielding better results than SVM, achieving a balanced accuracy of 93.16% vs 87.33% for the SVM in inferring the gender.

For future work, we intend to improve the author profile to include, besides the gender, the age group, educational background, and nationality information.

**Acknowledgments** The corresponding author would like to thank Miss Maryam for her assistance in proof-reading an earlier draft of the manuscript. This work was funded by a special fund of the Research Center of the College of Computer & Information Sciences (CCIS) at King Saud University for which the authors are thankful.

**Funding** This study was funded by a special fund in the research center of CCIS at King Saud University.

**Compliance with Ethical Standards**

**Conflict of Interest** The authors declare that they have no conflict of interest.

**Ethical approval** This article does not contain any studies with human participants or animals performed by the authors.

**Informed Consent** Informed consent was obtained from all individual participants included in the study.

## References

1. AbdulMageed M, Diab M, Kubler S. SAMAR: a system for subjectivity and sentiment analysis for Arabic social media. Comput Speech Lang. 2014;28(1):20–37.
2. Ahlqvist T, Back A, Halonen M, Heinonen S. Social media roadmaps: exploring the futures triggered by social media. Tech. Rep. VTT Tiedotteita Research Notes 2454. Espoo. 2008.
3. Al-Sibai A. The power of sleep. Arab News. http://saudigazette.com.sa/article/165716/The-power-of-sleep (2016). Accessed 3 Aug 2018.
4. Alabdullatif A, Shahzad B, Alwagait E. Classification of Arabic Twitter users: a study based on user behaviour and interests. Mob Inf Syst. 2016:Article 8315,281.
5. Aldayel HK, Azmi AM. Arabic tweets sentiment analysis a hybrid scheme. J Inf Sci. 2016;42(6):782–97.
6. Alowibdi JS, Buy UA, Yu P. Empirical evaluation of profile characteristics for gender classification on Twitter. In: 12th International Conference on Machine Learning and Applications (ICMLA). IEEE; 2013. p. 365–369.
7. Alowibdi JS, Buy UA, Yu P. Language independent gender classification on Twitter. In: IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM '13); 2013. p. 739–743.
8. Alowibdi JS, Buy UA, Yu PS. Say it with colors: language-independent gender classification on Twitter. Springer International Publishing. 2014. p. 47–62.
9. Alawsat A. "Aramco" conducts a workshop on media and social media sites (in Arabic). http://aawsat.com/node/493286 (2015). Accessed 17 Nov 2015.
10. Alwagait E, Shahzad B. Maximization of tweet's viewership with respect to time. In: World Symposium on Computer Applications & Research (WSCAR '14); 2014. p. 1–5.
11. Alwagait E, Shahzad B. When are tweets better valued? An empirical study. J Univ Comput Sci. 2014;20(10):1511–21.
12. Arabian Gazette. REVEALED: Saudi Arabia Social Media Statistics 2012. http://www.arabiangazette.com/revealed-saudi-arabia-social-media-statistics-2012/ (2013). Accessed 28 Nov 2015.
13. Argamon S, Koppel M, Pennebaker JW, Schler J. Automatically profiling the author of an anonymous text. Commun ACM. 2009;52(2):119–23.
14. Azmi AM, Aljafari EA. Modern information retrieval in Arabic—catering to standard and colloquial Arabic users. J Inf Sci. 2015;41(4):506–17.
15. Azmi AM, Almajed RS. A survey of automatic Arabic diacritization techniques. Nat Lang Eng (NLE). 2015;21(3):477–95.
16. BBC. Facebook has a billion users in a single day, says Mark Zuckerberg. http://www.bbc.com/news/world-us-canada-34082393 (2015). Accessed: 28 Aug 2015.
17. Burger JD, Henderson J, Kim G, Zarrella G. Discriminating gender on Twitter. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics; 2011. p. 1301–1309.
18. Cai F, Chen H. A probabilistic model for information retrieval by mining user behaviors. Cogn Comput. 2016;8(3):494–504.
19. Chen Z, Lu X, Shen S, Ai W, Liu X, Mei Q. Through a gender lens: an empirical study of emoji usage over large-scale android users. Computing Research Repository (CoRR) arXiv:1705.05546. 2017.
20. Dubai School of Government. Arab Social Media Report. http://www.mbrsg.ae/getattachment/46b34163-6221-4ee6-8813-c91334ccbbc6/2 (2011). Accessed: 28 Sep 2015.
21. Farghaly A, Shaalan K. Arabic natural language processing: challenges and solutions. ACM Trans Asian Lang Inf Process (TALIP). 2009;8(4):14:1–14:22.
22. GlobalWebindex. Stream social: quarterly social platforms update Q1 2013. https://www.emarketer.com/Article/Which-Social-Networks-Growing-Fastest-Worldwide/1009884 (2013).
23. Jiffry F. #Saudi Arabia world's 2nd most Twitter-happy nation. Arab News. http://www.arabnews.com/news/452204 (2013). Accessed 24 Oct 2015.
24. Jue AL, Marr JA, Kassotakis ME. Social media at work: how networking tools propel organizational performance. San Francisco: Jossey-Bass; 2010.
25. Kumar D. 7 Social media statistics that will get you thinking on your social media marketing strategy. http://dkspeaks.com/7-social-media-statistics/ (2015). Accessed 29 Nov 2015.
26. Larose DT. Discovering knowledge in data: an introduction to data mining. New York: Wiley; 2014.
27. Liu W, Ruths D. What's in a name? Using first names as features for gender inference in twitter. In: AAAI Spring symposium: analyzing microtext; 2013.

28. Mahalanobis P. On the generalized distance in statistics. Proc Natl Inst Sci (Calcutta). 1936;2(1):49–55.
29. Marquardt J, Farnadi G, Vasudevan G, Moens MF, Davalos S, Teredesai A, De Cock M. Age and gender identification in social media. In: Proceedings of CLEF 2014 Conference and Labs of the Evaluation Forum; 2014.
30. Miller Z, Dickinson B, Hu W. Gender prediction on Twitter using stream algorithms with N-gram character features. Int J Internet Sci (IJIS). 2012;2(24):143–8.
31. Pasquesoone V. Higher education: the path to progress for Saudi women. http://www.worldpolicy.org/blog/2011/10/18/higher-education-path-progress-saudi-women (2011). Accessed 18 Jul 2016.
32. Peersman C, Daelemans W, Van Vaerenbergh L. Predicting age and gender in online social networks. In: Proceedings 3rd International Workshop on Search and Mining User-generated Contents (SMUC '11). New York: ACM; 2011. p. 37–44.
33. Pennacchiotti M, Popescu AM. A machine learning approach to Twitter user classification. In: Proceedings of the Fifth International Conference on Weblogs and Social Media (ICWSM '11). Barcelona. Spain; 2011. p. 281–288.
34. Rajaraman A, Ullman JD. Mining of massive datasets. Cambridge: Cambridge University Press; 2012.
35. Santosh K, Bansal R, Shekhar M, Varma V. Author profiling: predicting age and gender from blogs. In: Notebook papers for PAN at CLEF (Conference and Labs of the Evaluation Forum). http://ceur-ws.org/Vol-1179/CLEF2013wn-PAN-SantoshEt2013.pdf. Accessed 9 Jun 2016; 2013.
36. Sap M, Park G, Eichstaedt JC, Kern M, Stillwell D, Kosinski M, Ungar LH, Schwartz HA. Developing age and gender predictive lexica over social media. In: Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha; 2014. p. 1146–1151.
37. Standage T. Writing on the wall: social media—the first 2,000 years. Bloomsbury. 2013.
38. Sun X, Ding X, Liu T. Gender identification on social media. Berlin: Springer; 2014, pp. 99–107.
39. Vinciarelli A, Esposito A, André E, Bonin F, Chetouani M, Cohn JF, Cristani M, Fuhrmann F, Gilmartin E, Hammal Z, Heylen D, Kaiser R, Koutsombogera M, Potamianos A, Renals S, Riccardi G, Salah AA. Open challenges in modeling, analysis and synthesis of human behavior in human–human and human–machine interactions. Cogn Comput. 2015;7(4):397–413.
40. Wilson Center Staff. Saudi Arabia's youth and the Kingdom's future. http://www.newsecuritybeat.org/2012/02/saudi-arabias-youth-and-the-kingdoms-future/ (2012). Accessed 5 Aug 2016.