IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 6, 2019, accepted July 26, 2019, date of publication July 30, 2019, date of current version August 26, 2019.

*Digital Object Identifier 10.1109/ACCESS.2019.2932026*

# Toward Identifying Features for Automatic Gender Detection: A Corpus Creation and Analysis

**SAAD AWADH ALANAZI**
Department of Computer Science, College of Computer and Information Sciences, Jouf University, Sakakah 72441, Saudi Arabia
e-mail: sanazi@ju.edu.sa

**ABSTRACT** The current paper aims to construct an inventory of stylometric and psychometric features for the automatic identification of the author's gender. These features are derived from an analysis of a manually developed Saudi Dialect Twitter Corpus (SDTwittC), consisting of four million words. Given that the study seeks to provide machine learning algorithms with the accurate set of features in solving the gender identification problem, word-based, character-based, syntactic, and function words are all considered during the selection stage. The word-based features constitute the largest category and they represent the possible gender discriminators from sociological, psychological and lexical perspectives. The results show that Saudi males use different styles that separate them from their female counterparts in terms of politeness (greeting, thanking, apology, congratulation, encouragement, best wishes etc), impoliteness (profanity and sarcasm), uses of intensifiers, hedges, color, emotion, reason, emoji among many others.

**INDEX TERMS** Automatic gender detection, feature extraction, Saudi dialects.

## I. INTRODUCTION

With the rapid growth of social media platforms (e.g. Twitter, Facebook, Instagram etc), the anonymity of authors raises a cyber-security concern worldwide [1], [2]. In cyberspace, users can conceal their personal information such as name, gender, age and location, and they become undetectable to the security forces. Under such anonymity, the perpetrators may misuse their accounts and commit online crimes. According to the FBI's Internet Crime Complaint Center (IC3), unlawful acts over the internet such as fraud, non-payment, non-delivery scams, personal data breaches and exhortation have been responsible for a $2.7 billion loss in 2018.[1] In attempt to minimize cyber-threats, local law enforcement agencies implement procedures and fund research that help in tracking the identity of users involved in terrorism, sexual exploitation, child trafficking and other violence acts [3].

Previous studies have executed author profile tasks either to identify threats [4] or to determine their authors' demographics [5]. To profile an author, many studies seek to identify a number of their characteristics such as identity [6]–[9]; native language [10]; dialect [11]; gender [3], [12]–[16]; and political affiliation [17].

Gender identification (GI) is among the natural language processing (NLP) problems aiming to determine the authors' gender of a given text. This problem has gained prominence in a wide range of applications such as e-commerce, marketing, security, forensics etc. In 2013, for instance, the Halt Abuse Organization reported that 30% of the harassers' gender is unknown (while 40% are males and 30% are females).[2] In light of textual data analysis, identifying the authors' gender becomes a promising solution. Male and female authors employ special writing styles and use different stylometric features. These features may be (i) word-based, (ii) character-based, (iii) structure-based, (iv) syntactic or (v) function words. Given that the performance of machine learning tasks heavily relies on the careful selection of features, the study aims to identify the most appropriate features in solving the GI problem.

The rest of the paper is organized as follows. Section (II) presents an overview of the literature done on the GI problem in Arabic and other languages. The methodology and the dataset under study will be introduced in section (III).

The associate editor coordinating the review of this article and approving it for publication was Yonghong Tian.

[1] https://www.ic3.gov/media/annualreport/2018_IC3Report.pdf

[2] http://www.haltabuse.org/resources/stats/2013Statistics.pdf

 This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see http://creativecommons.org/licenses/by/4.0/ 

In section (IV), we will discuss the set of features that separates Saudi users on a gender basis. Concluding remarks are laid out in section (V).

## II. RELATED WORKS

Psycholinguistic studies have shown that an author tends to use a unique style and a set of words that disclose his mental and physical health [18], [19]. It has been argued that males and females employ a myriad of distinct styles. Given that sex is biology-based whereas gender is socially constructed [20], each gender uses special markers to fulfill their social roles. Lakoff [21], for instance, has introduced the field of language and gender, providing lexical, morpho-syntactic and pragmatic features for each class. Similarly, Talbot [22] has explored further sociolinguistic properties that draw a dividing line between both genders, particularly in the workplace. This field has thrived ever since and many characteristics have been proposed on gender bases [23]–[25].

Although the identification of an author's gender has been the focus of many studies worldwide [3], [12], [26] inter alia, it has not received an adequate attention in the Arabic-related literature. Little work has been carried out on Arabic and gender [13], [16], [27]. Alsmearat *et al.* [27] analyze a dataset that consists of 500 articles written in Modern Standard Arabic (MSA) by male and female Arab authors (15 for each). The articles are manually collected from khaberni.com and sawaleif.com. For feature extraction, Alsmearat *et al.* [27] follow two approaches: (i) Bag-of-Words (BOW) and (ii) Sentiment and Emotion-based feature approach. The study does not find conclusive evidence for the common stereotype that female authors compose emotional texts than their male counterparts.

In the same vein, Alsmearat *et al.* [13] continue to address the problem of GI, drawing on another manually collected Arabic-news corpus derived from alrai.com, addustour.com and sawaleif.com. The corpus consists of 2177 articles written in MSA: 2120 by males and 1057 by females. Comparing two approaches (i) BOW and (ii) stylometric features (SF), they found that SF approach scores a higher level of accuracy, i.e. 80.4%, in comparison to BOW approach, i.e. 73.9%.

While the two earlier studies investigated articles from the internet newswires, Hussein *et al.* [16] analyzed textual data from social media applications. They examined an Egyptian Dialect Gender Annotated Dataset (EDGA) manually retrieved from Twitter and they propose a text classification solution to the GI problem. Their corpus consists of 70.000 tweets belonging to 140 active accounts located in Egypt. The GI in their study achieves 87.6% in accuracy and their proposed classification model was accurate by 77.4% in PAN-AP' 2017 dataset.

As far as we know, no previous work has explored the GI problem in datasets comprising Saudi Arabic-written texts. Similarly, no works have consulted sociolinguistic or psycholinguistic studies in feature selection stage. Both fields motivate the current study given that they constitute rich avenues of features that yield accurate results in author identification tasks. The current paper aims to fill this gap and contribute with a manually developed corpus of Saudi dialects from all over the kingdom. It also seeks to investigate whether sociolinguistic and psycholinguistic findings are corroborated by our in-house dataset, and whether new features can be constructed as part of the solution to the GI problem.

Although some studies have assigned a special attention to face attributes, making a great progress in the demographic estimation studies on gender, age, race etc [28], [29], the current study overlooks face images as a gender-discriminating feature due to cultural limitations. While it is more helpful to augment textual corpus with author profile pictures, Saudi females hold to an Islamic belief that woman should not uncover their faces, let alone posting it in a public platform. To maintain as much image-based information as possible, we preserve emojis during the data processing (see section III). We assume that emojis can transcend this limitation as Saudi females represent themselves and their feelings using female figures in online contexts.

## III. METHODOLOGY AND DATASET

In the Arab world, NLP-related problems, including GI, are in their early stage due to the lack of natural Arabic corpora. A few datasets for GI are publicly available but they suffer from some drawbacks. Articles written in MSA by both genders, as is the case in [27] and [13], might have been influenced by the formal lexicon of large news agencies such as Reuters, BBC, CNN and their Arabic-translated services. The lexicon of a political news story, for instance, is standardly constructed by unknown genders. Due to the non-idiosyncratic styles in these domains, the gender authentication of news writers might be less achievable.

Likewise, the creation of Twitter-based datasets, as is the case in Author Profiling Tasks at PAN [14], [15], and [16], draws on tweets as a textual source. Tweets, however, do not necessarily represent the natural language of a given user. They may contain unnatural, literary or copy-and-paste quotes composed by unidentifiable gender such as news, songs, poems, prayers, scriptures' verses among many others. Thus, they may lead to confusion in data analysis and/or author identification.

One of the major contributions of the current paper is to create a large-enough Saudi Arabic corpus with open access to all interested scholars. Thus, it consists of raw data that can be used for training and testing tasks for future studies. The developed corpus is manually gleaned from Twitter, a microblog service which allows users to share 280-character tweets. According to the social clinic (2013), the most users of Twitter in the world are Saudis posting 500 million tweets per month.[3] We name the corpus Saudi Dialect Twitter Corpus (SDTwittC) as it is an inclusive collection of Arabic texts written by users of different dialects spoken in Saudi Arabia.

[3]https://www.thesocialclinic.com/saudi-arabia-ranks-first-on-twitter-worldwide/

**TABLE 1.** The statistical information of SDTwittC.

| | Males | Females |
|---|---|---|
| # of Authors | 100 | 100 |
| # of Replies | 233926 | 219740 |
| # of Characters | 8458174 | 7957387 |
| # of Spaces | 2208300 | 2436565 |
| # of Words | 1624306 | 1732183 |

For the purpose of the study, speakers of other Arabic dialects in the Gulf, the Levant or North Africa are disregarded during the data collection.

SDTwittC consists of 200 authors evenly balanced by gender (100 for each). We identified the gender of the tweeters via their names and profile pictures. As potential copy-and-paste texts, both tweets and retweets are discarded in the first place. Only replies are compiled. The number of replies for each author varies from hundreds to thousands. Male authors produced 233926 replies whereas 219740 replies are generated by the female group. To the best of our knowledge, replies are the most reliable source of natural language as they represent both genders more accurately than composing original public tweets or sharing others' texts. Unoriginal data, if found in replies at all, is drastically minimized in comparison to (re)tweets. Moreover, replying to friends does not need the same amount of time and effort as composing for the public. Although composing in online contexts requires a carefully written prose as the audience varies, replies typically target well-known audience and they are produced for natural social interactions. In this corpus, we have compensated for the low number of subjects (i.e. 200) by eliciting an enormous amount of replies per each author, leading to a four-million-word dataset. Thus, we assume that this sufficient corpus can disclose genuine features that discriminate both genders on a linguistic basis.

SDTwittC has also undergone a cleansing process where all Twitter-specific noises are removed. Any gender-neutral text that is not naturally produced by the users under study is eliminated such as hashtags, URLs, and username mentions. Language other than Arabic, images and diacritics are also banished from the corpus for the accuracy of word frequency. To preserve as many gender identifiers as possible, the text is not further normalized. Punctuation marks, alternating letters and emojis are maintained. The final version of data is put in a plain text file and is annotated based on the gender of the authors. Consider the statistics of SDTwittC in Table 1.

Given that the data is retrieved from a public domain, the data collection process does not violate the ethical rules set by Institutional Review Board (IRB). According to Twitter policies, users are given the right to share their content or to lock and protect their accounts against any intruders. According to Vitak *et al.* [30], an IRB-based study on social computing research shows that researchers are not required to obtain informed consent to collect data from public spheres. Also, an overview of several hundreds of Twitter-based studies reveals that only a few papers have discussed ethical obligations [31]. In the current paper, we have collected data only from unlocked open access accounts, not to mention that we have removed all informants' identifying names plus other user symbols during the reprocessing phase. Therefore, no violations of IRB requirements are committed.

## IV. FEATURE SELECTION

Given that the accurate selection of features improves the robustness of author profiling tasks and other induction methods [32], it follows that the second major contribution of the current paper is to supply machine learning algorithms/classifiers with a stock of features that may predict the authors' gender. Inspired by earlier studies, this section will discuss all possible gender identifiers and will highlight the other non-discriminating markers. Thus, the gender-related features under study aim to provide scholars with preliminary directions in their future research.

In this section, we classify our gender-related features into four categories: (i) word-based, (ii) character-based, (iii) syntactic and (iv) function words. Word-based features will be presented in section (A) and they constitute the largest category covering a diverse array of intensifiers, hedges, and many other terms relevant to color, emoji, emotion, religion etc. As for character-based and syntactic features, they will be taken up in section (B) and they encompass tabs, spaces, special characters such as %, *, &, etc plus the punctuation marks. Function words include pronouns, demonstratives, wh-interrogatives, negation markers, the definite article /al/, feminine and plural words, and they will be finally discussed in section (C).

The second aim of this section is to examine whether our corpus confirms or negates earlier studies on gender-dependent behaviors. Thus, all the studies cited within these subsections will be from sociolinguistics or psycholinguistics. For the calculation of the feature values, we will count the frequency of tokens focusing on the most common words in each category. Some of these words are extracted from the literature and some others are manually selected based on our linguistic knowledge. Due to space limitations, we will provide representative examples for each class and calculate the total of other related tokens as a separate input. The percentage of the total is also computed per 10.000 words.

### A. WORD-BASED FEATURES

Word-based features can be categorized into four sub-sets: (i) sociology-based, (ii) psychology-based and (iii) lexicon-based. The sociology-based category discusses politeness as a social phenomenon. Thus, it covers both polite and impolite speech acts. Politeness-related acts include greeting, gratitude, apology, congratulation, encouragement, best wishes and laughter. As for impoliteness-related acts, they include derogatory expressions such as taboos, curses, swearwords as well as sarcasm. Both politeness and impoliteness require a dyadic interaction and they are implemented for social communication. Thus, the afore-mentioned speech acts aim

to highlight the sociological distinctions between Saudi males and females.

As for the psychology-based category, it is less likely to be used for communication. Given that it involves tokens used to express one's self, it is more expressive than communicative. It contains expressions related to individual preferences and feelings, personal perception of the surrounding environment and personal stances towards others, be they human or non-human. Thus, it covers terms related to color, emoji, emotion, reason, religion and words that intensify or lessen the speech force. We add this category with a view to separate the two genders on psychological grounds.

Concerning the lexicon-based category, it addresses the linguistic content that distinguishes Saudi males from their female counterparts. It consists of expressions from other Arabic varieties such as MSA, Gulf Arabic Pidgin as well as Arabicized English words, and it also involves topic-specific words in sport, health, education, economy, politics and the like. These linguistic differences can be compounded by other features such as character-based, syntactic and function words.

#### 1) SOCIOLOGY-BASED FEATURES

##### *a: POLITENESS-RELATED FEATURES*

In the past decades, many politeness-related acts have attracted a vast amount of research [25], [33], [34]. These acts include greeting, gratitude, apology, congratulation, encouragement, best wishes and laughter. They are used as ''rapport-sensitive speech acts'' [35, p. 18]. In all these respects, the findings suggest that women are more polite than men [25], [36], [37]. In contrast to these studies, our corpus provides conflicting results as shown in Table 2 below. The number followed by ‰ refers to the percentage of token count per 10.000 words.

As far as we know, no study has investigated greeting or congratulation in terms of frequency and gender correlations. The existing works only focus on the various strategies of greeting and congratulation between the two genders, e.g. [38] and [39] on greeting and [40] and [41] on congratulation. As demonstrated in Table 2, our dataset indicates that males greet 18.28 ‰ more than females do, i.e. 14.57 ‰ . Still, males congratulate the most (i.e. 26.71 ‰ ) in comparison to females who score 09.95 ‰ . These new findings suggest that Saudi males are more polite than their female counterparts in greeting and congratulation in online contexts.

The SDTwittC also shows new findings in contrast to earlier studies on apology and empathy (i.e. encouragement and best wishes). Although Schumann and Ross [42] argued that female apologize more, our dataset suggests an alternative view. Saudi males tend to use almost a double number of apology-related phrases than females, i.e. 1237 words (07.61 ‰) for males while 757 words (04.37‰) for their female counterparts. However, this might still indicate that males commit more errors and apologize in reaction. We will find indications for this hypothesis in the following section on impoliteness (§41.2). As for encouragement and best wishes, previous studies have proposed that females support and encourage more than males [43], [44]. In our dataset, we have drawn striking results. Saudi males are by far inclined to encourage and express their wishes to others than females: 17.86‰ vs. 09.94‰ for encouragement and 17.15‰ vs. 07.75‰ for best wishes.

As for thanking phrases, our corpus advocates the earlier results. Hesabi and Azima [45] and Yusefi [46] argued that males express more gratitude than females. Similarly, our dataset shows a gratitude tendency on the male part over the female one. Saudi males express their gratitude (27.26‰) whereas females assign only 14.81‰ for appreciation in their replies.

Although laughter is hardly accepted as a polite action, the literature shows that it enhances the social relations within a speech community and displays positive intentions towards unknown people [47]. Moreover, while speech interruption is sometimes considered rude, laughter interruption is always a gesture of positive feedback [48]. Thus, we treat laughter as a politeness-related action.

Sociolinguistic studies report that men are joke tellers and more funny than women [49]. Lakoff [21, p. 56] takes these facts to the extreme and argues that women ''have no sense of humor''. However, other studies reveal that women are also dragged to laughter in pursuit of affection and intimacy [50]. Given that we cannot calculate the number of jokes, we resort to the counts of laughs manifested in elongated consonants هههههه /hhhhhh/ and خخخخخ /khkhkhkh/. In calculation, we consider all words prefixed with three letters of these consonants. Also, we compute emoji laughing faces as part of the laughing process. The findings in Table 3 demonstrate that no significant difference is detected between the two genders.

##### *b: IMPOLITENESS-RELATED FEATURES*

As for impoliteness, it has been correlated with men who exercise more power, influence, confrontation and challenge [51]. In this section, we will focus on profanity and sarcasm. As a sociolinguistic phenomenon, profanity has received a special attention in the field of language and gender. It has been found that women shrink from vulgar expressions and use more refined language [21], [52], [53]. Regarding sarcasm, males are reported as more sarcastic [49], [54], [55]. Consider the results in SDTwittC below. For the sake of courtesy, we have written vulgar Arabic words reversely and used asterisks for their translations.

As shown in Table 4, Saudi males utter 13.37‰ of obscene words of the whole data more than their female fellows do (i.e. 08.16‰). As far as sarcasm is concerned, no sharp contrast is recorded, but the difference is still skewed towards males (i.e. 05.75‰) in comparison to females (03.64‰).

To bring these results together, it is obvious that there is no conclusive evidence for the mainstream view that women are more polite than men [25], [36], [37]. New findings in SDTwittC suggest the reverse. In online contexts,

**TABLE 2.** The frequency of politeness-related acts in SDTwittC.

| Greeting | Translation | Males | Females |
|---|---|---|---|
| صباح الخير | Good Morning | 898 | 828 |
| صباح النور | Enlightening morning | 313 | 402 |
| مساء الخير | Good Evening | 326 | 186 |
| السلام عليكم | Peace be upon you | 245 | 115 |
| أخرى: مساء النور، مرحبا، أهلا وسهلا، يا هلا، السلام عليكم ورحمة الله وبركاته إلخ | Others: Enlightening evening, welcome, Hi, peace and blessing be upon you etc | 1188 | 994 |
| المجموع | Total | 2970 | 2525 |
| نسبة ‰ | ‰ Percentage | 18.28‰ | 14.57‰ |
| **Gratitude** | | | |
| شكرا | Thanks | 2686 | 1246 |
| عفوا | You are welcome | 97 | 147 |
| يسلمو | Appreciated | 21 | 23 |
| أخرى: مشكور، تشكرات، ممتن، تسلم، بيض الله وجهك، ما قصرت إلخ | Others: Thanks, I'm grateful, God bless you, you did well etc | 1624 | 1150 |
| المجموع | Total | 4428 | 2566 |
| نسبة ‰ | ‰ Percentage | 27.26‰ | 14.81‰ |
| **Apology** | | | |
| آسف | Sorry | 244 | 66 |
| اعتذر | I apologize | 177 | 163 |
| اعذرني/اعذريني | Forgive me | 22 | 11 |
| اخرى: اعذرنا، أعذرينا، متأسف، معليش، سامحني، سامحيني، سامحنا، سامحينا إلخ | Others: forgive us, sorry, excuse me, forgive me/us etc | 794 | 517 |
| المجموع | Total | 1237 | 757 |
| نسبة ‰ | ‰ Percentage | 07.61‰ | 04.37‰ |
| **Congratulation** | | | |
| مبروك | Congrats | 2918 | 1110 |
| مبارك | Congratulation | 196 | 96 |
| تستاهل/تستاهلين | You deserve it! | 1225 | 518 |
| المجموع | Total | 4339 | 1724 |
| نسبة ‰ | ‰ Percentage | 26.71‰ | 09.95‰ |

**TABLE 2.** *(Continued.)* The frequency of politeness-related acts in SDTwittC.

| | | | |
|---|---|---|---|
| **Best Wishes** | | | |
| الله يوفقك | May God support you | 416 | 321 |
| بالتوفيق | Best wishes | 717 | 209 |
| الله يرزقك | May God bless you | 137 | 102 |
| أخرى: الله ييسر لك، فالك التوفيق/الفوز، الله يسلمك، توصل بالسلامة إلخ | Others: May Gold help you, you deserve success/victory, May God protect you, you reach safe etc | 291 | 194 |
| المجموع | Total | 2786 | 1344 |
| نسبة ‰ | ‰ Percentage | 17.15‰ | 07.75‰ |
| **Encouragement** | | | |
| كفو | Well done | 993 | 469 |
| الله عليك\ي\م | Good for you! | 347 | 315 |
| مبدع\مبدعة | You wonderful! | 246 | 146 |
| صح لسانك | Well said | 152 | 174 |
| أخرى: بطل، يا سلام عليك، منور، ونعم إلخ | Others: you hero!, Good for you, you are shining, great name etc | 1048 | 619 |
| المجموع | Total | 2786 | 1723 |
| نسبة ‰ | ‰ Percentage | 17.86‰ | 09.94‰ |

**TABLE 3.** The frequency of laughs in SDTwittC.

| Laughs | Translation | Males | Females |
|---|---|---|---|
| هههه | Hhh | 12440 | 16622 |
| خخخ | Khkhkh | 57 | 55 |
| 😂 | Laughing face | 42044 | 41973 |
| 🤣 | Tilted laughing face | 776 | 507 |
| المجموع | Total | 55317 | 59157 |
| نسبة ‰ | ‰ Percentage | 340.55‰ | 341.51‰ |

Saudi males score higher than females in almost all the politeness-related respects, i.e. greeting, gratitude, apology, congratulation, encouragement and best wishes. With respect to the two types of humor (laughter and sarcasm), no major difference is acknowledged but the number of sarcastic phrases is somehow tilted towards males in particular. Concerning profanity, it is much rampant in the male texts. Figure 1 illustrates all the sociology-based distinctions in SDTwittC.

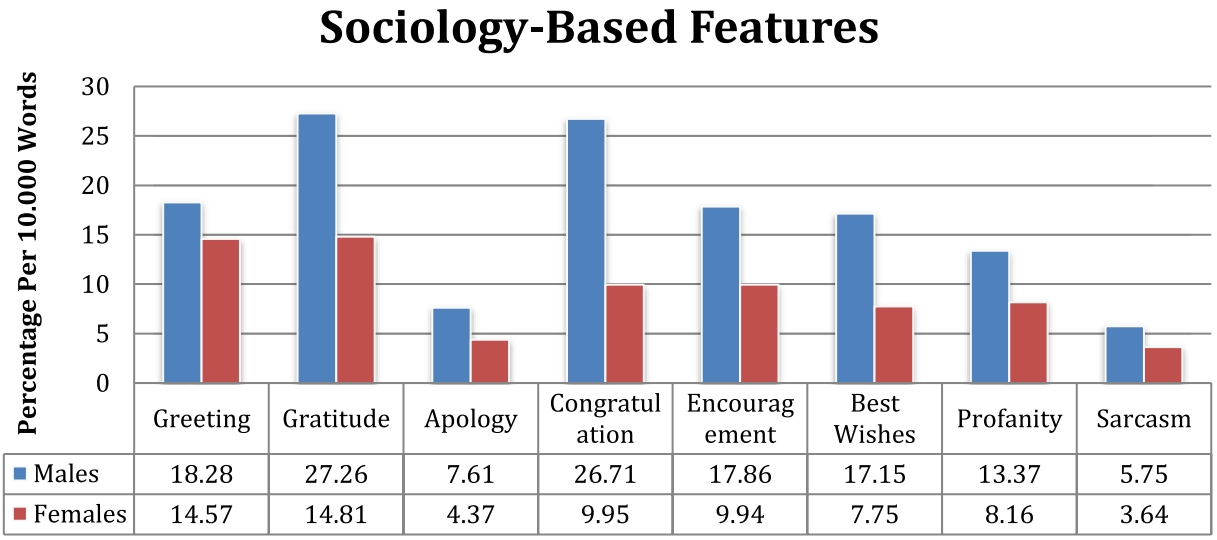



**FIGURE 1.** Summary of Sociology-Based Features in SDTwittC.

**TABLE 4.** The frequency of profanity and sarcasm-related expressions in SDTwittC.

| Profanity | Translation | Males | Females |
|---|---|---|---|
| قز | S**t! | 230 | 210 |
| وفت | Spit! | 216 | 179 |
| الله يلعنك | God curses you | 369 | 233 |
| ارخ | S**t | 172 | 114 |
| أخرى: علقنا، مطنا، فلذا، زط إلخ | Others: f**k off, shut up, get lost, etc | 1185 | 979 |
| المجموع | Total | 2172 | 1415 |
| نسبة ‰ | ‰ Percentage | 13.37‰ | 08.16‰ |
| **Sarcasm** | | | |
| ماش | Low! | 200 | 167 |
| امحق | Bad | 42 | 67 |
| هع | Huh! | 73 | 149 |
| يا ليل | Oh Night | 38 | 17 |
| أخرى: بسم الله عليك، بايخ، يا قدمك، ليتك راقد، نفسية، ها ها ها إلخ | Others: hope you're ok! Trite, you old! Keep sleeping! Psycho, ha ha ha | 581 | 232 |
| المجموع | Total | 934 | 632 |
| نسبة ‰ | ‰ Percentage | 05.75‰ | 03.64‰ |

**TABLE 5.** The frequency of intensifiers and hedges in SDTwittC.

| Intensifiers | Translation | Males | Females |
|---|---|---|---|
| جدا | Very | 1756 | 1365 |
| مرة | Very | 651 | 596 |
| كثير | Much | 1475 | 1420 |
| حلو | Cool | 1086 | 2119 |
| أكيد | Sure | 873 | 848 |
| بالضبط | Exactly | 288 | 402 |
| مستحيل | No way | 498 | 520 |
| أخرى: ولا شيء، بالتأكيد، ممتاز، يجنن، يهبل إلخ | Others: nothing, for sure, super, crazy etc | 498 | 556 |
| المجموع | Total | 7125 | 7826 |
| نسبة ‰ | ‰ Percentage | 43.86‰ | 45.06‰ |
| **Hedges** | | | |
| شوي | Little | 1278 | 1282 |
| مدري | Dunno | 1073 | 1513 |
| احتمال | Maybe | 104 | 60 |
| يمكن | Perhaps | 917 | 956 |
| صح؟ | Right? | 2315 | 1670 |
| ولالا؟ | Is not? | 19 | 12 |
| أخرى: يعني، اقصد، شكلها، عسى إلخ | Others: I mean, it seems, hopefully etc | 587 | 512 |
| المجموع | Total | 6293 | 6005 |
| نسبة ‰ | ‰ Percentage | 38.74‰ | 34.66‰ |

### 2) PSYCHOLOGY-BASED FEATURES

Research in psychology has put forward hypotheses that authors utilize diverse styles and behaviors stressing their mental and physical health [18], [19]. The psychology-based features differ from sociology-based ones in that they are more expressive than communicative. In other words, a speaker may express his or her psychology via different choices and views. Thus, we take psychometric features as a cover term for the uses of intensifiers and hedges, the personal feelings and beliefs and the choices of colors and emojis.

**TABLE 6.** The frequency of repeated consonants and vowels in SDTwittC.

| Repeated Letters | Translation | Males | Females |
|---|---|---|---|
| اااا | Aaaa | 2270 | 3593 |
| ووو | Wwww | 2654 | 3218 |
| يييي | Yyyy | 1582 | 2672 |
| ررر | Rrrr | 654 | 1335 |
| ددد | Dddd | 74 | 191 |
| سسسس | Ssss | 231 | 319 |
| ففف | Ffff | 86 | 169 |
| جججج | Jjjj | 49 | 99 |
| ككك | Kkkk | 117 | 401 |
| للل | Llll | 229 | 388 |
| أخرى: قققق، ببب، تتتت الخ | Others: qqqq, bbbb, tttt etc. | 1357 | 3352 |
| المجموع | Total | 9303 | 15737 |
| نسبة ‰ | ‰ Percentage | 57.27‰ | 90.85‰ |

**TABLE 7.** The frequency of comparative and superlative adjectives in SDTwittC.

| Adjectives | Translation | Males | Females |
|---|---|---|---|
| افضل | Better | 1972 | 628 |
| اجمل | More beautiful | 582 | 781 |
| احسن | Better | 800 | 729 |
| اكبر | Bigger | 887 | 722 |
| اصغر | Smaller | 71 | 47 |
| اهم | More important | 799 | 608 |
| اول | First | 1895 | 1606 |
| اخر | Last | 1145 | 1071 |
| اخس | Worse | 30 | 34 |
| احلى | More beautiful | 537 | 687 |
| أخرى | Others | 1086 | 966 |
| المجموع | Total | 9804 | 7879 |
| نسبة ‰ | ‰ Percentage | 60.35‰ | 45.48‰ |

### *a: USE OF INTENSIFIERS AND HEDGES*

Intensifiers are lexical items that reinforce the force of statements such as *completely*, *definitely*, and *absolutely* in English. They are also known as upgraders [56] or strengtheners [57]. As for hedges, also known as down-graders [56] or weakeners [57], they are mitigating words that reduce the power of utterances, e.g. *somewhat*, *maybe*, and *so and so* in English. Tag questions such as *is not it? right? does he?*etc are also subsumed under hedges [58]. Psychology research has found that women tend to use more intensifiers and hedges than males [59, p. 300]. Women exaggerate and use intensifiers to attract the interlocutors' attention and they also lessen the strength of their speech due to their lack of confidence [21], [59], [60].

Contra earlier studies, our corpus records no difference between the two genders as shown in Table 5: Saudi males produce 43.86‰ as intensifiers and 38.74‰ as hedges in comparison to females who generate 45.06‰ and 34.66‰ respectively. However, exact numbers suggest that

**TABLE 8.** The frequency of romantic and religious phrases in SDTwittC.

| Romantic Phrases | Translation | Males | Females |
|---|---|---|---|
| احبك | I love you | 652 | 1028 |
| قلبي | My heart | 1022 | 2641 |
| حياتي | My life | 476 | 823 |
| حبيبي | My love (M) | 1335 | 463 |
| حبيبتي | My love (F) | 96 | 762 |
| عمري | My days | 207 | 396 |
| بعد عمري | After my days! | 16 | 57 |
| بعد قلبي | After my heart! | 101 | 140 |
| أخرى: اشتقت لك، أموت فيك، عيوني، إلخ | Others: I miss you, I die for you, my eyes, kiss! | 451 | 1073 |
| المجموع | Total | 4356 | 7383 |
| نسبة ‰ | ‰ Percentage | 26.81‰ | 42.62‰ |
| **Religious Phrases** | | | |
| والله | BY Allah | 5944 | 6968 |
| اللهم | O Allah | 3727 | 5767 |
| يارب | O God | 3133 | 5282 |
| ما شاء الله | Majesty to Allah | 439 | 144 |
| الحمد لله | Praise be to Allah | 956 | 645 |
| بإذن الله | God willing | 1189 | 512 |
| ان شاء الله | God willing | 1272 | 906 |
| استغفر الله | I seek forgiveness from Allah | 309 | 374 |
| آمين | Amen | 2337 | 2471 |
| سبحان الله | Glory be to Allah | 555 | 657 |
| أخرى: لا حول ولا قوة إلا بالله، لا إله إلا الله، الله المستعان، حسبي الله إلخ | Others: no power but from Allah, No god but Allah, Allah is the supporter, God is enough for me etc | 2692 | 3100 |
| المجموع | Total | 22553 | 26826 |
| نسبة ‰ | ‰ Percentage | 138.84‰ | 154.86‰ |

males tend to down-grade their utterances whilst females intensify them. As for tag questions, and in support of the findings in [61], Saudi male authors raise more of such questions. As shown in Table 5, Saudi males produced صح؟ *2315*times as opposed to the 1670 times yielded by females.

**TABLE 9.** The frequency of number and time terms in SDTwittC.

| Numbers | Translation | Males | Females |
|---|---|---|---|
| 1-2 | 1-2 | 4142 | 3056 |
| 20-90 | 20-90 | 242 | 142 |
| مائة، مية، مئة | Hundred | 59 | 65 |
| ألف | thousand | 2364 | 1061 |
| مليون | Million | 533 | 280 |
| مليار | Billion | 86 | 33 |
| عشرات | Tens | 23 | 5 |
| مئات | Hundreds | 29 | 15 |
| آلاف | Thousands | 139 | 91 |
| ملايين | Millions | 78 | 40 |
| مليارات | Billions | 11 | 4 |
| المجموع | Total | 7706 | 4792 |
| نسبة ‰ | ‰ Percentage | 47.44‰ | 27.66‰ |
| **Time** | | | |
| اليوم | Today | 2458 | 2053 |
| امس | Yesterday | 594 | 510 |
| بكرة | Tomorrow | 126 | 109 |
| الحين | Now | 1447 | 1439 |
| الساعة | The hour | 183 | 169 |
| أخرى | Others | 4808 | 4280 |
| المجموع | Total | 8365 | 5968 |
| نسبة ‰ | ‰ Percentage | 51.49‰ | 34.45‰ |
| **Days** | | | |
| الاحد | Sunday | 186 | 150 |
| الاثنين | Monday | 206 | 112 |
| الثلاثاء | Tuesday | 105 | 50 |
| الاربعاء | Wednesday | 98 | 55 |
| الخميس | Thursday | 277 | 156 |
| الجمعة | Friday | 213 | 122 |
| المجموع | Total | 1085 | 645 |
| نسبة ‰ | ‰ Percentage | 06.67‰ | 03.72‰ |
| **Months** | | | |
| يناير | January | 26 | 20 |
| فبراير | February | 23 | 24 |
| مارس | March | 82 | 27 |
| ابريل | April | 151 | 42 |
| مايو | May | 26 | 27 |
| أخرى | Others | 146 | 43 |
| المجموع | Total | 454 | 183 |
| نسبة ‰ | ‰ Percentage | 02.79‰ | 01.05‰ |
| المجموع الكلي | Overall Total | 17610 | 11588 |
| نسبة ‰ | ‰ Percentage | 108.41‰ | 66.87‰ |

**TABLE 10.** The frequency of basic and spectral colors in SDTwittC.

| Basic colors | Translation | Males | Females |
|---|---|---|---|
| اسود | Black | 158 | 176 |
| ابيض | White | 160 | 158 |
| احمر | Red | 157 | 118 |
| اخضر | Green | 135 | 86 |
| ازرق | Blue | 105 | 101 |
| اصفر | Yellow | 155 | 66 |
| المجموع | Total | 870 | 705 |
| نسبة ‰ | ‰ Percentage | 05.35‰ | 04.07‰ |
| **Spectral colors** | | | |
| بني | Brown | 116 | 99 |
| وردي | Pink | 45 | 97 |
| بنفسجي | Purple | 8 | 17 |
| رمادي | Grey | 13 | 34 |
| برتقالي | Orange | 9 | 9 |
| خشبي | Wooden | 6 | 7 |
| غامق | Dark | 1 | 8 |
| المجموع | Total | 198 | 271 |
| نسبة ‰ | ‰ Percentage | 01.21‰ | 01.56‰ |
| المجموع الكلي | Overall Total | 2136 | 1952 |
| نسبة ‰ | ‰ Percentage | 13.15‰ | 11.26‰ |

Two linguistic behaviors can be considered among the intensifying mechanisms: word elongation and comparative/superlative adjectives. Word elongation or keystroke repetition occurs when a user repeats one letter as a sign of emphasis or intensity. This feature in the literature has been mostly associated with female tweeters [62], [63]. Given that capital letters which are used in English for screaming are not possible in Arabic orthography, letter replication is rather used in Arabic for the same purpose. In our corpus, women are more users of letter duplication (90.85‰ of the whole data) as in Table 6. It should be noted that we did not count repetitive consonants relevant to other functions such as laughs /hhhh/ or /khkhkhkk/ (see §1.1). As apparent in Table 6, unlike other consonants, vowels and glides are the most frequently duplicated letters.

Comparative and superlative adjectives may as well fall under the category of intensifiers. Comparison tips the scale for one item over the other(s), thus supplying the utterance with more emphasis. In contrast to earlier works [64] that reserve adjectives for females' language, Table 7 illustrates that males compare the most (i.e. 60.35‰ vs. 45.48‰ for females).

In sum, no significant differences between the two genders in terms of intensifiers or hedges exist. However, Saudi males tend to compare more (i.e. intensifiers) and raise tag questions (i.e. hedges) than females. As for elongated letters, they are exclusively among the salient properties of the Saudi females' language.

**TABLE 11.** The frequency of emojis in SDTwittC.

| Emojis | Translation | Males | Females |
|---|---|---|---|
| 😉 | Winking face | 652 | 244 |
| ❤ | Red heart | 9158 | 16762 |
| 😔 | Pensive face | 80 | 643 |
| 🎶 | Music | 44 | 509 |
| 🚶 | Person walking | 321 | 711 |
| 😂 | Loudly crying face | 10628 | 21033 |
| 💔 | Broken heart | 4768 | 10761 |
| 💚 | Green heart | 1453 | 823 |
| 😷 | Face with medical mask | 163 | 366 |
| 💦 | Splash of water | 9 | 87 |
| 😊 | Smiling face | 1114 | 329 |
| 🎼 | Music | 105 | 191 |
| 😘 | Face blowing a kiss | 157 | 1360 |
| 😁 | Beaming face with smile | 1052 | 639 |
| 💕 | Two hearts | 134 | 24 |
| 😅 | Grinning face with sweat | 1074 | 523 |
| أخرى | Others | 54150 | 56095 |
| المجموع | Total | 85062 | 111100 |
| ‰ نسبة | ‰ Percentage | 523.68‰ | 641.38‰ |

**TABLE 12.** The frequency of lexical items in different domains in SDTwittC.

| Random Phrases | Translation | Male | Females |
|---|---|---|---|
| MSA Phrases لا شك، يا إلهي، حسنا، أحمق إلخ | Definitely, oh my god, alright, idiot etc. | 3098 19.07‰ | 3496 20.18‰ |
| Arabic Pidgin Phrases قرقر، سيم سيم، نفر، كويس إلخ | Gibberish, same, a person, good etc. | 89 00.54‰ | 154 00.88‰ |
| Arabicized English Phrases هكر، واتس، بلوك، سيلفي إلخ | Hacker, WhatsApp, block, selfie etc | 3941 24.26‰ | 5059 29.20‰ |
| Political Phrases حكومة، مواطن، فساد، وزير إلخ | Government, civilian, corruption, minister etc | 8278 50.96‰ | 6875 39.68‰ |
| Economic Phrases راتب، فلوس، سعر، غالي إلخ | Salary, money, price, expensive etc | 3932 24.20‰ | 4135 23.87‰ |
| Sport Phrases مباراة، الهلال، النصر، تسلل، هدف إلخ | Game, Al-hilal, Al-nasser, off-side, goal etc | 18027 110.98‰ | 10226 59.03‰ |
| Social Phrases أسرة، أمي، أبي، عرس، زواج إلخ | Family, mother, father, wedding, marriage etc | 4788 29.47‰ | 7548 43.57‰ |
| Health Phrases دايت، طبخ، أكل، بشرة، إلخ | Diet, cooking, food, compulsion, etc | 6454 39.73‰ | 9252 53.41‰ |
| Education Phrases جامعة، ماجستير، دكتوراه، محاضرة إلخ | University, Master, PhD, lecture etc | 1522 09.37‰ | 1470 08.48‰ |

### *b: EMOTION AND REASON-RELATED FEATURES*

Emotion (i.e. heart) and reason (i.e. mind) are integral parts of the human psychology. There is a widely established view that women are more emotional than males [65], [66]. In contrast, men focus more on facts, logic and reason [67]. Given that the number of emotion-related words (i.e. love, sadness, anger, hate etc) is too large for our analytical purposes, we confined our attention to two emotional aspects that emphasize gender differences: (i) romantic and (ii) religious phrases. The former involves personal feelings towards human beings while the latter encompasses personal beliefs in a non-human power, i.e. God. Consider the results in Table 8 for both romantic and religious phrases.

As shown in Table 8, our dataset reveals that Saudi females use more romantic phrases (42.62‰) than their male counterparts, i.e. (26.81‰).The statistics of romantic phrases confirm the validity of the past findings that females are more emotional [68], [69]. However, they are in conflict with Alsmearat *et al.* [27] who obtains no decisive evidence in this regard.

Part of Alsmearat *et al.*'s [27] misanalysis may follow from the fact that they drew on emotion-bearing words from Mohammad and Turney's [70] Emotion Lexicon (EmoLex). Although EmoLex consists of English lexemes, the authors simply translated them into Arabic using Goolge Translate service. We take issues with this methodology due to the predictable inaccuracy of Google Translate as well as the irrelevance of English terms to Arabic. Going through Emolex, we have also noticed some phrases irrelevant to emotion.

Concerning religious phrases, several studies have postulated that women are more religious than men [71]–[74]. Our corpus provides similar findings. 154.86‰ of the females' speech is religious whereas 138.84‰ of the same phrases is only observed in the males' content.

In contrast to emotion, reason is manifested in facts, logic and statistics. Limiting our focus to the frequency of numbers (such as 1-10, 20-90, 100, 1000, 1000.000) and timing terms, we have found remarkable differences between the two genders. Consider the results in Table 9.

**TABLE 13.** The frequency of punctuation marks and other characters in SDTwittC.

| Punctuation Marks/Characters | Males | Females |
|---|---|---|
| ! | 14132 | 7772 |
| : | 12848 | 4815 |
| . | 72225 | 56945 |
| ( | 3275 | 1653 |
| ) | 3589 | 2041 |
| " | 6923 | 7568 |
| * | 1446 | 1141 |
| ' | 638 | 620 |
| % | 543 | 71 |
| @ | 79 | 170 |
| # | 47 | 41 |
| Others | 8414 | 4130 |
| Total | 124159 | 86991 |
| ‰ Percentage | 764.38‰ | 502.20‰ |

As evident in Table 9, Saudi male speech includes a high frequency of cardinals: 47.44‰ for males vs. 27.66‰ for females. It also follows naturally that males are more accurate in their perception of timing such as days and months.

As an overall conclusion, Saudi males embed 108.41‰ as number and time-related terms in their streams whereas females incorporate only two thirds of the same percentage, viz. 66.87‰.

### c: USE OF COLORS AND EMOJIS

The perception of color has been associated with the emotional component of a human being [75], [76]. Previous studies confirm that females are capable of identifying more colors than males [77]–[79]. Yet, our results in Table 10 do not support this line of research.

As manifested in Table 10, males (13.15‰) make use of colors than females (11.26‰), contrary to the earlier findings. Upon close inspection, however, Saudi females are still the winners in terms of spectral colors (i.e. 01.56‰) while males produce 05.35‰ as basic colors. The males' high number of basic colors might be a result of the females' tendency to employ spectral colors only; in other words, the exclusive use of basic colors by males might increase the statistics and give the male users an advantage in frequency.

In twitter and other social networking applications, users also supplement their texts with emojis to illustrate their feelings graphically and to apply more emphasis on their emotions. Emojis are smileys and ideograms used to convey facial expressions and represent objects such as body parts, animals, flags among many others. In SDTwittC, and in support of earlier research [80], [81], Saudi females embed 641.38‰ as emojis in their twitter-feeds whereas their male fellows allocate only 523.68‰ of their texts to emoticons. Consider the results in Table 11.

**TABLE 14.** The frequency of function words in SDTwittC.

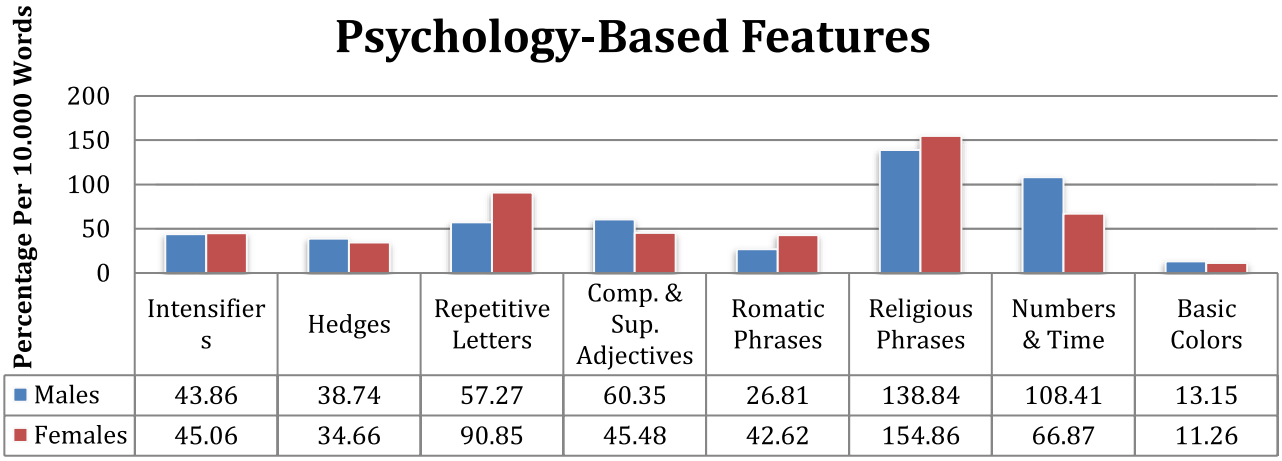
| Function Words | Translation | Males | Females |
|---|---|---|---|
| Pronouns أنا، انتو، انتم، هم، حنا إلخ | I, you, they, we etc | 19118 | 22089 |
| Demonstratives هذا، هذي، هاذي، هذولا، ذاك إلخ | this, these, that etc | 12041 | 8505 |
| Wh-words ايش، ليش، شلون، كم إلخ | What, why, how, how much etc | 9236 | 9913 |
| Negation Markers ما، مو، مب، لا، إلخ | no, not, etc | 28878 | 29852 |
| Yes/No Responses نعم، لا، ايه، ايوا، يب، يس، يب، نو إلخ | Yes, no, yea, yeah, yep, yes, no etc | 15478 | 16015 |
| Feminine Words ending with -at فاطمة، مدرسه، قرية إلخ | Fatima, school, village etc | 167458 | 178751 |
| Plural Words ending with –u:na/i:na and aat مسلمين، امهات، إلخ | Male Muslims, mothers etc | 52305 | 52073 |
| المجموع ‰ نسبة | Total ‰ Percentage | 599792 3692.60 ‰ | 624483 3605.17‰ |

In sum, Saudi males generally use more basic colors than females yet females use more spectral colors than males. Regarding emoticons, females generate a high amount of emojis than their male counterparts.

### 3) LEXICON-BASED FEATURES

This section discusses lexical items derived from other Arabic varieties such as MSA, Arabic Pidgin and Arabicized English. It also discusses well-known words and phrases in different domains such as politics, economy, sport, health and education. Given that males and females differ in their interests, a divergence in lexical selection is expected [82]. We restricted our search to the most common and predictable 20 vocabulary items in each field. The results are given in Table 12 below.

As presented in Table 12, no radical difference is noted in most of the categories. However, males tend to delve into sport and government topics whereas females prefer to tackle issues related to health and social life. Bringing all the psychology-based features together, consider Figure 2.




| | Intensifiers | Hedges | Repetitive Letters | Comp. & Sup. Adjectives | Romatic Phrases | Religious Phrases | Numbers & Time | Basic Colors |
|---|---|---|---|---|---|---|---|---|
| Males | 43.86 | 38.74 | 57.27 | 60.35 | 26.81 | 138.84 | 108.41 | 13.15 |
| Females | 45.06 | 34.66 | 90.85 | 45.48 | 42.62 | 154.86 | 66.87 | 11.26 |

**FIGURE 2.** Summary of Psychology-Based Features in SDTwittC.

### B. SYNTACTIC AND CHARACTER-BASED FEATURES

This section is concerned with the frequency of punctuation marks and other special characters. Recall that hashtag and username symbols (# and @ respectively) are removed from the dataset. However, they are still used in other irrelevant contexts. As seen in Table 13, a sharp contrast is detected. Saudi male tweeters tend to embed 764.38‰as characters in their replies than their female counterparts, i.e. 502.20‰usage.

### C. FUNCTION WORDS

As for function words, they include all expressions with grammatical functions such as pronouns, demonstratives, wh-interrogatives, definite article, negation markers, yes-no responses, feminine words (ending in -at or -ah) and plural words (ending with -u:na/i:na or -a:t). The results are given in Table 14.

The overall results indicate that no significant differences are observed between the two genders in terms of function words. However, Saudi women tend to use more pronouns than their male counterparts who are the most users of demonstratives.

## V. CONCLUSION

In conclusion, the contribution of the current paper is twofold: (i) a manually created Saudi dialectal dataset, and (ii) an inventory of features for machine learning-based tasks. The corpus is sufficient enough in that it includes 4-million-word texts from twitter replies, which represent the most authentic natural linguistic source. As for the set of features, they are proposed in light of the gender-based distinctions in the fields of sociolinguistics and psycholinguistics. The study derives its importance from these features that can feed machine learning algorithms and improve their performance and accuracy in author identification. Due to space limitations, the current paper does not measure the impact of using these features on machine and deep learning classifiers. Thus, we recommend that future works pursue this line of research and take our proposed features into consideration.

The corpus presents new findings contra the past studies. From the sociological point of view, for instance, we have found that Saudi males are more polite than their fellow citizens in terms of greeting, gratitude, apology, congratulation, encouragement and best wishes. Nonetheless, Saudi males are still impolite in terms of profanity and sarcasm. As for the psychometric features, and in line with the past literature, we have noted that Saudi females are more emotional and psychologically aware, due to their extensive use of romantic and religious phrases, emoji and spectral colors. No significant distinction has been drawn between the two genders as far as intensifiers and hedges are concerned. As part of the intensifiers, however, duplicated letters have been attested among the females whereas comparative and superlative adjectives are mostly used by the males. Regarding lexicon-based features, we have shown that Saudi males are more involved in government and sport topics while their female fellows are more concerned with health and social life. In terms of syntactic and character-based features, we have seen that Saudi men use more of these items. In the final section dedicated to function words, males are reported as the most users of demonstratives whereas females are seen as the most users of pronouns.

## REFERENCES

[1] A. Abbasi and H. Chen, "Visualizing authorship for identification," in *Proc. 4th IEEE Int. Conf. Intell. Secur. Inform.*, San Diego, CA, USA, May 2006, pp. 60–71.

[2] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, Feb. 2006.

[3] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digit. Invest.*, vol. 8, no. 1, pp. 78–88, Jul. 2011.

[4] L. C. Cagnina and P. Rosso, "Detecting deceptive opinions: Intra and cross-domain classification using an efficient representation," *Int. J. Uncertainty, Fuzziness Knowl. Based Syst.*, vol. 25, no. 2, pp. 151–174, Oct. 2017.

[5] F. Rangel and P. Rosso, "On the impact of emotions on author profiling," *Inf. Process. Manage.*, vol. 52, no. 1, pp. 73–92, Jan. 2016.
[6] P. Juola, "Authorship attribution," *Found. Trends Inf. Retr.*, vol. 1, no. 3, pp. 233–334, Mar. 2008.
[7] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, 2009.
[8] M. AL-Smadi, M. Al-Ayyoub, H. Al-Sarhan, and Y. Jararweh, "Using aspect-based sentiment analysis to evaluate Arabic news effect on readers," in *Proc. IEEE/ACM 8th Int. Conf. Utility Cloud Comput. (UCC)*, Limassol, Cyprus, Dec. 2015, pp. 436–441.
[9] J. Albadarneh, B. Talafha, M. Al-Ayyoub, B. Zaqaibeh, M. Al-Smadi, Y. Jararweh, and E. Benkhelifa, "Using big data analytics for authorship authentication of Arabic tweets," in *Proc. IEEE/ACM 8th Int. Conf. Utility Cloud Comput. (UCC)*, Limassol, Cyprus, Dec. 2015, pp. 448–452.
[10] J. Tetreault, J. Burstein, and C. Leacock, Eds., *The 8th Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, GA, USA: Association for Computational Linguistics, 2013.
[11] O. F. Zaidan and C. Callison-Burch, "Arabic dialect identification," *Comput. Ling.*, vol. 40, no. 1, pp. 171–202, Mar. 2014.
[12] A. M. Rezaei, "Author gender identification from text," M.S. thesis, Dept. Comput. Eng., Eastern Medit. Univ., Gazimağusa, Northern Cyprus, 2014.
[13] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, "Author gender identification from Arabic text," *J. Inf. Secur. Appl.*, vol. 35, no. 1, pp. 85–95, Aug. 2017.
[14] F. Rangel, P. Rosso, M. Potthast, and B. Stein, "Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter," in *Proc. Work. Notes Papers CLEF*, Sep. 2017, pp. 1–26.
[15] F. Rangel, P. Rosso, M. Montes-y-Gómez, M. Potthast, and B. Stein, "Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter," in *Proc. Work. Notes Papers CLEF*, 2018, pp. 1–38.
[16] S. Hussein, M. Farouk, and E. Hemayed, "Gender identification of Egyptian dialect in Twitter," *Egyptian Inform. J.*, vol. 20, no. 2, pp. 109–116, Jul. 2019.
[17] M. Koppel, N. Akiva, E. Alshech, and K. Bar, "Automatically classifying documents by ideological and organizational affiliation," in *Proc. IEEE Int. Conf. Intell. Secur. Inform.*, Dallas, TX, USA, Jun. 2009, pp. 176–178.
[18] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Personality Social Psycho. Bull.*, vol. 29, no. 5, pp. 665–675, May 2003.
[19] F. Peng, D. Schuurmans, V. Keselj, and S. Wang, "Automated authorship attribution with character level language models," in *Proc. 10th Conf. Eur. Chapter Assoc. Comput. Linguistics*, Budapest, Hungary, Apr. 2003, pp. 1–8.
[20] M. Crawford, *Talking Difference: On Gender And Language*. London, U.K.: Sage, 1995.
[21] R. Lakoff, *Language and Woman'S Place*. New York, NY, USA: Harper & Row, 1975.
[22] M. Talbot, *Language and Gender: An Introduction*. Hoboken, NJ, USA: Wiley, 1998.
[23] D. Brouwer, *Gender Variation in Dutch: A Sociolinguistic Study of Amsterdam Speech*. Dordrecht, The Netherlands: Foris, 1989.
[24] J. Coates, *Women, men, and language*. London, U.K.: Longman, 1993.
[25] J. Holmes, *Women, Men and Politeness*. London, U.K.: Longman, 1995.
[26] F. Rangel, P. Rosso, M. Koppel, E. Stamatatos, and G. Inches, "Overview of the author profiling task at PAN 2013," in *Proc. CLEF Conf. Multilingual Multimodal Inf. Access Eval. (CELCT)*, 2013, pp. 352–365.
[27] K. Alsmearat, M. Shehab, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, "Emotion analysis of Arabic articles and its impact on identifying the author's gender," in *Proc. IEEE/ACS 12th Int. Conf. Comput. Syst. Appl. (AICCSA)*, Nov. 2015, pp. 1–6.
[28] H. Han, C. Otto, X. Liu, and A. K. Jain, "Demographic estimation from face images: Human vs. machine performance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1148–1161, Jun. 2015.
[29] H. Han, A. K. Jain, F. Wang, S. Shan, and X. Chen, "Heterogeneous face attribute estimation: A deep multi-task learning approach," Jun. 2017, *arXiv:1706.00906*. [Online]. Available: https://arxiv.org/abs/1706.00906
[30] J. Vitak, N. Proferes, K. Shilton, and Z. Ashktorab, "Ethics regulation in social computing research: Examining the role of institutional review boards," *J. Empirical Res. Hum. Res. Ethics*, vol. 12, no. 5, pp. 372–382, Dec. 2017.
[31] M. Zimmer and N. J. Proferes, "A topology of Twitter research: Disciplines, methods, and ethics," *Aslib J. Inf. Manage.*, vol. 66, no. 3, pp. 250–261, May 2014.
[32] L. Mitchell, T. M. Sloan, M. Mewissen, P. Ghazal, T. Forster, M. Piotrowski, and A. Trew, "Parallel classification and feature selection in microarray data using SPRINT," *Concurrency Comput., Pract. Exper.*, vol. 26, no. 4, pp. 854–865, Sep. 2012.
[33] P. Trudgill, "Sex, covert prestige and linguistic change in the urban British English of Norwich," in *Language and Sex: Difference and Dominance*, B. Thorne and N. Henley, Eds. Rowley, MA, USA: Newbury House, 1975, pp. 88–104.
[34] D. Spender, *Man Made Language*. London, U.K.: Pandora Press, 1980.
[35] H. Spencer-Oatey, Ed., *Culturally Speaking: Managing Rapport Through Talk Across Cultures*. New York, NY, USA: Continuum, 2000.
[36] S. Okamoto, "Ideology and social meanings: Rethinking the relationship between language, politeness, and gender," in *Gender Practices in Language*, S. Benor, D. Sharma, and M. Rose, Eds. Stanford, CA, USA: CSLI, 2002, pp. 91–113.
[37] N. Lorenzo-Dus and P. Bou-Franch, "Gender and politeness: Spanish and British undergraduates' perceptions of appropriate requests," in *Género, Lenguaje y Traducción*, J. Santaemilia, Ed., Valencia, Spain: Univ. Valencia, 2003, pp. 187–199.
[38] Z. Gharaghania, A. E. Rasekh, A. Dabaghi, and I. Tohidian, "Effect of gender on politeness strategies in greetings of native speakers of Persian; English and EFL learners," *Cypriot J. Educ. Sci.*, vol. 3, no. 1, pp. 93–117, Sep. 2011.
[39] M. M. Bagwasi, "The effect of gender and age in Setswana greetings," *Southern Afr. Linguistics Appl. Lang. Stud.*, vol. 30, no. 1, pp. 93–100, Jul. 2012.
[40] H. Allami and M. Nekouzadeh, "Congratulation and positive politeness strategies in Iranian context," *Theory Pract. Lang. Stud.*, vol. 1, no. 11, pp. 1607–1613, Nov. 2011.
[41] Y. Al-Shboul and I. F. Huwari, "Congratulation strategies of Jordanian EFL postgraduate students," *Indonesian J. Appl. Linguistics*, vol. 6, no. 1, pp. 79–87, Jul. 2016.
[42] K. Schumann and M. Ross, "Why women apologize more than men: Gender differences in thresholds for perceiving offensive behavior," *Psychol. Sci.*, vol. 21, no. 11, pp. 1649–1655, Nov. 2010.
[43] G. Cordoni, E. Palagi, and S. B. Tarli, "Reconciliation and consolation in captive Western gorillas," *Int. J. Primatol.*, vol. 27, no. 5, pp. 1365–1382, Oct. 2006.
[44] T. Romero, M. A. Castellanos, and F. B. de Waal, "Consolation as possible expression of sympathetic concern among chimpanzees," *Proc. Nat. Acad. Sci. USA*, vol. 107, no. 27, pp. 12110–12115, Jul. 2010.
[45] A. Hesabi and M. Azima, "Speech act of thanking: A contrastive analysis among Iranian EFL learners in terms of gender and level of proficiency," *Int. Lett. Social Humanistic Sci.*, vol. 59, no. 1, pp. 76–84, Sep. 2015.
[46] K. Yusefi, H. Gowhary, A. Azizifar, and Z. Esmaeili, "A pragmatic analysis of thanking strategies among Kurdish speakers of Ilam based on gender and age," *Procedia Social Behav. Sci.*, vol. 199, pp. 211–217, Aug. 2015.
[47] K. Grammer, "Strangers meet: Laughter and nonverbal signs of interest in opposite-sex encounters," *J. Nonverbal Behav.*, vol. 14, no. 4, pp. 209–236, Dec. 1990.
[48] S. Petridis and M. Pantic, "Audiovisual discrimination between laughter and speech," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Las Vegas, NV, USA, Mar./Apr. 2008, pp. 5117–5120.
[49] P. E. McGhee, "The role of laughter and humor in growing up female," in *Becoming Female* (Women in Context: Development and Stresses), vol. 2, C. B. Kopp, Ed. Boston, MA, USA: Springer, 1979, pp. 183–206.
[50] G. Jefferson, H. Sacks, and E. A. Scheglo, "Notes on laughter in the pursuit of intimacy," in *Talk and Social Organisation*, G. Button and J. R. E. Lee, Eds. Clevedon, U.K.: Multilingual Matters, 1987, pp. 152–205.
[51] D. Bousfield and M. A. Locher, *Impoliteness in Language: Studies on its Interplay with Power in Theory and Practice*. Berlin, Germany: Mouton de Gruyter, 2008.
[52] T. Jay, *Why We Curse*. Amsterdam, The Netherlands: John Benjamins, 1999.
[53] J. Broadbridge, "An investigation into differences between women's and men's speech," M.S. thesis, Centre English Lang. Stud., Univ. Birmingham, Birmingham, U.K., 2003.
[54] A. Bowes and A. Katz, "When sarcasm stings," *Discourse Processes*, vol. 48, no. 4, pp. 215–236, May 2011.
[55] A. Drucker, O. Fein, D. Bergerbest, and R. Giora, "On sarcasm, social awareness, and gender," *Humor, Int. J. Humor Res.*, vol. 27, no. 4, pp. 551–573, Oct. 2014.

[56] J. House and G. Kasper, "Politeness markers in English and German," in *Conversational Routine*, F. Coulmas, Ed. The Hague, The Netherlands: Mouton, 1981, pp. 157–185.
[57] P. Brown and S. Levinson, *Politeness: Some Universals in Language Usage*. Cambridge, U.K.: Cambridge Univ. Press, 1987.
[58] K. L. Blankenship and T. Y. Craig, "Language and persuasion: Tag questions as powerless speech or as interpreted in context," *J. Exp. Social Psychol.*, vol. 43, no. 1, pp. 112–118, Jan. 2007.
[59] J. Holmes, *An Introduction to Sociolinguistics*. Harlow, U.K.: Pearson, 2008.
[60] B. Preisler, *Linguistic Sex Roles in Conversation*. Berlin, Germany: Mouton de Gruyter, 1986.
[61] P. M. Fishman, "Conversational insecurity," in *Language: Social Psychological Perspectives*, W. P. Robinson and P. Smith, Eds. Oxford, U.K.: Pergamon, 1980, pp. 127–132.
[62] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *Proc. 2nd Int. Workshop Search Mining Generated Contents*, Toronto, ON, Canada, Oct. 2010, pp. 37–44.
[63] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender in Twitter: Styles, stances, and social networks," Oct. 2012, *arXiv:1210.4567v1*. [Online]. Available: https://arxiv.org/abs/1210.4567v1
[64] S. Delgado and I. Lisbeth, "Gender differences in the use of English lexicon of the teachers at PUCESE I semester 2015," M.S. thesis, Dept. Appl. Linguistics, Pontifical Catholic Univ. Ecuador, Esmeraldas, Ecuador, 2016.
[65] L. R. Brody and J. A. Hall, "Gender and emotion in context," in *Handbook of Emotions*, M. Lewis and J. M. Haviland-Jones, Eds., 2nd ed. New York, NY, USA: The Guilford Press, 2000 pp. 395–408.
[66] J. R. Kelly and S. L. Hutson-Comeaux, "Gender stereotypes of emotional reactions: How we judge an emotion as valid," *Sex Roles*, vol. 47, nos. 1–2, pp. 1–10, Jul. 2002.
[67] D. Graham, "Gender styles in communication," Univ. Kentucky, Lexington, KY, USA, Tech. Rep., Sep. 2018.
[68] R. Parkins, "Gender and emotional expressiveness: An analysis of prosodic features in emotional expression," in *Proc. Griffith Working Papers Pragmatics Intercultural Commun.*, Jan. 2012, vol. 5, no. 1, pp. 46–54.
[69] A. Garimella and R. Mihalcea, "Zooming in on gender differences in social media," in *Proc. Workshop Comput. Modeling People's Opinions, Personality, Emotions Social Media (PEOPLES)*, Osaka, Japan, Dec. 2016, pp. 1–10.
[70] S. M. Mohammad and P. D. Turney, "Crowdsourcing a word–emotion association lexicon," *Comput. Intell.*, vol. 29, no. 3, pp. 436–465, Aug. 2013.
[71] D. P. Sullins, "Gender and religion: Deconstructing universality, constructing complexity," *Amer. J. Sociol.*, vol. 112, no. 3, pp. 838–880, Nov. 2006.
[72] L. Woodhead, "Why so many women in holistic spirituality? : A puzzle revisited," in *A Sociology of Spirituality*, F. Flanagan and P. C. Jupp, Eds. Aldershot, U.K.: Ashgate, 2007 pp. 115–126.
[73] M. Trzebiatowska and S. Bruce, *Why are Women More Religious Than Men?*. Oxford, U.K.: Oxford Univ. Press, 2012.
[74] R. N.-B. Shahar, "'At 'amen meals' it's me and God' religion and gender: A New Jewish women's ritual," *Contemp. Jewry*, vol. 35, no. 2, pp. 153–172, Jul. 2015.
[75] J. H. Xin, K. M. Cheng, G. Taylor, T. Sato, and A. Hansuebsai, "Cross-regional comparison of colour emotions Part I: Quantitative analysis," *Color Res. Appl.*, vol. 29, no. 6, pp. 451–457, Dec. 2004.
[76] F. Leichsenring, "The influence of color on emotions in the Holtzman Inkblot technique," *Eur. J. Psychol. Assessment*, vol. 20, no. 2, pp. 116–123, Sep. 2006.
[77] J. Lyons, "Colour in language," in *Colour: Art and Science*, T. Lamb and J. Bourriau, Eds. New York, NY, USA: Cambridge Univ. Press, 1995, pp. 194–224.
[78] H. Arthur, G. Johnson, and A. Young-Jones, "Gender differences and color: Content and emotion of written descriptions," *Social Behav. Personality Int. J.*, vol. 35, no. 6, pp. 827–834, Jan. 2007.
[79] J. S. Alowibdi, A. U. Buy, and P. Yu, "Language independent gender classification on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Niagara, ON, Canada, Aug. 2013, pp. 739–743.
[80] Y. Nishimura, "A Sociolinguistic analysis of emoticon usage in Japanese blogs: Variation by age, gender, and topic," in *Proc. 16th Annu. Meeting Assoc. Internet Researchers*, Phoenix, AZ, USA, Oct. 2015, pp. 1–17.
[81] Z. Chen, X. Lu, W. Ai, H. Li, Q. Mei, and X. Liu, "Through a gender lens: Learning usage patterns of emojis from large-scale Android users," in *Proc. World Wide Web Conf.*, Apr. 2018, pp. 763–772.
[82] A. Baquee, "Influence of gender roles in language choice: A study on male and female students of private universities in Dhaka city," M.S. thesis, Dept. Eng. Humanities, BRAC Univ., Dhaka, Bangladesh, 2016.

**SAAD AWADH ALANAZI** received the B.S. degree in computer science from Jouf University, Saudi Arabia, in 2007, the M.S. degree in computer science from Ball State University, in 2011, and the Ph.D. degree in artificial intelligence from Staffordshire University, in 2017. He is currently an Assistant Professor and the Chairman of the Computer Science Department, Jouf University. His research interests include natural language processing, text mining, and image processing.