ArabGend: Gender Analysis and Inference on Arabic Twitter

Hamdy Mubarak, Shammur Absar Chowdhury and Firoj Alam
Qatar Computing Research Institute, HBKU, Qatar
{hmubarak, shchowdhury, fialam}@hbku.edu.qa

Abstract

Gender analysis of Twitter can reveal important socio-cultural differences between male and female users. There has been a significant effort to analyze and automatically infer gender in the past for most widely spoken languages' content, however, to our knowledge very limited work has been done for Arabic. In this paper, we perform an extensive analysis of differences between male and female users in the Arabic Twitter-sphere. We study differences in user engagement, topics of interest, and the gender gap in professions. Along with gender analysis, we also propose a method to infer gender by utilizing usernames, profile pictures, tweets, and networks of friends. In order to do so, we manually annotated gender and locations for ∼167K Twitter accounts associated with ∼92K user location, which we make publicly available.[1] Our proposed gender inference method achieves an F1 score of 82.1% (47.3% higher than the majority baseline). We also developed a demo and made it publicly available[2].

1 Introduction

Demographic information (e.g., age, gender) has proven to be useful in many different decision-making processes such as, from business decisions (e.g., personalized online advertising), forensic investigation to policy-making purposes (Li et al., 2016; Volkova et al., 2013; Mukherjee and Liu, 2010; Soler and Wanner, 2016). For example, social media platforms and e-commerce sites are using customers' gender and other demographic attributes for targeted advertising (Tuan et al., 2019). In the past decade, there have been extensive research efforts to automatically infer demographic attributes of the social media users using their social media footprints (e.g., users' posts, names, and other attributes) (Chen et al., 2015; Volkova et al., 2015). In addition, to evaluate the performance of the model's fairness for different tasks it is important to have such attributes (Chakraborty et al., 2021; Wang et al., 2019). Given that such attributes are often removed from the original source for privacy and ethical reasons, however, having such attributes through inference is a possible way to evaluate the model's fairness.

Major research efforts for such attributes inference are mostly done for English, and very little effort has been given to non-English languages (Chakraborty et al., 2021). The research for Arabic demographic inference such as gender is relatively rare for social media users, specifically on Twitter's content. With approximately 164 million monthly active users, Twitter is one of the most popular social media platforms in the Arab region (Abdelali et al., 2021). The large volume of tweets produced represents the social and cultural characteristics of the region. Even though there is a large number of Twitter users, however, usage of Twitter differs in volume, topics, and engagement depending on the users' gender role. Another important factor is that social media users often provide misleading demographic information (e.g., name, age, location and marital status), which is highlighted in a survey conducted in the Arab region (Salem, 2017). Hence, self-declared information might not be always reliable. Though some studies argue that the proportion of such misleading self-reported information is relatively lower (Herring and Stoerger, 2014). While the availability of Twitter data and its large user base provides opportunities to understand such information, however, for privacy reasons, Twitter does not share users' gender information (Mueller and Stumme, 2016). Such factors stress the need to have automatic methods for gender inference, and here our focus is Twitter-sphere for the Arabic region. In addition, there is a gap in the literature in a thorough analysis of Arabic Twitter (e.g., linguistic content) for gender, even

[1] https://alt.qcri.org/resources/ArabGend.zip
[2] https://asad.qcri.org/

Proceedings of the 2022 COLING Workshop: The 8th Workshop on Noisy User-generated Text (W-NUT 2022), pages 124–135

though Arabic is a morphologically rich language where linguistic markers are present to distinguish genders in many cases (see Section 3.1).

To address the gap of gender analysis and automatic inference, in this paper, we perform an extensive analysis of Arabic Twitter data where we identify key distinguishing properties of male/female authorship. We experiment with different features to identify the gender of Twitter users. We examine the usage of friendship networks, profile pictures, and textual information such as usernames, user descriptions, and tweets to classify gender. The contributions of our work are as follows:

- We developed a new dataset of ∼167K Twitter accounts that are manually annotated for their gender and location, which we make publicly available for research purposes.
- We show differences between the two genders from different angles such as presence on Twitter in different Arab countries, language usage, etc.
- We show signs of gender gaps in the labor market which align with some official reports.
- We study automatic gender identification of tweets, user accounts, and user descriptions. We also study how profile pictures and networks of friends can influence results.
- Using our models we developed a demo, which is publicly available.

2 Related Work

Gender inference is a well-studied problem in English. Liu and Ruths (2013) present a dataset of 13K gender-labeled Twitter users and propose the use of first names as features for gender inference. Screen_names, full names, user descriptions, and tweets have also been used as features for gender inference (Burger et al., 2011). Rao et al. (2010) use stacked SVMs for identifying gender and other latent attributes of Twitter users. Semi-supervised methods that exploit social networks have also been used for gender classification (Li et al., 2016).

Gender inference has also received attention for a few other languages. Sakaki et al. (2014) combine the output of text processor and image processor to infer the gender of Japanese Twitter users. Taniguchi et al. (2015) propose a hybrid method that uses logistic regression to combine text and image features. Ciot et al. (2013) label 1000 users for gender in each of the following languages: Japanese, Indonesian, Turkish, and French. The authors use Support Vector Machines (SVMs) for classification. Sezerer et al. (2019) present a dataset consisting of 5.5K Twitter users labeled for their gender. Tuan et al. (2019) proposes clustering-based approaches for demographic analysis to support advertising campaigns. Very recently Liu et al. (2021) provided a large-scale study that investigate different inference techniques (e.g., classic machine learning to deep learning models) using Twitter data. The authors highlight that a simpler model performs well to infer age, however, sophisticated models (e.g., sentence embeddings) are important for gender.

For Arabic, on the other hand, work is relatively less explored. Malmasi (2014) use first names to classify the gender of Arabic, German, Iranian and Japanese names. ElSayed and Farouk (2020) uses neural networks to differentiate male and female authors of tweets in Egyptian dialect. Hussein et al. (2019) use classic machine learning classifiers such as Logistic Regression and Random Forest classifiers to identify gender in Egyptian tweets. Habash et al. (2019) use deep learning for gender identification and uses Machine Translation for reinflection. Bsir and Zrigui (2018) use the gated recurrent unit (GRU) for gender identification in Facebook and Twitter posts. Zaghouani and Charfi (2018) collect a corpus of 2.4M multi-dialectal tweets from 1600 accounts that are tagged for gender, age, and language. Wang et al. (2019) propose a new multilingual (32 different languages), multimodal, multi-attribute deep learning system for inferring different demographic attributes.

Our work differs from previous work on gender analysis and inference for Arabic in a number of ways *(i)* it uses a much bigger dataset for male and female users; *(ii)* it has no bias towards a specific country as it covers users from all Arab countries; *(iii)* it uses a generic method for collecting users and their names as opposed to starting with a specific list of names, which can be skewed towards some countries or cultures; *(iv)* in addition to gender inference, we perform a thorough analysis of gender differences in their profile descriptions, topics of interest, the profession gender gap among other things.

3 Dataset

3.1 Background

In Arabic, typically nouns and adjectives have gender markers such as Taa Marbouta letter "ة" as a feminine (f) suffix, and in case of absence, they can be considered as masculine (m). There are special cases where a word can have the feminine marker and it's gender is unknown (e.g., داعية - religious scholar (m and f)). Also, there are some cases where words are feminine without explicit gender markers (e.g., أنثى ، بنت - female, girl). Except for some special cases, converting gender from masculine to feminine can be done by appending the Taa Marbouta suffix "ة", e.g., words like مديرة ، شاعرة (manager(f), poet(f)) are the feminine forms of مدير ، شاعر (manager(m), poet(m)) in order.

It's widely observed that many Arabic users on Twitter describe themselves in the description field in their profiles. This description expresses several identity features such as nationality (NAT), profession or job (PROF), interest (INT), social role (SOC), religion (RELIG), and ideology (IDEO) among others. We provide a few examples in Table 1.

Description	Translation	Class
عراقي وأفتخر	Iraqi (m) and proud	NAT
مواطنة سعودية	Saudi citizen (f)	NAT
طبيبة أسنان	Dentist (f)	PROF
طالب دكتوراه	PhD student (m)	PROF
عاشقة الطبيعة	Nature lover (f)	INT
مهتم بأخبار التقنية	Interested (m) in IT news	INT
زوجة وأم	Wife and mother	SOC
شاب متفائل	Optimistic young man	SOC
مسلم وأفتخر	Muslim (m) and proud	RELIG
مسيحية عربية	Arab Christian (f)	RELIG
سياسي معارض	Opposition politician (m)	IDEO
ليبرالية أحب بلدي	Liberal (f), love my country	IDEO

Table 1: Examples of user description with gender (m/f) and identity label (class).

3.2 Data Collection

For the data collection, we used Twitter API to crawl Arabic tweets using a language tag to Arabic ("lang:ar"), back in January 2018. We collected data in two phases. *First*, we collected 4.35M tweets (*termed as former set*), which covers tweets

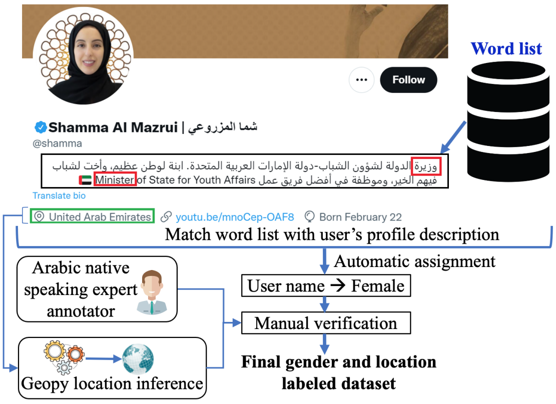


Figure 1: Our pipeline to develop **ArabGend** – labeling gender and location.

from 2008 to January 2018.[3] Using this dataset we developed a word list using a gender marker (see Section 3.3.1). In the *second* phase, we collected additional 100M millions tweets (*termed as later set*), dated from 2018 to 2020, to develop final annotated dataset (see Section 3.3.2). The purpose of the *former set* of tweets was to create a gender marker word list, the purpose of the *later set* of tweets was to create a large annotated dataset with gender and location labels. We used such an approach to avoid any biases that may appear due to the word list selection.

3.3 Annotation

3.3.1 Creating Word List with Gender Info

For the annotation, we first created a word list of gender markers. In order to do that we first extracted all profile information of users who posted these tweets. From the user's profile description, we obtained a list of all the first words that users used to describe themselves.[4] We obtained a unique list of 10K words. We then excluded words that appeared only once, which resulted in a list of ∼2,500 words out of 10K. We used the publicly available Farasa tool (Darwish and Mubarak, 2016) to initially detect the gender of each word in the list. Then, a native speaker revised gender information and provided both the masculine and feminine word forms and their different writings to have better coverage. For example, for the feminine form "محامية - lawyer (f)", the masculine form and its different writings "محامي ، محامى ، محامٍ - lawyer (m)" was also added if they did not appear in the word

[3]Note that our data collection might not consist of all of the tweets posted on Twitter during this period, which is because Twitter's free API has a limit.

[4]the First word is a very strong signal in identity description and can be mapped to gender.

list. The final gender marker word list contains 713 words, in which 56% of them indicate masculine and 44% indicate feminine gender.[5] The list can be found in our publicly released dataset.

3.3.2 Gender and Location Annotation

For gender and location annotation, we first collected another set of 100M tweets, *the later set*, which dated from 2018 to 2020.

Gender: We annotated 100M tweets with gender and location information in several steps. We used the word list, discussed in the previous section, and matched the words at the beginning of each user's profile description. The matching approach resulted to assign a gender label to ∼167K users. We could not able to assign the gender label for the rest of the users due to the mismatch between our created word list, and the empty user's profile description. We then manually revised the assigned gender labels of these 167K users by a native Arabic-speaking expert annotator. In Figure 1, we present *ArabGend* development pipeline that demonstrates how the user profile appears, how we used profile description with the word list to the assign gender marker and location information to assign a specific location. Note that we developed the *word list*, highlighted in blue, at the first phase of our dataset development, as discussed in Section 3.3.1. In this profile, user location is clearly visible, however, this is not always the case for which location inference is needed.

Location: Out of these 167K users we extracted 28K unique locations, which are then mapped into Arab countries with geographic location information using *GeoPy toolkit*.[6] Similar to gender annotation, the output of GeoPy is then manually revised by the same annotator. The annotation process resulted in identifying the countries for 92K users (55.08% of all users) out of 167K users. We could not identify the rest of user locations as they were either empty (38%) or cannot be mapped to a specific country (6.92%).

Removing Ambiguous and Inappropriate Accounts The manual annotation process consists of another step to remove ambiguous, inappropriate (e.g., adult and spam) accounts. Typically Arabic words are written without diacritics, which causes ambiguity in many cases, e.g., the word مدرسة can be interpreted as Teacher (f) or School. As we are interested in collecting personal accounts using their profile description, therefore, we excluded organizations' accounts from our data collection. Also, there are some titles that can be used to describe males and females, which we removed. For example, دكتور، مدير (Doctor, Manager) are used for both genders.

To filter adult and spam accounts we used the publicly available APIs from ASAD system (Hassan et al., 2021). It is a social media analysis toolkit consisting of eight modules to classify dialects, sentiment, emotion, news category, offensiveness, hate speech, adult content, and spam in Arabic tweets.[7] Based on the classified output from ASAD and a manual inspection during the annotation process, we removed *inappropriate* accounts. We use the term *appropriateness* to refer to the labels non-adult and spam content in the rest of the paper. In this phase, after filtering non-personal and inappropriate accounts, we ended up with 167K users (80% are males and 20% are females).

3.3.3 Annotation quality

To assess the quality of the annotation, we manually annotated 500 users' accounts. We selected a random sample of 500 users and then manually assigned gender labels by checking their accounts on the Twitter platform. Agreement with manual annotation was ∼99%. Similarly, for location, we randomly selected another sample of 500 unique user locations and checked their mappings to countries. The accuracy was 98%, which indicates annotation quality is very high for gender and location labels. Note that, Twitter user locations are typically noisy, and mapping them to countries is not always trivial.

Accounts	**Count**	**User Loc.**
Male	133,192 (80.0%)	75,539 (81.5%)
Female	33,348 (20.0%)	17,115 (18.5%)
Total	166,540 (100%)	92,654 (56.0%)

Table 2: Statistics of the dataset.

3.3.4 Statistics

In Table 2, we report number of final male and female accounts and percentage of successful map-

[5] Words like شخص، كاهن، زول (person, priest, man) have no corresponding feminine words.

[6] https://pypi.org/project/geopy/, It is a python client for several geocoding web services including Nominatim (https://nominatim.org), which uses OpenStreetMap data to find location.

[7] https://asad.qcri.org

User Name	Description	User Loc.	G	C
صفية الشحي (Safia Alshehi)	إعلامية ـ كاتبة (journalist (f) and writer (f))	UAE - Dubai	F	AE
Ahmed Azhar	إنسان بسيط جدا (very simple person (m))	جدة (Jeddah)	M	SA

Table 3: Annotation: Description was mapped to Gender (G), and User Loc. was mapped to Country (C).

pings of user locations to countries for both genders. According to a report from the World Bank in 2015,[8] the gender gap in Middle East and North Africa region can reach to 34% in internet usage. This gap comes second after the largest gender gap in Sub-Saharan Africa region (45%). Further, while 52% of females (91M) have mobile phones, this ratio increases to 56% for males with additional 8M male users. These factors can explain the less presence of female users on Twitter as shown in our study. In Table 3, we present some annotation examples from our dataset. We use ISO 3166-1 alpha-2 for country codes.[9]

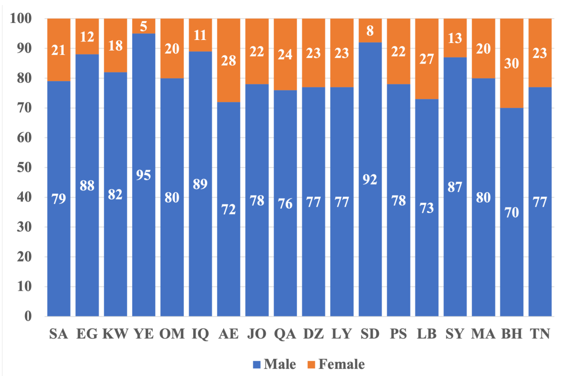


Figure 2: Gender distribution in Arab countries.

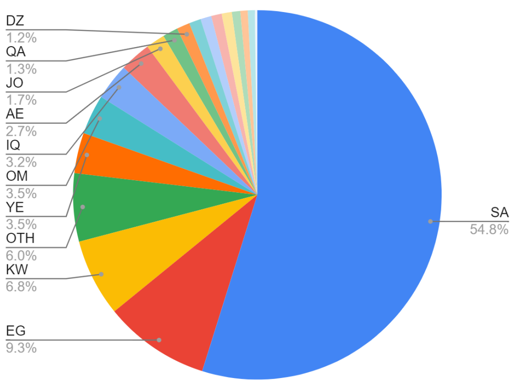


Figure 3: Country distribution of Twitter accounts.

4 Analysis

4.1 Gender and Location Distribution

In Figure 2, we present gender distribution of Twitter users in Arab countries. We observe that the top three countries that have higher percentages of female users. For BH (Bahrain), AE (United Arab Emirates) and LB (Lebanon) are 30%, 28% and 27%, respectively. The lowest percentages of female users from YE (Yemen), SD (Sudan) and IQ (Iraq) are 5%, 8% and 11%, respectively.

In Figure 3, we present country distribution of all accounts in our dataset. We observe that more than half of Twitter users are from SA (Saudi Arabia) and 70% of accounts are from Gulf region (SA, KW, OM, AE, QA and BH) followed by accounts from EG, YE, etc. Distributions are very similar to what was reported in a previous study to collect dialectal tweets in a different time span and using a different approach (Mubarak and Darwish, 2014).

We mapped user locations to OTHER (OTH) for the countries that are outside Arab World. They represent 6% of all user locations. Top five countries that are outside Arab World include US, GB, TR, DE and FR in order. In addition, we found that the dataset has 1,495 verified accounts, out of which 90% are male and 10% are female. Such numbers represent 1% and 0.45% verified male and female accounts, respectively.

4.2 User Engagement

We extracted the date of joining on Twitter for all accounts to study their engagement with Twitter. As shown in Figure 4 (in appendix), we can see that many accounts joined Twitter between 2010 and 2012, then the number of users who joined Twitter between 2013 and 2018 was almost stable for male and female accounts. Starting from 2019, there was an increasing number of users. We notice that there is a slightly increasing number of female accounts who join Twitter over time, however, Twitter was always dominated by male accounts and the gap between the two genders seems to increase in the future as shown in the cumulative chart in Figure 5 (in Appendix).

4.3 User Connections

Figure 6 shows an average number of followers and followees (friends) of male and female accounts in our dataset. We can see that on average, female accounts tend to attract more followers than males (more than double). Further, females have ∼30%

[8] https://blogs.worldbank.org/ar/arabvoices/ten-facts-about-women-arab-world

[9] https://en.wikipedia.org/wiki/List_of_ISO_3166_country_codes

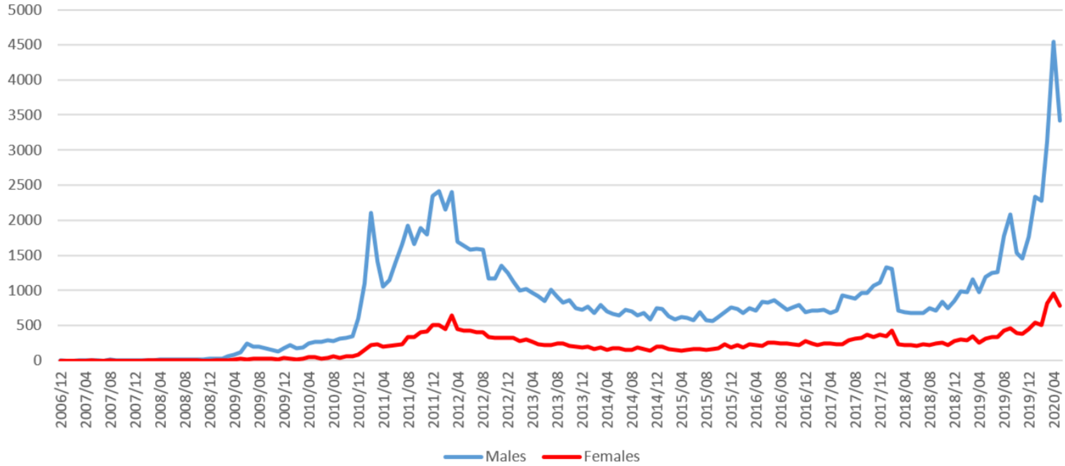


Figure 4: Distribution of Twitter joining date

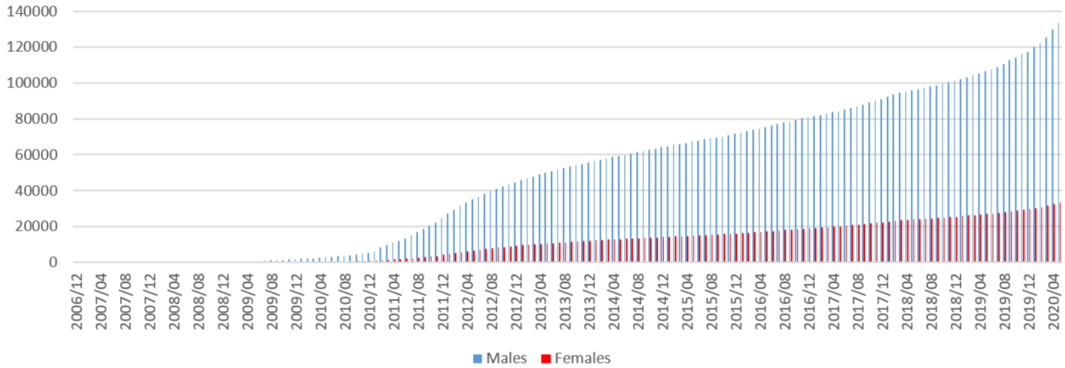


Figure 5: Accounts distribution over time

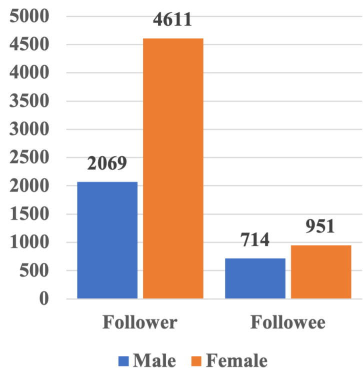


Figure 6: Followers and followees distribution.

more friends than males, which may indicate that females prefer to have a larger community and friends than males on Twitter.

4.4 Person Names

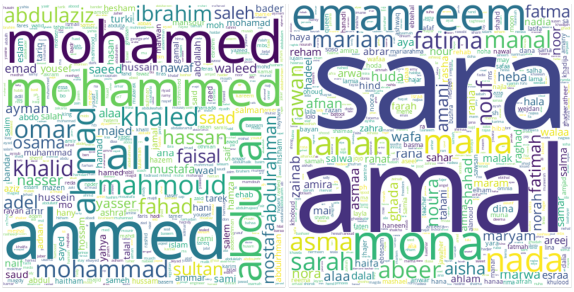
A person's name is a very important feature in identifying gender. To understand the demographics of Twitter users, prior studies have been using a seed list of names to collect male and female accounts. Mislove et al. (2011) used the most common 1000 male and female names in the US to collect Twitter user information. Such an approach, i.e., using a pre-specified list of person names, can create bias in the resulting data collection. In our study, we attempted to follow a different approach to avoid such a bias. We created *initial* dataset to create word list, and used a different set (i.e., the *later* 100M) to create the final list. We further normalized the names by removing diacritics, mapping Alif shapes, Taa Marbouta and Alif Maqsoura letters to plain Alif, Haa, and Yaa letters respectively, and mapping decorated letters to normal letters.

From the obtained lists, we can extracted names that are used for both genders when they are written in Arabic (e.g., نور، صباح، شمس - Nour, Sabah, Shams), or due to transliteration ambiguity, e.g., the names علاء(m) and آلاء(f) both are transliterated to "Alaa", also أمجد(m) and أمجاد(f) have the same transliteration "Amjad".

In Figure 7 we show the most common male and female names are written in English. Mostly, they have similar distribution as their Arabic counterparts with different ways of transliteration.

4.5 Interests According to User Description

In Figure 8 we present most common words used in user's profile description for males and fe-



(a) Male names **(b) Female names**

Figure 7: Common Arabic names (in English).

males in order. This gives an indication about jobs and interests for both genders. We can see that females tend to describe their social role (e.g., بنت، أم، فتاة، صديقة - daughter, mother, girl, friend) more than males. For comparison, while more than 1000 female accounts describe themselves first as أم (mother), less than 200 accounts describe themselves as أب (father). We can also see that a good portion of Twitter users is young (e.g., طالب، فتاة، خريجة - student, young woman, graduate) as opposed to few accounts who describe themselves as متقاعد (retired). From our analysis, we observed that self-description can be used to predict the age group of Twitter users. We leave this for future work.

(a) Male accounts **(b) Female accounts**

Figure 8: Description of male and female accounts. The top five for males are: engineer, student, lover, interested (in), and teacher. The top five for females are: student, graduate, teacher, girl, and mother.

4.6 Topics of Interest

In Figures 9 we present the common distinguishing words in tweets written by male and female accounts in our dataset. We computed the valence score discussed in (Conover et al., 2011; Chowdhury et al., 2020) with a threshold of 0.5 to obtain these words.

While tweets from males have many words related to politics (e.g., اليمن، الإخوان - Yemen, Muslim Brotherhood) and sports (e.g., الدوري، الهلال - league, Hilal club), tweets from females have many words related to family and society (e.g., أمي، أبناء، زميلات - my mother, children, colleagues) and feelings (e.g., قلبي، شعور، حبيبتي - my heart, feeling, my love).

4.7 Gender Gap in Professions

We can observe from Figure 8 that the most frequent profession for males was مهندس (engineer) while it was معلمه (teacher) for females. In Table 4, we report the distribution of some professions for male and female accounts in different domains. We observe that the Sport domain is overwhelmingly dominated by males, and other domains (e.g., Management, Software, Health, etc.) have less representation of females (percentages are from 9% to 20%). The best domain that has a good representation of females is the Translation domain with a percentage of 36%.

According to the World Bank's report in June 2020,[10] the labor force participation rate of females in the Middle East and North Africa region is around 20% with a slight improvement from 17.4% in 1990. Our study supports this report by showing that females are less represented in many job domains, and participation rates can be roughly quantified in different sectors of job markets. The same report also mentions that only 11% of females hold managerial positions compared to the world average of 27%.[11] The ratio of female managers to all managers in our dataset is 9% based on the self-description of the profile.

(a) Male tweets **(b) Female tweets**

Figure 9: Most common words.

5 Experiments

For the classification experiments, we focused only on the gender inference and leave the location in-

[10] https://data.worldbank.org/indicator/SL.TLF.CACT.FE.ZS?locations=ZQ

[11] www.dw.com/ar/, shorturl.at/vLOQT

Prof.	Translation	G	Freq.	%	Domain
لاعب	player	m	1,096	98	Sport
لاعبة		f	19	2	
مهندس	engineer	m	6,619	94	Engineering
مهندسة		f	404	6	
مدير	manager	m	2,982	91	Management
مديرة		f	286	9	
مبرمج	programmer	m	153	91	Software
مبرمجة		f	16	9	
محاسب	accountant	m	580	90	Finance
محاسبة		f	61	10	
طبيب	doctor	m	2,265	80	Health
طبيبة		f	577	20	
مترجم	translator	m	177	64	Translation
مترجمة		f	98	36	

Table 4: Profession gaps examples.

ference study as for a future study. We measure the performance of the classification models using accuracy (Acc), macro-averaged precision (P), recall (R) and F1 score. We use macro-averaged F1 score as a primary metric for comparison in our discussion.

5.1 Datasets

We used two datasets for training to provide a comparative study. We used our developed *ArabGend* dataset only for training. We also used *ARAP* dataset (Zaghouani and Charfi, 2018), which consists of 1,600 Twitter accounts labeled for their gender along with country and language. We used half of the *ARAP* dataset for training, and half for the evaluation. Hence, in our experiments, models are evaluated using the half of the *ARAP* dataset, which we considered as our test set.

5.2 Classification Models and Features

We used Support Vector Machines (SVMs) as our classifier. Our choice of SVM was influenced by a reasonable accuracy and a system deployment in a low computational resource setting. As features, we used character n-gram vectors weighted by term-frequency-inverse document term frequency (tf-idf). We experimented with different n-gram ranges. Only character [2-5] n-gram results are reported in this paper since they yielded the best results.

In addition, we also varied different types of input to the classifiers. We experimented with *(i)* a single tweet from each user, *(ii)* aggregate all tweets from a user, *(iii)* usernames of the Twitter users. We also experimented by balancing the *ArabGend* training set, to have equal number males and females, to understand the affect on the performance of the classifiers. Since ARAP dataset is balanced in terms of gender, hence, we do not apply any sampling to balance the data any further. Since there was not any significant improvement in performance after balancing to equal distribution, therefore, we do not report that results.

5.3 Results

In Table 5, we report the classification results on ARAP test set. From the results, we observed that for both ARAP data and *ArabGend* data, best results are obtained when usernames are used as opposed to aggregation of tweets or user descriptions. In general, aggregating tweets do not improve results in general by a significant margin. The usernames in the *ArabGend* dataset have a significant performance improvement over all other settings, resulting in an F1 score of 82.1.

Train Data	Features	Acc.	P	R	F1
Majority Baseline		53.3	26.7	50.0	34.8
	Usernames	67.2	67.1	66.8	66.8
ARAP (Baseline)	Description	58.2	58.5	58.5	58.2
	Tweets	69.8	70.9	70.4	69.7
	All Features	59.9	65.3	61.6	57.9
	Usernames	**82.4**	**82.7**	**82.0**	**82.1**
ArabGend	Description	64.1	65.4	62.7	61.8
	Tweets	63.1	62.9	62.9	62.9
	All Features	78.0	80.2	77.1	77.1

Table 5: Performance on ARAP test data

5.4 Additional Experiments

Predicting Gender from Profile Images To evaluate the efficiency of profile image based gender detection model we used Gender-and-Age-Detection model (Levi and Hassner, 2015) on ARAP test set. It uses deep learning to identify the gender and age of a person from face image, which was trained on ∼27K images from Flickr (Adience dataset) (Levi and Hassner, 2015).[12]

For comparison, we manually annotated the same ARAP test set for gender prediction using profile images and the accuracy was 81%. This shows that profile image can be one of the powerful features to predict gender. It is worth to mention that 87% of the package errors are due to misclassification of female users as males. We plan to use profile images with textual features in future.

[12] Accuracy of this model is 64%. Some images are hard for gender prediction, e.g., flag, natural scene, incomplete face, kid image, cartoon, mixed, etc.

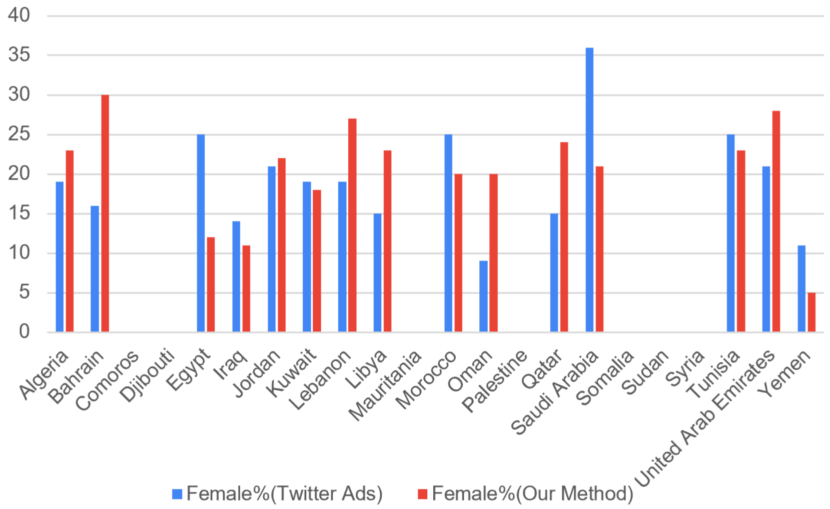


Figure 10: Distribution of female accounts

Predicting Gender from Friends Network Homophily (meaning love of the same) is a tendency in social groups for similar people to be connected together (McPherson et al., 2001). Homophily has predictive power in social media (Bischoff, 2012). We anticipated that female users on Twitter tend to have more female friends than male users and vice versa. To experiment this assumption, we collected a list of up to 100 friends for all accounts in the ARAP test set, and from their usernames, we used our classifier to predict their gender. We experimented with different thresholds on ratio of predicted male to predicted female friends to decide gender of our target users. The best results were obtained when 1/3 of friends of an account are predicted as females. In these cases, we propagated the label "female" to the account and propagated "male" otherwise. By doing so, we could achieve 56% accuracy. This shows that gender distribution of friends network has limited impact on determining gender of a user.

We also explored if information about friend's gender can improve the performance of the model from the earlier section. We adopt the following procedure: if the classifier is not confident that the instance is male, we apply the threshold technique above and take the classifier's predicted label otherwise. By doing this, we were able to improve the performance from 82.1% to 82.9% indicating that friend's gender might be helpful in cases where the classifier is not confident. However, obtaining a list of friends for all accounts needs a significant amount of time. This limits the usage of friends' gender in cases where fast response is needed.

Comparison with Twitter Ads API Advertisers on Twitter can target their campaigns based on geolocation, gender, language, and age. Twitter uses the gender provided by people in their profiles, and extends it to other people based on account likeness. We used Twitter Ads API to get total number of users in all Arab countries and their gender distribution. Figure 10 shows distribution of female users as obtained from Twitter Ads and our method. Although there are some differences between the two methods, the average percentages of female users are similar (19% using Twitter Ads vs. 20% using our method). This can show that our method is close to Twitter Ads for gender prediction of users although Twitter has much larger information to use. Note that Twitter Ads results (also our method) may have limitations in terms of accuracy.

6 Conclusion

In this paper, we have presented *ArabGend*, a new dataset of Twitter users labeled for their gender and location. To the best of our knowledge, this is the largest Arabic dataset for gender based analysis. We analyzed the characteristics of the users from a gender perspective. We identified key differences between male and female accounts on Arabic Twitter such as user connections, topics of interest, etc. We also studied the gender gap in professions and argued that results obtained from our dataset are aligned with recent reports from the World Bank and Twitter Ads information. We also showed that our dataset yields the best inference results on a publicly available test set. In the future, we plan to enhance our data collection method by considering gender markers in the whole user description and other profile fields.

Ethical Concern and Social Impact

User Privacy For privacy protection and compliance with Twitter rules, we make sure that Twitter account handles and tweets are fully anonymized. We share tweets by their IDs, and we share a list of names written in Arabic and English as first names only.

Biases and Limitations Any biases found in our dataset are unintentional, and we do not intend to cause harm to any group or individual. In our study, we tried to remove biases in data collection by providing all forms of male and female description words. But, because Twitter is widely used in some regions (e.g., Gulf) and less used in other regions (e.g., Maghreb), we acknowledge that our statistics and results may be less accurate for some Arab countries in the real world. However, they give rough estimates about the actual presence of users from those countries on Twitter. The bias in our data, for example towards a particular gender, is unintentional and is a true representation of users on Twitter as obtained also from Twitter Ads. Gender label (male/female) is extracted from the data and might not be a true representative of the users' choice.

Further, we heavily depend on users' self-disclosure (first words only) which covers a small portion of Twitter users. Therefore, the statistics presented in our paper provides an estimate of the whole picture. In the future, we plan to consider better methods for data collection with greater diversity and coverage.

References

Ahmed Abdelali, Hamdy Mubarak, Younes Samih, Sabit Hassan, and Kareem Darwish. 2021. QADI: Arabic dialect identification in the wild. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kerstin Bischoff. 2012. We love rock'n'roll: analyzing and predicting friendship links in last. fm. In *Proceedings of the 4th Annual ACM Web Science Conference*, pages 47–56.

Bassem Bsir and Mounir Zrigui. 2018. Enhancing deep learning gender identification with gated recurrent units architecture in social text. *Computacion y Sistemas*, 22:757–766.

John D. Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1301–1309, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Joymallya Chakraborty, Suvodeep Majumder, and Tim Menzies. 2021. Bias in machine learning software: why? how? what to do? In *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, pages 429–440.

Xin Chen, Yu Wang, Eugene Agichtein, and Fusheng Wang. 2015. A comparative study of demographic attribute inference in twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, 1, pages 590–593, California, USA. AAAI.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J. Jansen, and Joni Salminen. 2020. A multi-platform Arabic news comment dataset for offensive language detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6203–6212, Marseille, France. European Language Resources Association.

Morgane Ciot, Morgan Sonderegger, and Derek Ruths. 2013. Gender inference of Twitter users in non-English contexts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1136–1145. Association for Computational Linguistics.

Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. 2011. Political polarization on twitter. In *Proceedings of the International AAAI Conference on Web and Social Media*, pages 89–96, Barcelona, Spain. AAAI.

Kareem Darwish and Hamdy Mubarak. 2016. Farasa: A new fast and accurate Arabic word segmenter. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, LREC'16, pages 1070–1074, Portorož, Slovenia. European Language Resources Association (ELRA).

Shereen ElSayed and Mona Farouk. 2020. Gender identification for egyptian arabic dialect in twitter using deep learning models. *Egyptian Informatics Journal*, 21(3):159–167.

Nizar Habash, Houda Bouamor, and Christine Chung. 2019. Automatic gender identification and reinflection in Arabic. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 155–165. Association for Computational Linguistics.

Sabit Hassan, Hamdy Mubarak, Ahmed Abdelali, and Kareem Darwish. 2021. ASAD: Arabic social media analytics and unDerstanding. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 113–118, Online. Association for Computational Linguistics.

Susan C Herring and Sharon Stoerger. 2014. Gender and (a) nonymity in computer-mediated communication. *The handbook of language, gender, and sexuality*, 2:567–586.

Shereen Hussein, Mona Farouk, and ElSayed Hemayed. 2019. Gender identification of egyptian dialect in twitter. *Egyptian Informatics Journal*, 20(2):109–116.

Gil Levi and Tal Hassner. 2015. Age and gender classification using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 34–42.

Shoushan Li, Bin Dai, Zhengxian Gong, and Guodong Zhou. 2016. Semi-supervised gender classification with joint textual and social modeling. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2092–2100. The COLING 2016 Organizing Committee.

W. Liu and D. Ruths. 2013. What's in a name? using first names as features for gender inference in twitter. In *AAAI Spring Symposium: Analyzing Microtext*.

Yaguang Liu, Lisa Singh, and Zeina Mneimneh. 2021. A comparative analysis of classic and deep learning models for inferring gender and age of twitter users. In *Proceedings of the International Conference on Deep Learning Theory and Applications*.

Shervin Malmasi. 2014. A data-driven approach to studying given names and their gender and ethnicity associations. In *Proceedings of the Australasian Language Technology Association Workshop*, pages 145–149.

Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444.

Alan Mislove, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela, and James Rosenquist. 2011. Understanding the demographics of twitter users. In *Proceedings of the International AAAI Conference on Web and Social Media*, Barcelona, Spain. AAAI.

Hamdy Mubarak and Kareem Darwish. 2014. Using twitter to collect a multi-dialectal corpus of arabic. In *Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP)*, pages 1–7.

Juergen Mueller and Gerd Stumme. 2016. Gender inference using statistical name characteristics in twitter. In *Proceedings of the The 3rd Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016*, pages 1–8.

Arjun Mukherjee and Bing Liu. 2010. Improving gender classification of blog authors. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 207–217. Association for Computational Linguistics.

Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents*, SMUC '10, page 37–44, New York, NY, USA. Association for Computing Machinery.

Shigeyuki Sakaki, Yasuhide Miura, Xiaojun Ma, Keigo Hattori, and Tomoko Ohkuma. 2014. Twitter user gender inference using combined analysis of text and image processing. In *Proceedings of the Third Workshop on Vision and Language*, pages 54–61, Dublin, Ireland. Association for Computational Linguistics.

Fadi Salem. 2017. Social media and the internet of things. *The Arab Social Media Report*.

Erhan Sezerer, Ozan Polatbilek, and Selma Tekir. 2019. A Turkish dataset for gender identification of Twitter users. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 203–207. Association for Computational Linguistics.

Juan Soler and Leo Wanner. 2016. A semi-supervised approach for gender identification. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, pages 1282–1287. European Language Resources Association (ELRA).

Tomoki Taniguchi, Shigeyuki Sakaki, Ryosuke Shigenaka, Yukihiro Tsuboshita, and Tomoko Ohkuma. 2015. A weighted combination of text and image classifiers for user gender inference. In *Proceedings of the Fourth Workshop on Vision and Language*, pages 87–93. Association for Computational Linguistics.

Tran Anh Tuan, Tien-Dung Cao, and Tram Truong-Huu. 2019. Dirac: A hybrid approach to customer demographics analysis for advertising campaigns. In *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 256–261.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Svitlana Volkova, Theresa Wilson, and David Yarowsky. 2013. Exploring demographic language variations to improve multilingual sentiment analysis in social media. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1815–1827. Association for Computational Linguistics.

Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. 2019. Demographic inference and representative population estimates from multilingual social media data. In *The world wide web conference*, pages 2056–2067.

Wajdi Zaghouani and Anis Charfi. 2018. Arap-tweet: A large multi-dialect Twitter corpus for gender, age and language variety identification. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Appendix

A Demo

Using the developed model, we also built a demo that takes a person's name written in Arabic or

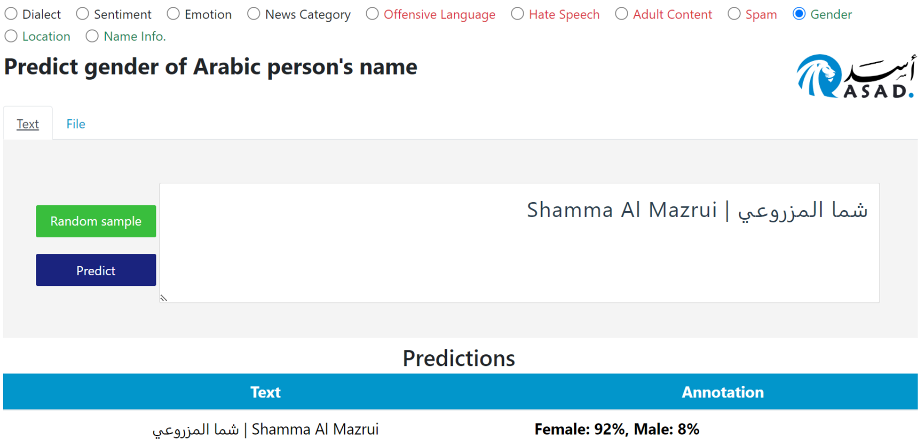


Figure 11: Demo interface for gender inference using our proposed models.

English and predicts a gender label with probabilities. The demo can be accessed using the link: `https://asad.qcri.org/demo` (part of ASAD tools (Hassan et al., 2021)). A screenshot of the demo is presented in Figure 11.