Investigating the Use of Machine Learning Algorithms in Detecting Gender of the Arabic Tweet Author

Emad AlSukhni
Computer Information Systems Department
Yarmouk University
Irbid, Jordan

Qasem Alequr
Computer Information Systems Department
Yarmouk University
Irbid, Jordan

***Abstract*—Twitter is one of the most popular social network sites on the Internet to share opinions and knowledge extensively. Many advertisers use these Tweets to collect some features and attributes of Tweeters to target specific groups of highly engaged people. Gender detection is a sub-field of sentiment analysis for extracting and predicting the gender of a Tweet author. In this paper, we aim to investigate the gender of Tweet authors using different classification mining techniques on Arabic language, such as Naïve Bayes (NB), Support vector machine (SVM), Naïve Bayes Multinomial (NBM), J48 decision tree, KNN. The results show that the NBM, SVM, and J48 classifiers can achieve accuracy above to 98%, by adding names of Tweet author as a feature. The results also show that the preprocessing approach has negative effect on the accuracy of gender detection. In nutshell, this study shows that the ability of using machine learning classifiers in detecting the gender of Arabic Tweet author.**

***Keywords*—Social Networking; Data Mining; Sentiment Analysis; Sentiment Classification; Gender Detection; Twitter**

I. INTRODUCTION

Nowadays, the existence of many social websites such as Twitter, Facebook, Myspace and blogs that make the internet a large repository of different type of data. These media allow different type of users from different cultures and languages to communicate and share their opinions, and experience with others.

These opinions represent many kind of information (political, sport, technology, etc.) that come from different sources. Such a large repository of data and information sparked the attention of researchers and companies to take advantage of this data for various purposes. Sentiment analysis or opinion mining is a field aims to extract or predict the polarity of people opinions in specific areas. This is considered as a challenging task for sentiment analysis.

Gender detection is a sub-field of sentiment analysis for extracting and predicting the gender of a Tweet author. Most researchers studied gender detection for Tweet writers in different language such as English, European and other languages. However, in Arabic language there is a few researchers studied gender detection. In this study, we focused on Arabic opinions Twitter. Some of these studies have been investigated only gender aspect as a core attribute which can be a good indicator of the author of Tweet as in [1, 2]. Other studies investigated not only gender but also other attributes such as age for example in [3, 4].

Twitter website is our interest of study. We analyzed Tweets, which are small texts that consist of maximum 140 characters each). The Tweets are classified based on their writers' gender into two classes male and female. Twitter is considered as one of the largest social media website widespread in the world that has a huge number of users and a large amount of data in different languages from different places. Many researchers have studied the users Tweets for many purposes such as extracting political opinions, spam detection, etc.

The importance of knowing the gender of the Tweet author may help governments to make their policies and help companies in handling commercial issues. Thus, many social websites collect some information about their users when register such as age, gender, location, and others.

The main purpose of this research is to detect the gender of the writer of an Arabic Tweet by classifying them into two classes (male or female). This problem can be considered as a binary text classification (TC) problem. In this study, we used five classifiers KNN, NB, NBM, SVM, and J48 decision tree to test their ability in predicting the gender of the Tweet author.

Research Questions

In this study we are trying to answer the following questions:

Q1: Are data mining techniques able to identify the gender of an Arabic Tweet author with a significant accuracy?

Q2: What are the best classifier(s) to predict the gender of a Tweet author?

Q3: What is the effect of preprocessing techniques on classification accuracy in gender detection domain?

Q4: What is the effect of adding an author name as a feature on classifiers accuracy in gender detection domain?

Q5: What is the effect of adding the number of words and average word length in the Tweet as features on classifiers accuracy in gender detection domain?

The reset of the paper is organized as following: Section II reviews the previous works. Section III presents the methodology. Section IV discusses the experimental results. Finally, section V presents conclusions and future work.

II. Literature Review

Many researchers have studied gender detection of the writer of Twitter website and other social media users in different languages. However, a few of them have investigated Arabic language Tweets. In this section we list the most important of these studies with their results.

A. Gender Detection research on Multi-language

Rao et al. in [3] studied many author attributes such as age, gender, regional, and political orientation to classify Twitter users based on each attribute. They investigated the use of SVM algorithm over a set of features to classify user attributes (e.g. age, gender, regional). They built a large dataset manually and also used crawling Tweets. Their task was to detect a gender of Tweet author whether male or female based on the content of the Tweet. Their goal was to show if the language has an impact on detecting attributes of the author based on his/her Tweets. They used three classification models, first sociolinguistic-feature model; which is based on finding a lot of keywords. They also studied the writing styles effects in Tweet author gender and age detection. They extracted a list of words to be used in SVM classifier. The second model they used the N-gram feature with SVM classifier. The results of detecting gender of author showed that the SVM classifier slightly outperformed than sociolinguistic-feature and also n-gram with 72.33%.

Burger et al. in [5] used statistical models to detect the gender of unknown users from different places with different language. They used a huge dataset of Tweets from Twitter website labeled with male and female. The experiments on this dataset were conducted to show the accuracy of these models. They used WEKA tool to apply machine learning algorithms such as SVM, NB, and balanced Winno2. The result of the first experiment showed that the NB accuracy is 67.0%, and balanced Winnow2 accuracy is 74.0%, and SVM accuracy is 71.8%.

Liu and Ruths in [6] studied the relationship between the first name and detecting the gender of users who write English Tweets, and how this can improve the accuracy of gender detection. They collected a dataset from Twitter website randomly. Then, they have introduced idea of knowing and labeling the gender of each Tweet throughout profile picture. To ensure the accuracy of labeling they used the Amazon mechanical truck, which approved that the accuracy gives a good indicator of labeling. The core classifier they used was SVM; they applied some of the features as methods to be used in SVM such as top keywords, key-top stems, key-top n-gram, that all differentiate between the two genders. The results of these methods achieved high accuracy with SVM as 87.1%.

Marquart et al. in [7] investigated how to increase the predictive way of detecting users with both age and gender attributes from different social media such as Twitter, blogs, reviews, and others based on English and Spanish languages. They have used three features; content-based feature related to frequency of words, and a stylistic feature related to readability and spelling issues. In the evaluation step, they used SVM as the core classifier and used two approaches first label-power set which transforms multi-label problem into single label problem. They also used chain classifiers; which determine the dependency between two classes, and determine which class is good predictor to the other. They showed that gender is a good feature to use in predicting age.

Modak and Mondal in [8] studied gender classification using machine learning techniques; such as Naïve Bayes, maximum entropy and decision tree. In their study, they focused on the name of a user rather than on content-based of texts to classify it into male and female written. They collected different names from the web and form a labeled corpus. In their study they tested the three classifiers. The results showed that maximum entropy has achieved the highest accuracy in comparison with other classifiers.

Deitrick et al. in [1] studied gender identification of Tweets author for the English language using simple stream-based neural network. They collected a huge amount of data from Twitter website and then divided it into three different feature groups, 1-gram, 2-gram, and other features. After that they split the dataset into two files; one file containing the training set used in modified balanced Winnow. While the other file, containing testing dataset, used to evaluate the balanced Winnow. The algorithm has achieved 82% accuracy using entire set of features and 97.89% precision.

Mikros in [9] investigated the authorship of attribution and author gender detection or author profiling using Greek blogs. Blogs were chosen because people can write their opinion on the blogs. He collected the corpus from different blogs. They focused on two features of text content; first classical stylometric features, which depends on the vocabulary "richness", word length, and word frequencies. As for the second type of features, they used modern features which depends on character bigram, word gram. The classifier they used is SVM which is suitable for binary classification problem. The results of their experiment showed that accuracy of gender identification achieved 82.6%

Koppel et al. [10] studied automatically detect the gender of formal document authorship. They focused in their research on classification based on the writing style. The research tested two assumptions based on some previous research. First they assumed that no difference in writing between man and woman in formal texts. But in the second assumption, there is a difference between the two genders where it can be used to classify text of unknown authors. They built a dataset named BNC (Britch national corpus), and applied machine learning algorithms. Finally, they proved that male differs in writing pronouns and some types of noun modifiers in comparison with females.

Sap et al. [11] derived predictive lexica for age and gender using regression and classification models from words based on social media websites such as Facebook and Twitter. They collected a dataset mainly from Facebook in English language. The lexica has achieved 91.9% accuracy in gender detection.

Volkova et al. in [12] introduced an analysis of important difference between male and female in subjective language in Twitter website using three languages English, Russian and Spanish. They studied how the gender of Tweets play an important role in the sentiment classification. They developed two corpuses one for gender detection while the second for sentiment analysis. In their research, they showed that included author gender as a feature can significantly improve subjectivity and polarity classification with all tested languages.

Ugheoke in [13] studied gender detection for Tweet author. He focused on Twitter website because of its popularity in the world. Some of features that helped to be as indicators of Tweet author gender such as user profile, behavior of Tweets user which is related to number of Tweets per day and the number of replies, the linguistic style, and the social network were used. he relies on the name of user profile that checked by US census data (American names) for manual labeling the dataset. They divided the texts into separated words then also used a stemmer to reduce the number of keywords. The experiments show that, SVM classifier has achieved 86.8% accuracy with no name inference, and 95.3% accuracy with associated author name.

B. Gender Detection on Arabic Language

Few researchers have studied gender detection for the Arabic language, here we show these studies as follows.

Estival, et al. in [14] developed an application which can detect author attributes or demographic information; such as name, age, gender, level of education, from Arabic emails. They used two email corpuses for Arabic and English languages. They used a questionnaire to check and analyze the personality of the author of email such as age, gender, name and level of education. Many machine learning classifiers were used in their experiments; such as SVM, KNN, and decision trees (J48) combined chi-square and information gain. In gender detection, the result showed that SVM without feature selection technique achieve high accuracy over other classifiers.

Alsmearat et al. in [2] investigated gender identification on Arabic articles using the Bag Of Words (BOW) feature in the selection phase. The proposed technique works by estimating each word frequency in each document. They collected their dataset from Arabic news websites manually. They also collected Modern Standard Arabic for both genders. To reduce the number of words they used light stemming technique. To reduce feature selection they applied four algorithms (correlation analysis, Principle Component Analysis, correlation-based subset evaluator, Relief F) after dividing the dataset into five versions to show if there are any relationships between words (stylistic differences). In the classification phase they used many classifiers such as Naïve Bayes, KNN, and SVM, and then applied them on the five versions separately. Results showed that NBM and SVM achieved high accuracy on the first version (original version). In other versions and by applying feature selection techniques, the results showed a negative impact compared with results of the original version due to lack of information. On the other hand, they studied the impact of stemming on the dataset (Arabic light stemmer), The results showed that no significant impact on the accuracy. But when they applied stemming with best feature selection technique (sub dataset) the results showed NBM achieved good results over other classifiers.

III. RESEARCH METHODOLOGY

This section describes the research methodology that consists of 4 steps as shown in Fig. 1: collect Tweets form Tweeter, text preprocessing, gender classification, and evaluate the result.

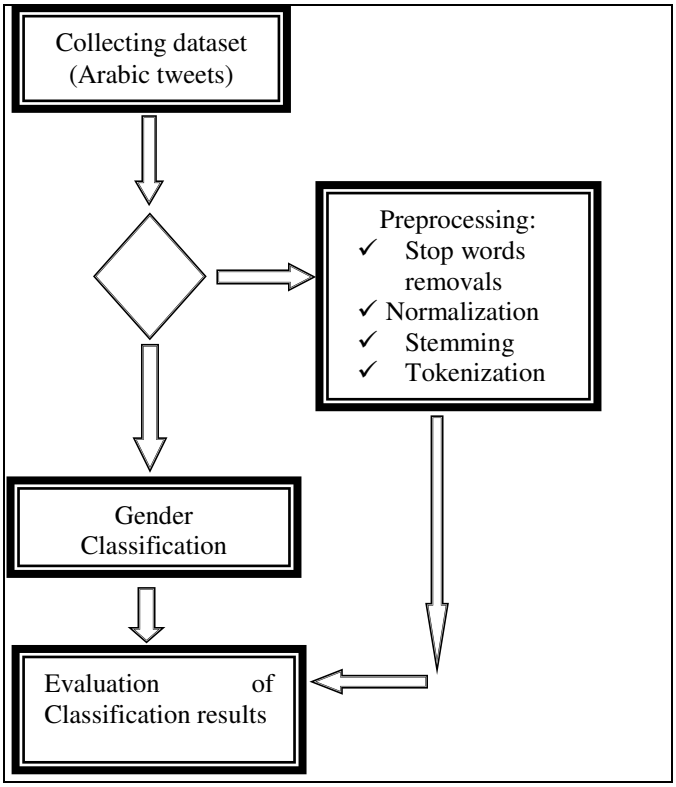

Fig. 1. Schematic overview of the methodology

In this research, we considered Twitter website as the target population to collect user Tweets from to be used in our experiments. There are many terminologies used in Twitter such as:

- Tweet: special text written by each user to represent his/her opinion about any topic.
- ReTweet: any user may republish any Tweet written from other users to be appeared in his/her profile.
- Followers: users follow individual user and see his/her Tweets.
- Hashtag (#): special symbol used to group any Tweets contain it such as (#sport) this Hashtag will group every Tweet that include this word.

In Twitter website when a user register, the user need just to enter username, email, and password to complete the registration, so no extra information could be used to determine the gender of the users. Figures 2 and 3 show two Twitter profiles one for a male user and another for a female user.

Fig. 2. Female Twitter Profile

Fig. 3. Male Twitter Profile

A. *Dataset of Tweets*

We collected large number of Tweets from Twitter website that exceeds 8000 Tweets in Arabic language, mainly in Jordanian dialects. We selected them from different domains for different users. The dataset consists 4017 Tweets written by males and 4017 Tweets written by females. Each tuple in our dataset consist of the following attributes:

1) The Tweets
2) Name of a Tweet author
3) Gender
4) Tweet Average length of word in
5) Number of words in the Tweet.

During the collecting phase, we considered some points to determine the gender of the Tweets' writers such as:

- Profile user name was used as a good indicator (User names are written either in Arabic or in English language) such as (‘علي’, ‘Ali’), and the profile picture to identify the correct gender (male or female).
- We used a lot of (Hashtages) to search about user profiles (‘رؤيا#’,’بس_بقول#’,’الجامعة الاردنية#’, etc.) and then we visit the followers to collect other Tweets.
- We focus on Tweets, which are written by original users and excluded Tweets that were reTweet it (which is wrote by other users).
- We excluded Tweets that are written in newspaper articles or television reports.
- During building the data collection, we took approximately 50 Tweets for each author.
- Tweets have English words are excluded.
- We classify the dataset into male and female manually.

Sample of male Tweets with their authors are shown in table I.

TABLE I. SAMPLE OF MALE TWEETS

Profile User Name	Arabic Tweets
Fadi ‘فادي ‘	“يعني لو درست لاب كان نجحت بس أنا متخلف”
Mohammed ‘محمد ‘	“ثلاث ساعات نايم من مبارح لا ومش عارف انام كمان”
Khaled ‘خالد’	“مفكرينا لاجئين هَوْن”
Omar ‘عمر’	"صباح الخير ياعرب"
Abd-Rahman ‘عبدالرحمن’	"البيت في الشتوية عبارة عن مكان ممل للغاية واحلى اشي فيه الأكل والنوم فقط لا غير"
Shaheen ‘شاهين’	"حد عنده راس للبيع. بدي راس ثاني"
Bahaa ‘بهاء’	"الدوام بكره زهق وملل ...بداية اسبوع ممتعه"
Mammon ‘مأمون’	"الجو بنعس كتير"
Tareq ‘طارق’	" هاي شكل واحدة تنتحر"
Moaied ‘مؤيد’	"الواحد بضل يتحمض للمخميس وبس ييجي الخميس بطلع مثل يوم الثلاثاء"
Salama ‘سلامه’	" طول ما الله موجود ! فالامل ابدا ما حيموت"
Nabil ‘نبيل’	" السواقة في هيك جو خرافية مش شايف متر قدامي"
Hasan ‘حسن’	" جمال العيون في النظرة مش في اللون"
Faris ‘فارس’	" نصيحة اليوم حبوا بعض وفي كل وقت من الاوقات بلابد ان تكرهو بعض"
Mosab ‘مصعب’	" خذو الحكمة من افواه المغردين"
Sameer ‘سمير’	" السناب عندي عبارة عن تغطية مباشرة لكل شوارع جدة و هي بتغرق"
Ahmed ‘احمد’	" حتى لو برد و تجمدنا و جلدنا بس برضه الشتى احلى من الصيف و ناموسه و قرفه"

Sample of female Tweets with their authors are shown in Table II.

TABLE II. SAMPLE OF FEMALE TWEETS

Profile User Name	Arabic Tweets
Aseel ‘ اسيل ‘	“اقنع امي انو انا ما باكل مقلوبه”
Hanaa ‘ هناء’	“ أكتر شخص مسالم وهادي بالحياة بس بحسك دايماً مكتئبة من الحياة”
Marh ‘مرح’	"مقدار السعادة انه لهلا سهرانة بدون ما اتذكر بكرة لازم اصحى على 6"
Salsabeel ‘سلسبيل’	“ ما أشنع البنت الي بتسوي حالها مهتمة ومحروقة عأشياء الرياضة خلص ماشي انت حلم كل شاب عربي بس اسكتي”
Sammer ‘سمر’	"بس عشان نكون واقعيين، الحياة بدها شخصية باردة ومكبرة عقلها"
Haneen ‘حنين’	"صباح الباص إلي راح علي"

Rand ‘رند’	"نفسي اغسل أموال بهل بلد عشان افيدها بمشاريعي"
Tasneem ‘تسنيم’	”زهقانة حالي“
Wasen ‘وسن’	"صباحكم جميل..بكلمات أخرى .. صباح القرود"
Araam ‘غرام’	"مش عارف ليش عندي شعور انه اليوم رح يكون احلى يوم"
Yara ‘يارا’	أسوأ اشي انه تحس حالك بتمشي بسرعة “ بتصير تضيّع بالوقت زي المحروم”
Amani ‘أماني’	" لا تجبر حالك على شي مسببلك قلق ، خلي قاعدتك في الحياة الشي اللي ما بسعدني ما بلزمني"
Randa ‘رندا’	" يلا حبايبي اللي مش عاجبه انفولو بسرعة"
Losi ‘لوسي’	" علمتني الحياه انو ما اصدق غير اللي بشوفو بعيني او بسمعو باذني غير هيك لا لانو هالعالم صارت تعشق الهشت عشق"
Ronza ‘رونزا’	" الإشي الوحيد الحلو بحياتي حالياً هو الأكل"
Rawan ‘روان’	" الجاكيتات الجوخ الطويله ، فقط للرجل صاحب القامه الطويله غير هيك عبث"
Jodi ‘جودي’	" عد ما رجعت من عنده و اتطلعوا ع بعض، لم كل صحابه و ما خلّى كلمة عليها و بلش يتسلى عليها"

B. Limitations and Assumptions

- The dataset represent the Jordanian dialects of Arabic language.
- Some users may use fake profile name that does not refer to the gender, such as a male user may use a female name.
- Collected profiles of famous users, newspapers, or any profile that uses the Arabic standard Arabic language are not considered.

C. The Preprocessing Phase

In this research, we study if the preprocessing stage has any impact on the quality of the results. Preprocessing stage consists of two major steps: 1) removing stop words and 2) stemming.

According to [18], using the Weka tool can make the preprocessing step by applying Saad light stemmer which performs the following things:

1- Normalized words
   - o Remove diacritics
   - o Replace ا أ آ with ا
   - o Replace ة with ه
   - o Replace ى with ي
2- Stem prefixes
   - o Remove Prefixes:كالـ, فالـ الـ, والـ.
3- Stem suffixes
   - o Remove Suffixes:ها , ان, ات, ون, ين.

D. Classifications

The dataset tuples are classified into two classes; male and female. We applied supervised machine learning classifiers to study the accuracy for each of them in detecting the author gender. Basically, classification is an approach aims to predict a class label that is unknown. Classification consists of two main stages: it builds the model from the training dataset, and then making a prediction.

In our research, we have used five data mining classifiers as listed below:

1) Key Nearest Neighbor (KNN): By using similarity and dissimilarity measures, the classifier works to estimate the distance between unlabeled documents and all documents in the training set as in [15]. For instance, if we want to classify the document x, it calculate the distance between x and documents in training set then after finding the k nearest documents to x, the classifier assign the document x to the class that have the large number of documents near of x. The Euclidean distance is used as a conventional method for measuring distance between two documents, d_1 (w_{11}, w_{12},…, w_{1n}) and d_2 (w_{21}, w_{22},…, w_{2n}):

$$E(d_1, d_2) = \sqrt{\sum_{i=0}^{n}(W_{2i} - W_{1i})^2} \quad \text{.....(1)}$$

2) Naïve Bayes (NB): Worked based on the probability theorem of conditional probability, mainly it is used for binary classification. In this classifier, the features of each document do not depend on the other features to predict the class, In the below the equation used estimate the probability of class.

$$P(Ci \mid X) = (P(X \mid C_i) P(C_i)) / P(X) \text{.....(2)}$$

3) Naive Bayes Multinomial (NBM) : The multinomial model of naïve Bayesian classification algorithm captures the word frequency information in document. NBM take into account the word frequency of each word as in [17].

4) Support vector machine (SVM): This algorithm works based on structural risk minimization principle from the computational learning theory. It divides the training set into two groups then try to find the hyperplane that is far from two groups as in [15]. Finding the optimal hyperplane based on the following formula:

$$F(X) = B_0 + B^T X \quad \text{..... (3).}$$

Where (B) is weight vector and (B_0) is a bias.

The closest training documents to hyperplane are called support vectors. Then the distance x and the hyperplane is estimated based on the below equation:

$$\text{Distance} = | B_0 + B^T X| / (\|B_0\|) \quad \text{..... (4).}$$

And then find the margins (distance) between the document (x) and the hyperplane from both sides of two groups, the margin that is represented by

$$M = 2/|B| \text{..... (5).}$$

According to [15], SVM has several advantages over other techniques, such as it is robust in high dimensional spaces, any feature is important, they are robust when there is a sparsely of samples.

5) Decision tree: the decision tree classifier works by creating a classification tree, where each non-leaf node corresponds to a feature name and its children corresponds to a feature value. The Decision Tree classifier is a supervised machine learning approach that often used in a text classification domain. It requires two sets: a training set and test set. the Decision Tree Classifier creates a binary tree where the child nodes are instances of the classifier. In other words, this algorithm partitions the training set from the bottom to the top and then it picks up one attribute each time

and then the most information gain attribute is used to split the tree.

IV. Experiment and Evaluation

A. Performance Measures

Yang and Liu in [19] lists many of measurements to test the performance of classifiers such as:

1) True Positive (TP): If the instance is a positive and classified as positive.

2) False Negative (FN): If the instance is a positive and classified as negative.

3) True Negative (TN): If the instance is negative and it is classified as negative.

4) False Positive (FP): If the instance is negative but it is classified as positive.

Accuracy: It is the ability to predict categorical class labels. This is the simplest scoring measure. It calculates the proportion of the classified instances correctly:

Accuracy = (TP + TN) / (TP+TN+FP+FN) …….(6)

Sensitivity/Recall: Sensitivity is the proportion of the actual positives, which are correctly identified as positives by the classifier. It is also called true positive rate.

Sensitivity = TP / (TP +FN) …….(7)

Precision: it is a measure of the retrieved instances that are relevant.

Precision= TP/ (TP+FP) ……. (8)

B. Experiment Results and discussion

We evaluate the performance of the selected classifiers in classifying the gender of the Arabic Tweets' author. The rest of this section describes the results of experiments that have been designed and conducted to answer the research questions of this study. Most of the research use cross validation technique and splitting percentage in their classification experiments. In this study, we use cross-validation (10 Folds) and splitting percentage (66% train, 33% test). Because of the limited space of this Paper, we include all the cross-validation based results and the summery of splitting percentage based results.

Experiment 1: Evaluation classifiers without preprocessing.

In this experiment we test the performance of the selected classifiers in classify the gender of the Arabic Tweets' author without applying preprocessing step. As shown in Table III, the NBM classifier outweighs to other classifiers so as to achieve a better of accuracy of (62.49%) and recall (63%) to 5021 instances that are correctly classified manually (2532 instances for females and 2489 for males) as shown in Table III. The SVM classifier is improved slightly the precision (64%) compared with other classifiers. We notice that the NBM classifier outperforms the other classifier in correctly classified female instances than male instances.

TABLE III. Rate of Classifiers Performance without Preprocessing Step

Classifier	Accuracy	Recall	Precision
KNN	54.00%	0.393	0.557
Naïve bayes (NB)	57.39%	0.554	0.577
J48 (decision tree)	57.91%	0.488	0.597
SVM	61.63%	0.529	0.641
NBM	62.49%	0.630	0.624

Experiment 2: Evaluate the effect of preprocessing classification accuracy in gender detection domain.

In this experiment, we test the performance of the selected classifiers in classifying the gender of Arabic Tweets' author with applying the preprocessing step. As shown in Table IV and Figure 4, the NBM classifier outperforms the other classifiers. It achieved promising results with 61.27% accuracy in which the total number of correctly classified instances was 4923 (including 2507 instances of them for female and 2416 for male). Based on the precision measure SVM classifier achieved good result with 61%. We notice that NBM classifier outperforms the other classifier in correctly classified female instances than male instances. In another hand; the accuracy of KNN classifier is the lowest result with 53.43%.

TABLE IV. Classifiers Performance with Preprocessing

Classifier	Accuracy	Recall	Precision
KNN	53.43%	0.407	0.546
J48 (decision tree)	57.09%	0.494	0.584
Naïve bayes (NB)	57.55%	0.575	0.576
SVM	59.99%	0.539	0.614
NBM	61.27%	0.624	0.610

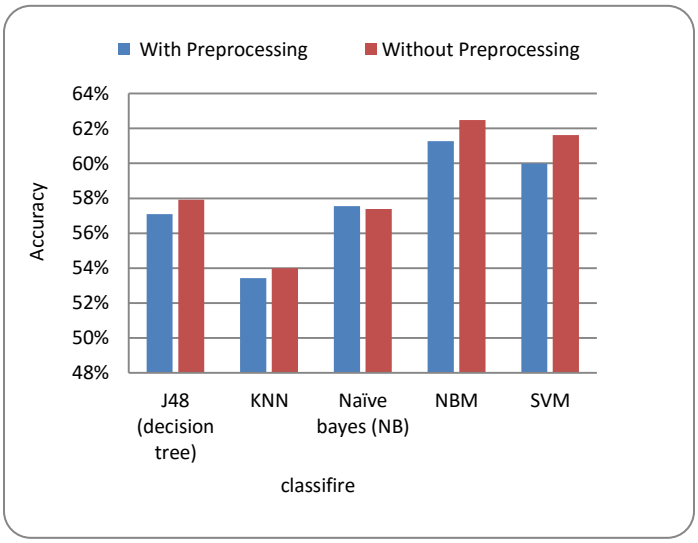

Fig. 4. Accuracy Results of Classifiers with and without Preprocessing

By comparing the accuracy of the classifiers, we conclude that the preprocessing step has a small negative effect on the accuracy of all classifiers. This results answer the third research question Q3.

Experiment 3: Evaluate the effect of adding an author name as a feature on classifiers accuracy in gender detection domain.

This experiment is designed to evaluate the effect of the author name feature on the performance of the selected classifiers which classifying the gender of Arabic Tweets' author with and without preprocessing. We add the name of the Tweet's author as a new feature in to the dataset to test the effect of this feature on the accuracy of the classifiers.

The result of this experiment shows the accuracy of detect the gender of the Tweet's author is significantly improved by adding Tweet's author name as a feature in the dataset as shown in the Figure 5. Moreover, we notice the same significant effect is achieved with and without applying preprocessing step. It is also clear that the accuracy of the top three classifiers become convergent.

TABLE V. EVALUATION OF CLASSIFIERS ACCURACY (WITH NEW TEXT FEATURES AND AUTHOR NAME ADDED)

Classifier	Accuracy (With Author Name added) without Preprocessing	Accuracy (With Author Name added) with Preprocessing	Accuracy Improvement ratio
Naïve bayes (NB)	77.29%	74.96%	25.75%
KNN	91.39%	83.67%	40.91%
NBM	98.49%	98.19%	36.55%
SVM	98.69%	98.25%	37.55%
J48 (decision tree)	98.69%	98.29%	41.32%

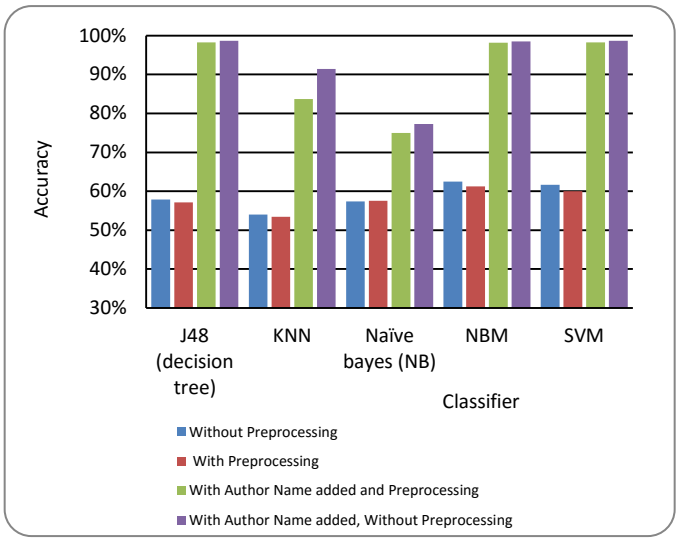

Fig. 5. Classifiers Accuracy (With Author Name added)

It is noticeable that after adding the author name feature to the dataset, the accuracy of the J48 classifier become very close to the accuracies of NBM and SVM classifiers. So, we can conclude that adding the author name feature to the dataset has the significant effect on the J48 classifier accuracy. Table V shows that the J48 and SVM classifiers outperform other classifiers with 98.69% accuracy either applying preprocessing or without applying. In this experiment the total number of correctly classified instances is 7929 by J48 (including 3994 female Tweets and 3935 male Tweets) and total number of correctly classified instances is 7929 by SVM (including 3990 female Tweets and 3939 male Tweets). We also notice that J48 and SVM classifiers have correctly classified both male and female written Tweets with the same accuracy.

According to Ugheoke T in [13], there is a relationship between the Tweets written in American English language and the name of Tweet's author that has an enhancement on accuracy of the gender detection. Thus, We can conclude from the result of this experiment that the effect of adding author names of Arabic language Tweets has the similar effect of adding author names of English language Tweets on the gender detection.

Experiment 4: Evaluate the effect of adding the number of words and average word length in the Tweet as features on classifiers accuracy in gender detection domain.

In order to get more improvement on classifiers accuracy, we added two features into our dataset that includes Tweet's author name. These two features are average word length and the number of words in the Tweet. So, this experiment is designed to evaluate the effect of the average word length and the number of words features on the accuracies of the classifiers into which classifying the gender of Arabic Tweets' author, Table VI gives such results.

TABLE VI. EVALUATION RESULTS OF CLASSIFIERS ACURACY WITH ADDING NAME OF TWEET'S AUTHOR AND THE NUMBER OF WORDS AND LENGTH OF WORD IN TWEET WITHOUT PREPROCESSING

Classifier	without preprocessing			preprocessing		
	Accuracy	Recall	Precision	Accuracy	Recall	Precision
KNN	69.60%	0.719	0.687	67.97%	0.722	0.666
Naïve bayes (NB)	73.38%	0.740	0.731	72.91%	0.733	0.727
NBM	99.06%	0.985	0.996	98.45%	0.977	0.992
SVM	99.35%	0.995	0.993	98.97%	0.987	0.992
J48 (decision tree)	99.50%	0.996	0.994	98.86%	0.989	0.988

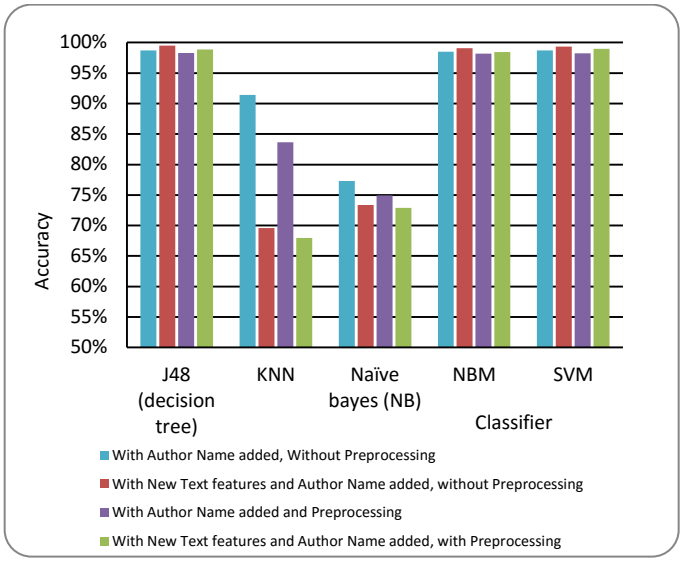

Fig. 6. Classifiers Accuracy adding the number of words and average word length in the Tweet as features

In this experiment, we applied the selected classifiers to study the impact of adding the name of the Tweet's author, the number of words and the length of word to the dataset without applying the preprocessing step. As shown in Table VI and Figure 4, the J48 classifier outperforms the other classifiers and it achieves the highest accuracy with 99.50% which the total number of 7994 correctly classified instances is 7994 (including 4001 for female and 3993 for male), with recall of 0.996%, but in precision we notice that the NBM classifier achieves slightly better results with 0.996%. Although the J48 classifier outperforms the other classifiers, the accuracies of SVM and NBM are very close to the accuracy of J48. We conclude adding the number of words and the average word length has a minor positive effect on top thee classifiers J48, NBM and SVM. KNN. On other hand, adding the number of words and the average word length has a negative effect on KNN and Naïve bayes classifiers as shown in Table 6 and Figure 6.

C. *Summary of classifiers Cross-Validation based accuracy results*

Table VII and Figure 7 present the summary of the effect of each studied feature on the classifiers accuracy. The results show that the accuracy of the classifiers without preprocessing are vary from 57% to 62%. The accuracy of all classifiers slightly decreased with applying preprocessing. The classifiers accuracy significantly increased by adding author names as a feature. The accuracy of the three beast classifiers slightly increased by adding two text features (number of words and average word length in the Tweet).

Form the table VII and Figure 7, we can answer all the research questions especially the first two questions; the answer of the first question is: yes, data mining techniques is able to identify the gender of an Arabic Tweet author because three classifiers got 99% accuracy. Regarding the second question answer, the three classifies (J48 (decision tree), NBM and SVM) have the best results in classification the Tweet author gender.

TABLE VII. SUMMARY OF CLASSIFIERS CROSS-VALIDATION BASED ACCURACY

Classifier	Without preprocessing	With Preprocessing	Adding Author Name	Adding New Text features
J48 (decision tree)	57.91%	57.09%	98.69%	99.50%
KNN	54.00%	53.43%	91.39%	69.60%
Naïve bayes (NB)	57.39%	57.55%	77.29%	73.38%
NBM	62.49%	61.27%	98.49%	99.06%
SVM	61.63%	59.99%	98.69%	99.35%

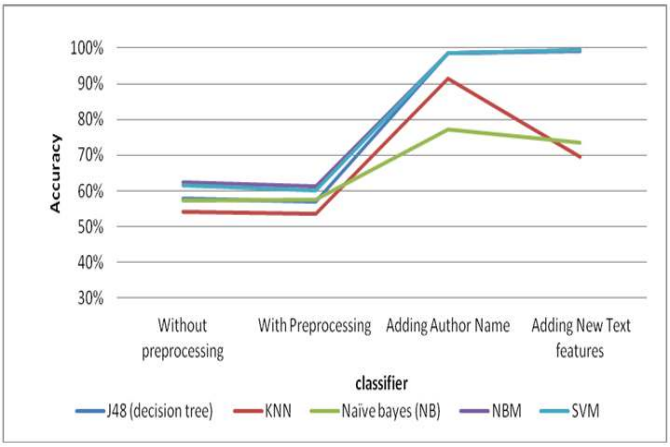

Fig. 7. Summary of classifiers Cross-Validation based Accuracy

D. *Summary of classifiers Splitting Percentage based accuracy results*

To give more creditability to our results, we rerun the experiments with Splitting Percentage of (66%).

In nutshell, we evaluate the performance of the selected classifiers in classifying the gender of Arabic Tweets' author based on the specific Splitting Percentage. From the Splitting Percentage based results, we can get the same above mentioned conclusions which give us more confidence in our findings.

TABLE VIII. SUMMARY OF CLASSIFIERS SPLITTING PERCENTAGE BASED ACCURACY

Classifier	without preproc-essing	with preproc-essing	Author Name added	With New Text features and Author Name added
KNN	53.69%	53.18%	77.85%	67.24%
J48 (decision tree)	57.79%	56.00%	88.17%	74.45%
Naïve bayes (NB)	57.39%	58.23%	98.64%	99.52%
NBM	60.72%	59.91%	98.79%	99.93%
SVM	60.68%	59.22%	98.79%	99.59%

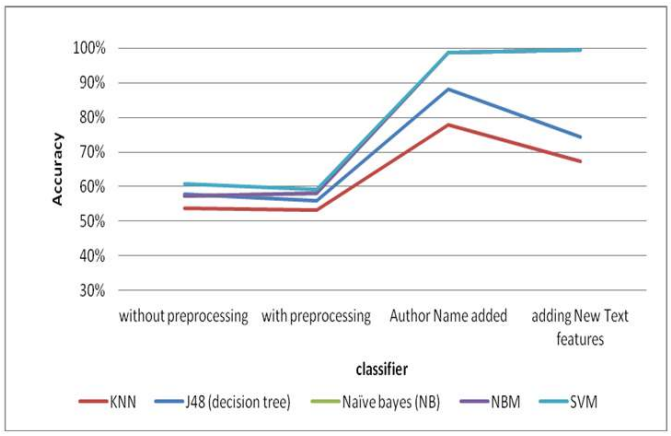

Fig. 8. Summary of classifiers Splitting Percentage based Accuracy

Table VIII and Figure 8 shows the results of many experiments. Let us start with the results of the classifiers without preprocessing on the dataset, the NBM classifier outperforms the other classifiers and achieved promising results with accuracy of 60.72% in which the total number of correctly classified instances was 1659 (including 854 instance of them for female and 805 for male). We notice that the NBM classifier outperforms the other classifier in correctly classifying female instances than male instances. Table 8 shows that KNN is the lowest classifier result with accuracy of 53.69%.

Table VIII, also shows the results of the experiment with apply preprocessing on the dataset. Notice that, the NBM classifier outperforms the other classifiers, into which the accuracy of 59.91% in which the total number of correctly classified instances 1637 (including 849 instance of them for female and 788 for male). We notice that the NBM classifier outperforms the other classifier in the correctly classified female instances than male instances. On the other hand, KNN is the lowest classifier with accuracy of 53.18%.

We test effect of adding the Tweet's author name to the dataset with and without preprocessing. As shown in Table 8 without preprocessing, the J48, SVM classifiers outperform other classifiers with 98.79% for both accuracy in which the total number of correctly classified instance is 2699 for J48 (including 1348 for female and 1351 for male) for both classifiers. We also notice that the J48 and SVM classifiers outperform the other classifier in correctly classifying male instances than female instances. On the other hand, NB classifier got the lowest accuracy of 77.85%.

In the last experiment, we add the number of words and the average word length to each Tweet in the dataset, which already has names of Tweet's authors. The last column in Table 8 shows the results this experiment. The results show that the NBM classifier outperforms the other classifiers and it achieves better results in accuracy with 99.93%, in which the total number of correctly classified instance is 2703 (including 1333 for female and 1370 for male). On other hand, KNN classifier has the lowest accuracy which is 67.24%.

V. Conclusion and Future Work

This research aims to test the ability of many machine learning classifiers, such as J48, KNN, Naïve Bayes, NBM and SVM in detecting the gender of Arabic Tweet's writers. We collect the dataset that contains 4017 Tweets as a first step for the purpose of this study. The results show that the classifiers can be used to detect the gender of the Tweet's author. We also test the effect of preprocessing on the accuracy of the classifiers that were under testing. The results show a negative effect of preprocessing on the accuracy of all classifiers. Moreover, this study tests the effect of adding author names and word features on the accuracy of the classifiers that were under testing. The results show significant positive effect of adding the names of Tweets' author on the accuracy of all classifiers, the accuracy of J48, NBM and SVM classifiers achieved above 98%. Overall, the results of all classifiers in recall and precision measures are significantly improved. The results also show that there is a slightly positive effect result in adding the number of words and the average length of Tweet's words on the accuracy of the J48, NBM and SVM classifiers. On other hand, the results shows a significant negative effect on the accuracy of KNN and Naïve Bayes.

Overall results demonstrate that it is possible to use machine learning classifiers to detect the gender of Arabic Tweet's author. We got the same findings with both cross-validation and splitting percentage (66%) on preparing the dataset for our experiments. During the experiments, we notice that NBM, J48 and SVM classifiers achieve the best results in ability to classify female instances more than male instances. This leads to conclude that the possibility to detect female Tweets written in more accurate than male Tweets.

In future, we plan to add different dialects of other Arab countries and also collect Tweets written in standard Arabic language in our dataset. We also planning to study gender detection of Tweet's author who uses both English and Arabic languages in the same Tweet.

References

[1] Deitrick, W., Miller, Z., Valyou, B., Dickinson, B., Munson, T., & Hu, W.,"Gender identification on Twitter using the modified balanced winnow". Communications and Network, 4(3), pp 189–195, 2012.

[2] Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R., "An extensive study of the Bag-of-Words approach for gender identification of Arabic articles". In proceeding of Computer Systems and Applications (AICCSA), 2014 IEEE/ACS 11th International Conference, IEEE, pp. 601-608. 2014.

[3] Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M., "Classifying latent user attributes in Twitter". In Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM, pp. 37-44, 2010.

[4] Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M. F., Davalos, S., Teredesai, A., & De Cock, M., "Age and gender identification in social media". Proceedings of CLEF 2014 Evaluation Labs, 2014.

[5] Burger, J. D., Henderson, J., Kim, G., & Zarrella, G.,"Discriminating gender on Twitter". In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguisticsl, pp. 1301-1309, 2011.

[6] Liu, W, and Ruths, D., "What's in a name? using first names as features for gender inference in Twitter". In Symposium on Analyzing Microtext. 2013.

[7] Marquardt, J., Farnadi, G., Vasudevan, G., Moens, M. F., Davalos, S., Teredesai, A., & De Cock, M., "Age and gender identification in social media". Proceedings of CLEF 2014, 2014.

[8] Modak, S, and Mondal, A., "A Comparative study of Classifiers Performance for Gender Classification", IJIRCCE, 2(5), pp 4214-4222, 2014.

[9] Mikros, G. K.., "Authorship Attribution and Gender Identification in Greek Blogs", Methods and Applications of Quantitative Linguistics, pp. 21–32, 2012.

[10] Koppel, M., Argamon, S., & Shimoni, A. R., "Automatically categorizing written texts by author gender". Literary and Linguistic Computing, 17(4), pp. 401-412, 2009.

[11] Sap, M., Park, G., Eichstaedt, J., Kern, M., Ungar, L., & Schwartz, H. A., "Developing age and gender predictive lexica over social media". In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP, pp. 1146-1151, 2014.

[12] Volkova, S., Wilson, T., & Yarowsky, D., "Exploring Demographic Language Variations to Improve Multilingual Sentiment Analysis in Social Media". In Proceedings of EMNLP, pp. 1815-1827, 2013.

[13] Ugheoke, T. O., "Detecting the Gender of a Tweet Sender", pp 1-60, 2014.

[14] Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B., "TAT: an author profiling tool with application to Arabic emails". In Proceedings of the Australasian Language Technology Workshop, pp. 21-30, 2007.

[15] Gharib, T. F., Habib, M. B., & Fayed, Z. T., "Arabic Text Classification Using Support Vector Machines". IJ Comput. Appl., 16(4), pp192-199, 2009.

[16] http://docs.opencv.org/2.4/doc/tutorials/ml/introduction_to_svm/introduction_to_svm.html.accessd-28/12/2015. Accessed in March 10th,2016.

[17] Saad, M. K.., "The impact of text preprocessing and term weighting on Arabic text classification", Doctoral dissertation, The Islamic University-Gaza, 2010.

[18] Motaz K. Saad and Wesam Ashour,"Arabic Morphological Tools for Text Mining", 6th ArchEng International Symposiums, EEECS'10 the 6th International Symposium on Electrical and Electronics Engineering and Computer Science, European University of Lefke, Cyprus, pp. 112-117, 2010.

[19] Yang, Y., and Liu, X., "A re-examination of text categorization methods". In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval. pp 42-49, 1999.

[20] Nguyen, D. P., Trieschnigg, R. B., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., & de Jong, F. M. G., "Why gender and age prediction from Tweets is hard: Lessons from a crowdsourcing experiment". In Proceedings of COLING, 2014.

AUTHORS PROFILE

Emad Mahmoud Alsukhni obtained his PhD in from Ottawa University in Canada in (2011), he obtained his Masters' degree in Computer and Information Science from Yarmouk University in (2003), and obtained his Bachelor degree in Computer Science from Yarmouk University in (2003). Alsukhni is an assistant professor at the Faculty of Information Technology and Computer Science at Yarmouk University in Jordan. Alsukhni research interests include Computer Networks, Information Retrieval, Sentiment analysis and Opinion Mining, and Data Mining. He is the author of several publications on these topics.

Qasem Ibrahim Alequr obtained his Master degree in Computer Information Systems from Yarmouk University in Jordan in (2016), he obtained his Bachelor degree in in Computer Information Systems from Yarmouk University in (2008), and . Alequr research interests include Sentiment analysis and Opinion Mining, and Data Mininge