Egyptian Informatics Journal 20 (2019) 109–116

Contents lists available at ScienceDirect

Egyptian Informatics Journal

journal homepage: www.sciencedirect.com

Gender identification of egyptian dialect in twitter

Shereen Hussein [a,*], Mona Farouk [a], ElSayed Hemayed [a,b]

[a] Cairo University, Faculty of Engineering, 1 University Street, Giza 12613, Egypt
[b] Zewail City of Science and Technology, Ahmed Zewail Road, October Gardens, 6th of October City, Giza, Egypt

ARTICLE INFO

Article history:
Received 6 October 2018
Accepted 19 December 2018
Available online 24 January 2019

Keywords:
Text classification
Egyptian Arabic
Gender identification
Author profiling
Gender annotated dataset

ABSTRACT

Despite the widespread of social media among all age groups in Arabic countries, the research directed towards Author Profiling (AP) is still in its early stages. This paper provides an Egyptian Dialect Gender Annotated Dataset (EDGAD) obtained from Twitter as well as a proposed text classification solution for the Gender Identification (GI) problem. The dataset consists of 70,000 tweets per gender. In text classification, a Mixed Feature Vector (MFV) with different stylometric and Egyptian Arabic Dialect (EAD) language-specific features is proposed, in addition to N-Gram Feature Vector (NFV). Ensemble weighted average is applied to the Random Forest (RF) with MFV and Logistic Regression (LR) with NFV. The achieved gender identification accuracy is 87.6%.

1. Introduction

We are now living in the era of social media, which has become an essential ingredient in our daily routine. Over two billion users around the globe are obsessed with social media.[1] While some use social media to interact with other people, some use it with fake accounts to commit illegal acts. GI for the account owners is needed for both types of users. For real accounts, GI is required for some digital market advertising where you can personalize the products and advertisements. On the other hand, for fake accounts, it is required to identify the real gender of the account owner. There are many types of crimes that can be committed with fake accounts such as cyber harassment, bullying, stalking, warfare and terrorism. The Halt Abuse Organization published statistics for cyber harassment indicating that the harassers, gender is 40% males, 30% females, while 30% are unknown.[2] The Middle East is strongly facing cyber-crimes especially with the recent widespread of social media.

Gender identification of authors has, thus, become more important with the increase of fake identities in social media accounts. Arabic language writers are increasing in social media. Taking into consideration the unique nature of the Arabic language and the work done on GI in general, Arabic in specific, GI methodologies and models need to be studied.

The work done in Arabic, especially dialectal Arabic, is still in its infancy. There is a lack of resources and lexicons, also the average accuracy achieved for dialectal Arabic needs improvement. This work builds a new gender annotated dataset EDGAD, to be available for future research. The paper at hand also proposes a complete GI model in addition to an engineered feature vector with integrates different EAD language-related features with N-gram FV achieving 87.6% accuracy.

The paper is organized as follows: Section 2 presents a brief of the related work done in gender identification from text in Arabic in specific and in other languages in general. Section 3 introduces the new collected dataset EDGAD, and also provides relevant statistics. Section 4 elaborates details the method used for gender identification and the MFV implemented. Section 5 presents the various experiments done and the corresponding results on EDGAD and PAN author profiling dataset for 2017 PAN-AP'17.[3]

* Corresponding author at: 44, Street 12 Repeated, ElMaadi Sarayat, ElMaadi, Cairo, Egypt.
E-mail addresses: shereenhussein@rocketmail.com (S. Hussein), mona_farouk@eng.cu.edu.eg (M. Farouk), hemayed@ieee.org (E. Hemayed).

Peer review under responsibility of Faculty of Computers and Information, Cairo University.

Production and hosting by Elsevier

[1] https://www.emarketer.com/Article/eMarketer-Updates-Worldwide-Social-Network-User-Figures/1016178.

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

[3] https://pan.webis.de/clef17/pan17-web/author-profiling.html.

https://doi.org/10.1016/j.eij.2018.12.002
1110-8665/

2. Literature review

This section summarizes the literature related to gender identification of authors. There is a gap between the amount of research done in English social media and that done in Arabic, namely in the areas of sentiment analysis, language understanding and gender identification.

Recently, a number of papers started addressing dialectical Arabic as those participating in PAN at CLEF [1,2]. Author profiling in the before mentioned dataset consists of four Arabic dialects including Egyptian dialect. Tellez et al. [3] applied the GI problem on a general classifier, MicroTC, which works regardless of the domain and language particularity [4], they achieved an accuracy of 83.78% on PAN-AP'17 Arabic dataset and 79.78% on PAN-AP' 18. Another approach of using N-grams for characters, words or POS tags is represented in a variety of research work for GI problem. Däniken et al. [5] used word unigrams, character 1-5 grams and emoji unigrams with LR and achieved 77.42% for English, 74.64% for Spanish, and 73.20% for Arabic tweets. However, AlRifai et al. [6] used the character 1-7 grams and added hashtags, links, mentions and the lengthened words ratio to their FV and achieved 72.25%. Alsmearat et al. [7] used a wide range of stylometric features which in fact achieved better results than the BOW approach they tried over the same dataset. Their dataset consisted of news written by famous journalists in MSA from Jordon and Palestine. The addition of the embedding words representation to the FV is used in [8,9] to discriminate the gender of the author. The emojis features were used as a discriminating feature for GI [6,5,10] and had proven to be effective. For classifiers, SGD classifier has rendered good results in Arabic text classification [11] when compared to other classifiers.

3. Dataset

Twitter is one of the factors that helped in the spread of information between Egyptians on the 25th of January revolution [12]. From this time on, a huge amount of Egyptian accounts were created on Twitter with a large number of tweets posted. The conducted experiments in this paper are reliant on a dataset retrieved from Twitter. The dataset initially contains around 283,690 gender-labeled tweets. The tweets are retrieved from 140 public accounts of active people and social media influencers who have more than 2000 followers per account. The selected accounts contain tweets about different topic categories; accounts that only tweet about specific topics, such as politics and sports, are avoided. Furthermore, the selected accounts are for Egyptian dialect speakers of both genders. The tweets were annotated as male and female according to the gender of the account's owner. The gender was determined by the name and profile picture of the account's owner. Since most of the public accounts were chosen for famous social media influencers, this was an additional verification for the gender of the owners of the accounts. The rest of the accounts were further verified by sampling 20 tweets from each author and manually annotating them. The collection of this dataset is based on tweets up until on May 2018. The non-Arabic tweets and retweets are filtered out. The total number of tweets decreased after the preprocessing as some tweets were less than five words and some had only links. Also, when trying to make EDGAD balanced and at the same time allow all accounts to have equal representation (i.e. to have the same number of accepted tweets), 1000 tweets were used per account as this was the minimum number remained for some accounts after preprocessing. The acceptance criteria are to have at least five words after preprocessing. The total number of tweets per gender is the sum of all tweets in all the accounts for the specified gender. The dataset has 27,499 unique words from both genders' tweets. The statistics calculated on EDGAD are represented in Table 1. The dataset has almost the same distribution of tweet lengths for both genders as represented in Fig. 1. The dataset is available for further research in Egyptian Arabic dialect[4].

4. Methodology

4.1. Approach overview

The author's GI problem is a text classification problem with two classes, male and female. Text classification needs a labeled (male/female) dataset representing both genders where significant features can be extracted to help to identify the author's gender. Our proposed methodology –illustrated in Fig. 2 – has two main phases. As an initial step the dataset is split by accounts into training and testing sets, this assures that no tweet from the same account is available in both the training and testing sets. Phase one is concerned with training the classifiers using the training set. As a first step for training feature selection is done and the MFV and NFV are formed. Then the classifiers are trained using these feature vectors as inputs to arrive at the desired classification model. Phase two is concerned with testing where it initially works on the testing set to get the features that will be input to the model to produce the output gender. Multiple classifiers are experimented, such as RF, Multinomial Naive Bayes (MNB), Bernoulli Naive Bayes (BNB), LR and Stochastic Gradient Descent (SGD). The next subsections provide more details of some of these steps.

4.2. Data splitting

The data is composed of male and female sets where each set has 70 accounts. Each account contains a variable number of tweets in case of the unbalanced dataset and limited to 1000 tweets in case of the balanced dataset. Cross-validation is applied with 7 folds. Each fold contains 10 accounts of each gender. As a result, the training set will always contain 120 accounts while the testing set contains 20 accounts. This will ensure that tweets from a certain account will not be available in both training and testing in the same fold.

4.3. Feature selection

Feature selection is the process of selecting features from Egyptian Arabic text that have an obvious effect in discriminating the genders. The first FV is the NFV where we conducted experiments on 1 gram, 1-2 gram and 1-3 grams. Another FV, MFV, is created as well which contains a set of effective features.The first set of features is the count of certain emojis that proved to be discriminating in GI. The second feature is the count of female suffixes available in the text. The third feature is categorized into 12 topics/lexicons i.e. swear words, emotions, politics. . .etc. also, they were selected according to experiments that appealed their effectiveness among other topics/lexicons. The last set is the representation of the input by an embedding layer.

4.3.1. Emojis-based features

These are the counts of 159 emojis. These emojis are chosen if the difference in the occurrences of an emoji in one gender is 20% higher than the other as represented in Fig. 3.

[4] https://github.com/shery91/Egyptian-Dialect-Gender-Annotated-Dataset.

Table 1
EDGAD (Unbalanced and Balanced) Statistics.

Metric	Unbalanced		Balanced	
	Female	Male	Female	Male
Number of Tweets	127,083	156,607	70,000	70,000
Number of Tweets After Preprocessing	101,733	124,273	70,000	70,000
Average Tweet Length	11	11	14	14
Vocabulary Per Gender	13,821	16,983	9,687	10,233
Dataset Vocabulary	27,499		18,094	

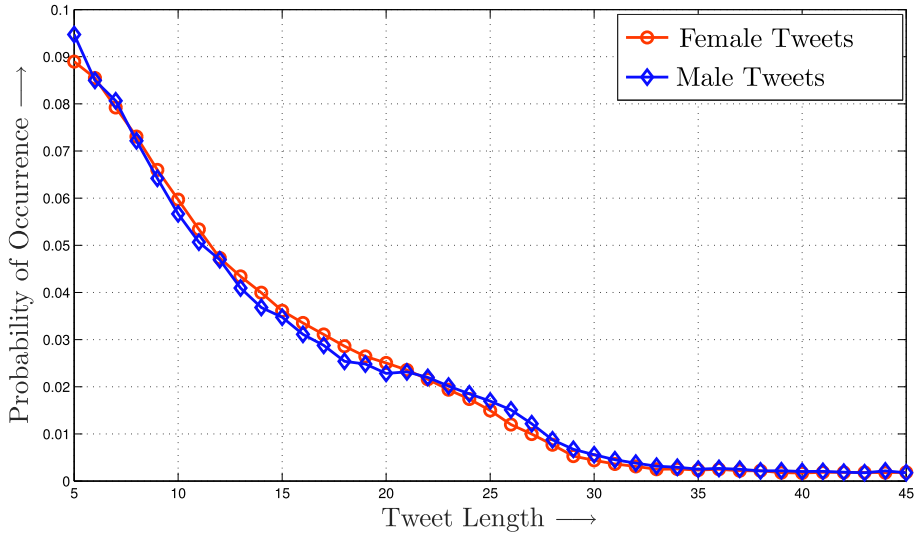

Fig. 1. The probability of occurrence of different tweets lengths in both genders calculated for EDGAD.

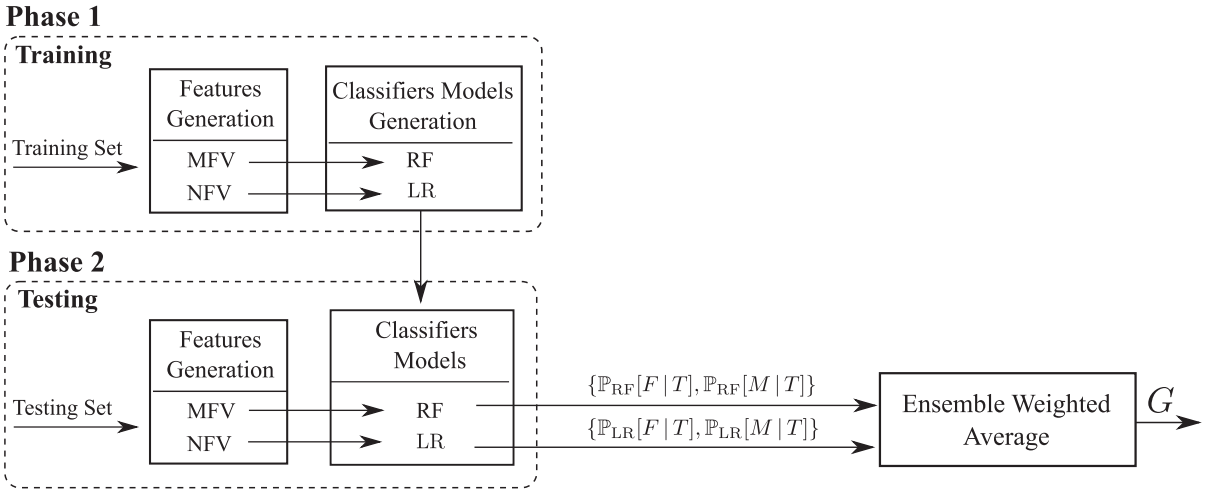

Fig. 2. The block diagram of the proposed classification model for author gender identification. The female gender is denoted by F, the male gender is denoted by M, and the output gender is $G \in \{F, M\}$. The posterior female and male probabilities for the RF and the LR classifiers given the tweet T are denoted as $\mathbb{P}_{RF}[F|T]$, $\mathbb{P}_{RF}[M|T]$, $\mathbb{P}_{LR}[F|T]$, and $\mathbb{P}_{LR}[M|T]$ respectively.

4.3.2. Female suffix feature

This is a single number that represents whether the input text contains words that end with "ه, ت , ة", if we found "أنا", which means "I", before it in the tweet. This feature also checks if some female related words occur, such as "بنت" which means "daughter" or "زوجة , مرات" which means "wife".

4.3.3. Function-based features

There are function words representing twelve different topics (swear words, emotions, football, flirtation, political, love and marriage, jewelry, technology and feelings). Each feature indicates the number of occurrence of words in each category. We gathered a list of words in Egyptian Arabic representing each topic. The swear words list, for example, is a list of 24 types of similar swear words. A list of almost all the swear words in MSA and Egyptian Arabic is made with their different pronunciations, suffixes and prefixes. We added more words to the emotions list published initially by Rabie and Sturm in 13.[5]

[5] https://github.com/shery91/Egyptian-Dialect-Gender-Annotated-Dataset/blob/master/wordLists.zip.

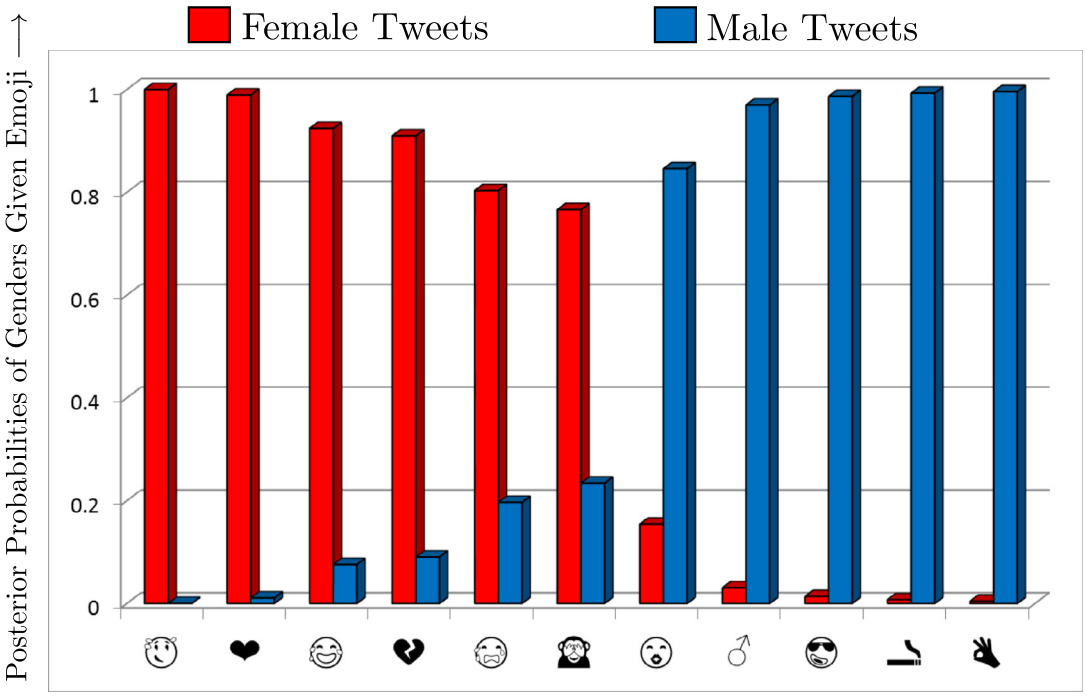

Fig. 3. Subset of emojis that are gender discriminative.

4.3.4. Embedding layer feature

This is a representation for the input text. More tweets are collected; 180,000 tweets are taken from re-tweets and others are retrieved from other 100 accounts in January 2018 that is not going to be used in the dataset. These tweets are used as an input to Gensim.[6] We used the skip-gram model to build our word2vec model. It is parameterized with iterations (5), minimum word counts (5), vector dimension (350), workers (4) and window (8). This embedding layer is used to get the vector representing each word in the input text, then an element-wise average is calculated to get a fixed length vector representing the whole input text, i.e. given a sentence of 4 words, each word is represented as a vector of 350 elements, then the average of each element over all the words is calculated and the final averaged 350 elements are added to the FV.

4.4. Classification

The two feature vectors MFV and NFV will be inputs to RF and LR classifiers respectively. The RF classifier builds multiple decision tree models and outputs the class which is the model of all the decision trees built. The use of multiple decision trees corrects the over-fitting to the training data that can happen in a single decision tree.The decision of using the RF and LR classifiers is based on several experiments that revealed their higher performance over other classifiers, this will be illustrated in details in the experiments section. According to this, the choice was to combine them. The combination is made using ensemble weighted average.

Ensemble weighted average is the process of creating multiple models and combining them to produce the desired output, as opposed to creating just one model. Frequently an ensemble of models performs better than any individual model because the various errors of the models average out [15]. It is used where each classifier is given a weight. Let w_j, where $j \in \{1, 2, \ldots, N\}$ be the weight associated with the jth classifier and N is the total number of classifiers considered. Furthermore, denote $\mathbb{P}_j[F|T]$ and $\mathbb{P}_j[M|T]$ as the posterior probabilities of the test tweet T being a female or male gender, respectively, as computed by the jth classifier. The ensemble weighted average results of all the classifiers are obtained by multiplying the classifier's weight and the classifier's posterior probability for each gender as follows:

$$\mathbb{P}[F|T] = w_1 \times \mathbb{P}_1[F|T] + w_2 \times \mathbb{P}_2[F|T] + \ldots + w_N \times \mathbb{P}_N[F|T], \quad (1)$$

$$\mathbb{P}[M|T] = w_1 \times \mathbb{P}_1[M|T] + w_2 \times \mathbb{P}_2[M|T] + \ldots + w_N \times \mathbb{P}_N[M|T]. \quad (2)$$

The final gender prediction G, is afterwards obtained as,

$$G = \begin{cases} F, & \text{if } \mathbb{P}[F|T] > \mathbb{P}[M|T] \\ M, & \text{otherwise} \end{cases}. \quad (3)$$

In the process of ensemble weighting, LR and RF are selected to be combined together with weights w_1 and w_2 respectively such that $(0 < w_1, w_2 < 1)$, the exact weight for each classifier is found by experiment. They are selected because they have the highest accuracies in the results in addition to the variation in the correctly classified tweets among the two classifiers. This is explained in details in Section 5.

5. Experiments and results

We performed several experiments on EDGAD and PAN-AP'17 in order to get the best results for gender identification for tweets. The experiments include data preprocessing, features selection, the comparison between the results using different features combinations to arrive at the best one, using different classifiers and adjusting their parameters.

5.1. Dataset preprocessing

In the preprocessing step, several transformations are applied to the tweets. Listed below are the preprocessing steps conducted on

[6] Gensim is a package used to create a language model where each word is represented by a vector:https://radimrehurek.com/gensim/.

Table 2
GI accuracy percentage for EDGAD achieved by different classifiers with 1-g and 1-3 g FV. The experiment is carried out over different number of tweets concatenated together with CV and TFIDF.

Classifier	Single		20 Tweets		40 Tweets		60 Tweets		80 Tweets	
	CV	TFIDF	CV	TFIDF	CV	TFIDF	CV	TFIDF	CV	TFIDF
Unbalanced EDGAD										
Baseline	55		54.8		54.7		54.8		54.6	
MNB (1G)	58.7	59.3	72.4	58.6	75.1	58.1	76.2	56.7	76.5	56.5
BNB (1G)	58.5	58.5	71.5	57	72.3	55.7	72.3	55.6	72.2	55.5
LR (1G)	**59.4**	**59.4**	**78.9**	**76.5**	**82.7**	**79.2**	**84.4**	**79.4**	**86.1**	**78.9**
RF (1G)	56.1	56	61.1	61.5	62.8	62.1	62.9	61.0	64.3	61.9
SGD (1G)	56.6	56	75.4	74.4	78.9	77	82.5	79.2	84.1	82.5
MNB (1-3G)	59.1	59.1	67.9	56.4	69.5	56.2	68.1	55.7	67.7	55.6
BNB (1-3G)	57.3	57.3	55.5	55.5	55.7	55.7	55.6	55.6	55.5	55.5
LR (1-3G)	**59.3**	**59.0**	**77.8**	**71.8**	**82.0**	**74.9**	**83.0**	**72.6**	**84.8**	**71.7**
RF (1-3G)	55.7	55.7	55.5	55.6	56.4	56.7	55.6	56.2	56.7	56.6
SGD (1-3G)	57.1	56.7	74.9	74.2	79.9	77.3	80.3	79	81.7	81.2
Balanced EDGAD										
Baseline	50									
MNB (1G)	57.9	58.3	73.5	72	76.5	72.2	77.1	72.5	77.0	71.5
BNB (1G)	57.7	57.8	74.3	59.3	77.1	52.9	77.8	51.4	78.6	50.9
LR (1G)	**58.2**	**58.4**	**78.5**	**75.4**	**82.1**	**77.5**	**83.2**	**78.2**	**85.0**	**77.9**
RF (1G)	52.8	52.2	61.1	59.4	63.2	62.0	63.6	61.9	61.2	62.4
SGD (1G)	55.8	55.5	75.1	73.9	78.8	78.0	79.7	79.3	81.5	80.2
MNB (1-3G)	57.9	58.3	73.5	70.5	76.0	71.4	77.1	71.1	76.6	70.1
BNB (1-3G)	57.0	57.5	72.4	50.9	75.4	50.9	75.5	50.9	72.7	51
LR (1-3G)	**58.1**	**57.7**	**77.2**	**70.8**	**80.7**	**71.8**	**82**	**71.0**	**83.6**	**70.7**
RF (1-3G)	50.1	50.4	53.6	51.4	53.5	52.9	54.2	52.1	54.4	55.7
SGD (1-3G)	56.3	56.1	75.0	73.8	77.6	76.5	78.7	80.8	78.6	79.5

Table 3
GI accuracy percentage for PAN-AP'17 achieved by different classifiers with 1-g and 1–3 g FV. The experiment is carried out over different number of tweets concatenated together with CV and TFIDF.

Classifier	Single		20 Tweets		40 Tweets		60 Tweets		80 Tweets	
	CV	TFIDF	CV	TFIDF	CV	TFIDF	CV	TFIDF	CV	TFIDF
PAN-AP'17 – Egyptian Dialect										
Baseline	50									
MNB (1G)	57.5	57.4	64.5	61.4	64.4	62.4	65.1	59.8	63.0	60.1
BNB (1G)	57.31	57.3	62.6	62.9	63.7	61.6	63.5	52.6	61.9	52.7
LR (1G)	**56.4**	**57.4**	**70.1**	**68.7**	**75.4**	**68.4**	**74.8**	**68.9**	**73.9**	**65.1**
RF (1G)	52.3	51.5	56.1	56.5	54.7	54.9	55.7	54.3	54.1	52.7
SGD (1G)	55.2	54.5	65.9	65.9	69.2	73.2	69.5	76.9	68.6	72.8
MNB (1-3G)	57.2	57.4	64.3	61.3	63.9	61.8	63.5	59.5	61.7	59.6
BNB (1-3G)	56.2	56.1	57.3	52.9	60.3	49.7	54.1	50.5	61.6	47.2
LR (1-3G)	**56.9**	**57.1**	**71.6**	**64.0**	**77.2**	**64.1**	**76.1**	**62.3**	**76.4**	**62.8**
RF (1-3G)	50.2	50.3	53.9	52.3	55.4	51.8	50.5	50.0	52.5	51.1
SGD (1-3G)	55.4	55.1	66.6	67.4	69.9	73.3	72.4	75.4	72.1	75.93

the dataset and an illustration of whether they were actually applied (Used) or found useless (Ignored).

- (Used) – The non-Arabic characters (English characters, digits, special characters and links) are removed.
- (Used in some features) – The repetition of characters in the word are removed. That is to say, "فرحاااااااانة" which means "I am haaaaaaappy". We created a list containing most of the words that originally have two repeated consecutive characters in order to normalize only the words that have extra repetitions
- (Used) – A whitespace between emojis is added to help the tokenizer to separate each emoji on its own.
- (Used) – Tweets less than five words are discarded along with duplicate tweets.
- (Used in some features) – Text normalization is useful for the function-based features and swear words. An example of Arabic text normalization is removing "ال", "ء" from each "إ", making "ة" – as only "ه" and keeping "ي" in only one form "ى"
- (Ignored) – Removing the stopping words for (MSA), a new list for Egyptian dialect stopping words is created with all their possible spelling mistakes. It is observed that the use of stop words can be an aiding identification feature for genders.
- (Used in some features) – Removing prefixes and suffixes. Prefixes can be the definite article "ال" which means "the" in English, "و" which means "and" that is sometimes mistakenly attached to the beginning of a word without a separating space and some prepositions that are attached to the word like "ل, ب, ف, ك,". Suffixes are syllables attached to the end of the word in Arabic, such as "ين, هم, هن, هما, نا, كم,". Some suffixes state that the speaker is singular or plural, male or female, which is the reason why removing them can degrade the accuracies.

5.2. Text classifier

In the process of text classification for gender identification, a baseline classifier was required in order to evaluate the perfor-

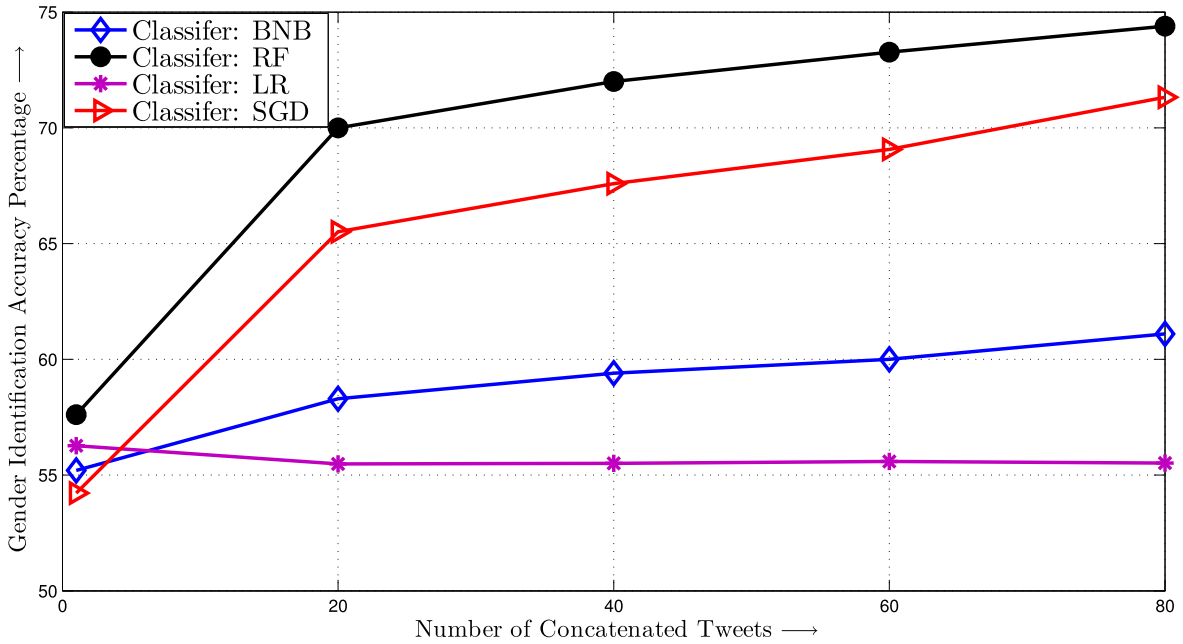

Fig. 4. GI accuracy percentage on EDGAD achieved by different classifiers using all the features and different number of concatenated tweets from single up to 80 tweets.

Table 4
The GI accuracy percentage achieved by the RF classifier using each feature separately for 40 concatenated tweets on EDGAD.

Feature	GI Accuracy %
Characters based feature	62.9
Pronouns	54.4
Definite article	55
Emoji	70
Emotions	65.1
Swear words	68.9
Function words	61.2
Embedding layer	66

mance of the proposed approach. The baseline used here is the majority class probability which is 55% for the male class in EDGAD in the case of unbalanced dataset (single tweet) and 50% in the balanced one.

The first step in our approach was to try different classifiers in order to get the highest classification accuracy. In Tables 2 and 3 we investigated the MNB, BNB, RF, LR and SGD classifiers using Sklearn package[7] with the different number of tweets concatenated together for both the balanced and unbalanced cases in EDGAD, the highest scores are emphasized by bold. The input vectors used in these experiments are the 1 g and the 1–3 g using Count Vectorizer (CV) and Term Frequency-Inverse Document Frequency (TFIDF)[8]. MNB was used with the default parameters. However, BNB was parameterized with 0.1 binarize. RF classifier was parameterized with 60 minimum sample leaf, 50 maximum depth and 10 N-estimators parameters. LR classifier is used with the default parameters. SGD parameters are configured to get the best results as follows; perceptron loss, l2 penalty, 1e-5 alpha, inv-scaling learning rate, 0.1 eta0, 10 maximum iterations and 42 random states. The next step is to find the best classifier for the proposed feature vector MFV. The next subsection presents the details of all the experiments conducted concerning the MFV.

[7] http://scikit-learn.org/stable/.

[8] It's a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.

5.3. Feature vector

There are several aspects that affect the results of classification using the MFV. To get the best results, the first step is to find the classifier that achieves the best accuracy with the full list of features in the MFV. The initial MFV consisted of both the features illustrated in Section 4.3 and some additional features that were eliminated afterwards as they were experimentally proved to have the negative effect on the classifier i.e. character-based features, count of pronouns...etc..The second step is to find the features subset that represents the best combination that can achieve the greatest enhancement of the results of the first step. Each step will be explained in details with all its conducted experiments.

Step 1: The classifier: The classifier that achieved the best results with NFV is not necessary to perform well with MFV. In Fig. 4, the initial MFV has experimented with different classifiers such as RF, LR, BNB and SGD. The RF achieved the best accuracy percentage when used with MFV for variable input length (number of concatenated tweets).

Step 2: Features combination: In the feature selection process, we investigated the classification results for each feature separately using the best classifier in the previous step on an average number of concatenated tweets (40 tweets) as shown in Table 4. Experiments were conducted using different feature combinations in order to eliminate the features that mislead the classifiers. As an example, we had a feature set for character-based features i.e. the characters count, the Arabic characters count, the digits count and another feature for pronouns count. Although these features have proven to be effective for gender identification [7,14], when we eliminated them from the feature vector, the results improved as shown in Fig. 5 where we represent two sets of features combinations. The function words features were applied for more than 20 topics i.e. family, work, food, medical, ⋯etc., but the chosen topics to be added to the feature vector are those whose difference in the occurrence percentage between male and female is more than 15% in order to ensure that the available topics are the most discriminative. For instance, swear words were observed clearly more in males' tweets than in females' ones in all the swear words categories.

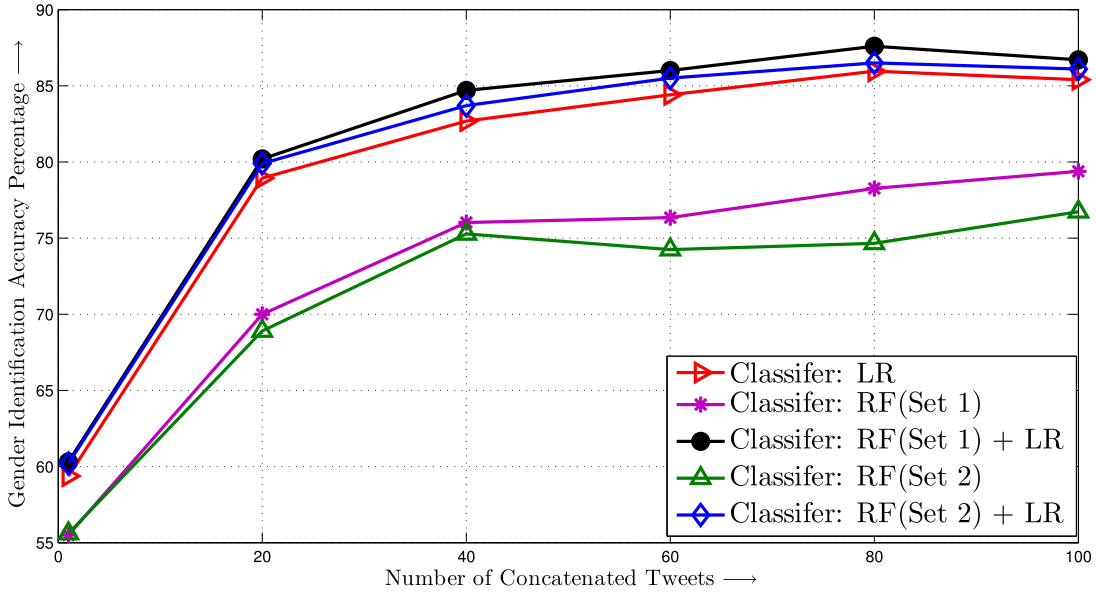

Fig. 5. The GI accuracy percentage on EDGAD for different classifiers: *i*) RF classifier with different features' sets, *ii*) LR classifier with NFV (1G), and *iii*) The ensemble weighted average of RF and LR classifiers. Set 1 represents the emoji, female suffixes, function words, and embedding layer. Set 2 includes the features in Set 1 as well as character based features, pronouns and definite article.

Table 5
The resulting GI accuracy percentage on the Egyptian dialect part from PAN-AP'17. Set 1 is a group of features, it consists of emoji, female suffixes, function words and embedding layer.

Classifier	Single Tweet		20 Tweets		40 Tweets		60 Tweets		80 Tweets	
	1G	1-3G	1G	1-3G	1G	1-3G	1G	1-3G	1G	1-3G
LR	56.4	56.9	70.2	71.2	74.8	77.0	75.2	76.4	73.9	76.3
RF(Set 1) + LR	**58.1**	**58.1**	**71.5**	**72.6**	**76.2**	**78**	**76.1**	**77.2**	**75.7**	**77.4**

5.4. EDGAD results

The experiments carried on the classifiers that achieved the highest accuracy on EDGAD with NFV and MFV, namely LR and RF respectively, showed that the combination of both classifiers with ensemble weighting achieved better results. As illustrated in Fig. 5, LR used NFV with 1 g and the RF was experimented with two different sets of features, the combination between RF with Set 1 and LR with NFV achieved a 1.3% increase in the accuracy, for the 80 concatenated tweets case, over LR that achieved the best accuracy, 85.4%. The reason behind this is that after analyzing the test cases with MFV and NFV, some of the tweets that were misclassified using the N-gram were correctly classified with MFV. Accordingly, the proposed model using ensemble weighting was a means to combine the correct outputs from both classifiers. To adjust the weights of both classifiers we worked over the range of [0.05, 0.95] with 0.05 step. The best results were achieved with 0.6 weight for RF classifier with MFV and 0.4 for LR classifier with NFV.

5.5. PAN-AP'17 egyptian dialect tweets

The weighted ensemble average between LR with NFV and RF with MFV further experimented with the PAN-AP'17 [1]. The Arabic part in PAN-AP'17 is composed of four Arabic dialects (Egyptian, Gulf, Levantine and Moroccan) with 600 author per dialect; each author has 100 tweets. Most of the proposed solutions in [1] grouped 100 tweets of the same author together to work as a document. We tested our approach over their Egyptian dialect tweets only as MFV depends totally on the Egyptian Arabic words. Several numbers of tweets concatenation were tried with different classifiers. The experiments proved that LR achieved the best accuracy as shown in Fig. 4 and our combined approach surpassed it as shown in Table 5. The weight of the RF classifier in the ensemble weighting is 0.6 as calculated before on EDGAD.

6. Conclusion

In this paper, a balanced dataset for Egyptian Arabic tweets labeled by the gender of the author is created. Several statistics of the dataset are obtained in order to be reusable in future research. A solution for the gender identification problem is proposed using the obtained dataset. A novel engineered mixed feature vector, as well as N-gram feature vector are utilized with ensemble weighting for two classifiers, the RF and LR respectively. The proposed feature vector involved emojis, female suffixes, and a group of well-selected function words i.e. swear words, emotions, politics,..etc. Furthermore, an average embedding representation for the tweet is added in the feature vector. Some features required additional preprocessing such as removal of extra repeated characters. Hence, a list of MSA and Egyptian dialect correct words with up to two repeated characters is created. In emotions features, additional words along with their synonyms and common spelling mistakes are appended to an already existing list. Moreover, lists of swear, politics, love and marriage, football, flirtation and technology words used in Egyptian dialect are created. The experiments carried employing weighted ensemble for the RF classifier with MFV and the LR classifier with NFV result in 87.6% accuracy on EDGAD. Furthermore, the proposed classification model achieved 77.4% accuracy on the PAN-AP'17 dataset.

References

[1] Rangel F, Rosso P, Potthast M, Stein B. Overview of the 5th author profiling task at PAN 2017: gender and language variety identification in twitter. In: Notebook for PAN at CLEF.

[2] Rangel F, Rosso P, Gomez MMy, Potthast M, Stein B. Overview of the 6th author profiling task at PAN 2018: multimodal gender identification in twitter. In: Notebook for PAN at CLEF.

[3] Tellez ES, Miranda-Jimenez S, Moctezuma D, Graff M, Salgado V, Ortiz-Bejar J. Gen-der Identification through Multi-modal Tweet Analysis using MicroTC and Bag of Visual Words. In: Notebook for PAN at CLEF.

[4] Tellez ES, Moctezuma D, Miranda-Jimenez S, Graff M. An automated text categorization framework based on hyperparameter optimization. Knowl-Based Syst 2018;149:110–23. Available:http://arxiv.org/abs/1704.01975.

[5] von Daniken P, Grubenmann R, Cieliebak M. Word unigram weighing for author profiling at PAN 2018. In: Notebook for PAN at CLEF.

[6] AlRifai K, Rebdawi G, Ghneim N. Arabic tweeps gender and dialect prediction. In: Notebook for PAN at CLEF.

[7] Alsmearat K, Al-Ayyoub M, Al-Shalabi R, Kanaan G. Author gender identification from Arabic text. J Inf Secur Appl 2017;35(8):85–95. Available: https://linkinghub.elsevier.com/retrieve/pii/S2214212616301715.

[8] Akhtyamova L, Cardiff J, Ignatov A. Twitter author profiling using word embeddings and logistic regression. In: Notebook for PAN at CLEF.

[9] Takahashi T, Tahara T, Nagatani K, Miura Y, Taniguchi T, Ohkuma T. Text and image synergy with feature cross technique for gender identification. In: Notebook for PAN at CLEF.

[10] Chen Z, Lu X, Shen S, Ai W, Liu X, Mei Q. Through a gender lens: an empirical study of emoji usage over large-scale android users. CoRR. vol. abs/1705.05546; 2017. Available:http://arxiv.org/abs/1705.05546

[11] Alotaibi F, Lee M. Mapping arabic wikipedia into the named entities taxonomy. In Proceedings of COLING, Mumbai, December 2012. pp. 43–52.

[12] Lotan G, Graeff E, Ananny M, Gaffney D, Pearce I. The Arab Spring the revolutions were tweeted: information flows during the Tunisian and Egyptian revolutions. Int J Commun 2011;5:2011.

[13] Rabie O, Sturm C. Feel the heat: emotion detection in arabic social media content. In: International Conference on Data Mining, Internet Computing, and Big Data, Kuala Lumpur, Malaysia.

[14] Cheng N, Chen X, Chandramouli R, Subbalakshmi KP. Gender identification from E-mails. In: IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, USA.

[15] Gretel Liz De la Pe~na Sarraćen, Ensembles of methods for Tweet Topic Classification, The Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017), 2017;1881. http://ceur-ws.org/Vol-1881/COSET_paper_1.pdf