Predicting author characteristics of Arabic tweets through Author Profiling *

Isabella Karabasz, Paolo Cellini, and Gonzalo Galiana

Universitat Politècnica de València, Spain
{iska1, paocel, gongafor}@masters.upv.es

Abstract. In this paper we present our team participation in the Author Profiling Task for the APDA@FIRE-2019 competition, using the bag of words technique and including a search for additional indicative vocabulary, we trained a random forest model to categorize the age, gender, and dialect of authors of Arabic tweets.

Keywords: Author Profiling · Arabic · Discriminatory Vocabulary · Bag of Words · Tweets

1 Introduction

Author Profiling is a method that is used to study the use of language through written text with the objective of collecting information about the author and identifying characteristic traits.

Nowadays, there is a tremendous amount of technology that is used to create and share text; moreover, there is a widespread availability to numerous platforms where every day millions and millions of texts are created and shared in digital format. Thanks to the liberty that allows us all to write whatever they want, many of users publish their daily thoughts or actions for every one to read and interpret, thereby building for themselves a separate digital identity. As such, it is becoming increasingly relevant to develop a system to hold individuals accountable for their actions online, just as in the physical world.

The acquisition of user information can be used for many different objectives. This task is motivated by the aim to improve cyber-security by detecting potentially threatening messages and identifying their author. Understanding how individuals of a certain age, gender and origin think and use language is a key part in fulfilling the following task. Nonetheless, to get the necessary data, we need to design and use algorithms that can learn recognize characteristics of each class of author, and this will be the objective of this project.

To achieve this objective, the social media platform Twitter will be used as a source for the recollection of texts from people of the Arabic world, with the purpose to analyze some users. This will be done through the design, implementation and use of Machine Learning algorithms, that utilize the method of Author Profiling to try to predict characteristics of each author based on his/her writing features.

2 Task Description

The APDA (Author Profiling And Deception Detection In Arabic) task is separated into two subtasks, of which our team chose to take part in Author Profiling [2]. Given a set of Arabic tweets, the goal of the task is to design and implement and algorithm in the R programming language to predict the gender, age, and dialect of the authors of these texts.

The given training corpus consists of a total of 2250 authors and 100 tweets per author (in XML format), whereby each author is assigned three labels: gender, age and dialect (in a separate text file). The gender category is a binomial class with labels 'Male' and 'Female'. The age category is divided into three classes: 'Under', 'Between', and 'Above', where the class 'Under' represent ages under 25 years, the class 'Between' represent ages between 25 and 34 years and the class 'Above' represents ages above 35 years, inclusively. Finally, the dialect category is divided into fifteen classes, such as as 'Sudan', 'Morocco', and 'Algeria', just to name a few. By studying how language is shared by people, we can learn to distinguish authors of different categories.

3 Proposed Approach

We propose an approach that builds on the Bag of Words technique as a baseline model for each category. From there, variations to the baseline model were made by adding additional features to encounter particular characteristics for each category class.

3.1 Bag of Words

The Bag of Words is a technique used Natural Language Analysis where texts are vectorized into a matrix representation for more effective computation. In the matrix, there is one row per author and one column per word. The words are based on a vocabulary created by calculating the N most frequent words of the entire corpus. This matrix is then used as input for a selection of machine learning models such as Random Forest, Support Vector Machines and Decision Trees, all of which are offered in the *caret* package in R.

3.2 Discriminant Vocabulary

One of the key modifications done to the baseline model was the addition of a discriminatory vocabulary for each category. That is, a list of words that are perhaps less frequent and not included in the original model but offer conclusive information about the gender, age, or dialect of the author.

Once the Bag Of Words and Discriminant Vocabulary techniques have been presented, below is the suggested workflow of our approach.

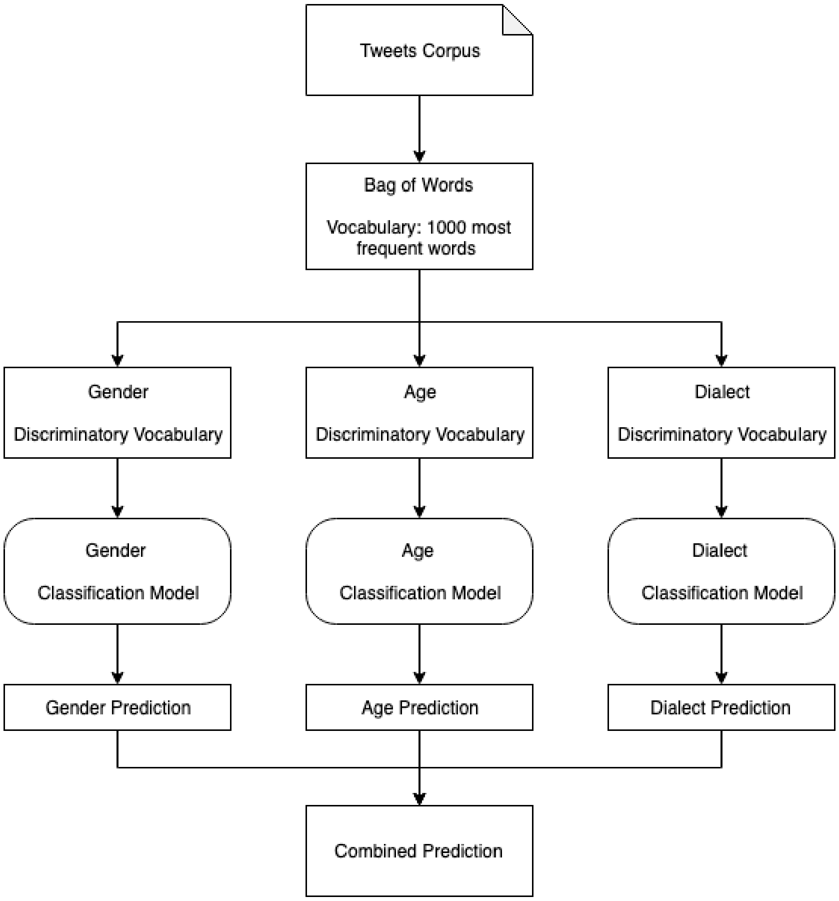
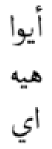


Fig. 1. Proposed Workflow.

The use of discriminatory words provides the bag of words with some vocabulary that may have been excluded, but with high information gain. Their low frequency but exclusive use serves to make definite predictions on the age, gender or dialect of the author. Below are some examples included in each discriminant vocabulary.

In the case of predicting dialect, we have selected some words that are unique to several regions of the Arab world [1]. For example, the list of words below all mean *yes* but are used in distinct regions; Egypt, the United Arab Emirates and Iraq respectively.



أيوا
هيه
اي

Fig. 2. Arabic word for *yes* in different dialects.

In the case of the prediction of the gender and age, as we are working on the Arabic language, we try to find the most common words based on the most common topics in the Arabic world.

For example, for the gender, we use words or phrases that only men or women would say, for example "my husband" would likely only be said by a female, and "my wife" in the case of a male. Further, Arabic pronouns differ for males and females, so some common pronouns were included as well. As for age, we used words that have been found to be common blog topics for each age group [4].

Gender	**Age**
my husband	*homework*
my wife	*bored*
us (f.)	*apartment*
us (m.)	*marriage*

Table 1. Some example words included in age and gender discriminant vocabularies.

4 Experiments and Results

4.1 Machine Learning Models

The base model uses the SVM linear kernel provided in the caret package in R, and yields a cross-validation accuracy of 84% for the dialect category. We have also experimented using SVM Polynomial, decision trees and random forest machine learning models. Despite taking considerable time to train, the random forest method was truly worth the wait as it increased the accuracy by around 6% from the dialect category. We have also experimented using SVM Polynomial, C5.0 decision trees and random forest machine learning models. Despite taking considerable time to train, the random forest method was truly worth the wait as it increased the accuracy by around 6% from the original SVM Linear model for the prediction of dialect. Similar improvements occurred for the other two categories.

4.2 Stopwords

Besides the inclusion of a discriminatory vocabulary, the inclusion of stopwords was one of the first modifications done to the model. These are words that occur in high frequency that offer little to no additional information about the content of the text. The stopwords include months, days of the week, spelled out numbers, prepositions of time and place, among many others to create a list of 750 words that are omitted when creating the bag of words vocabulary.

4.3 Absolute or Relative Frequency

The baseline bag of words uses the absolute frequency of each word in the feature matrix. Alternative variations include replacing the absolute frequencies with relative frequencies. That is the number of times a word occurs per author is divided by the total number of words written by that author.

Classifier	Frequency	Accuracy
Dialect	Absolute	88.57%
Dialect	Relative	89.02%

Table 2. Absolute and Relative Frequencies

When applied to the dialect classifier, the use of relative frequency increased the cross validation accuracy by 1%

4.4 Discriminatory Emoticons

The use of emoticons adds character, style, and of course emotion to a tweet. In fact, some of the most frequent words that ended up in the vocabulary were emoticons. Here they are:

🤣 🤔 🖤 ✨ 🤦🏻‍♂️

Fig. 3. Some of the most common words are in fact emoticons.

We decided to add some emoticons to the gender discriminatory vocabulary. While some emoticons shown in Figure 3 above are likely to be used equally by men and women, the one on the right is far more likely to be used by a male than a female, due to the nature of the image. It is for this reason that we added various emoticons to the gender discriminatory vocabulary. Below are some examples of the emoticons selected.

🤦‍♀️ 🤦‍♂️ 🙋‍♀️ 🙋‍♂️ 💅 👔

Fig. 4. Examples of selected gender discriminatory emoticons.

4.5 Length of Tweets and Number of Mentions

Besides the content of the tweet there are other quantitative characteristics that can be analyzed as well. One of these is the length of tweets. For each tweet, we calculated the word count per tweet and aggregated the results per author by calculating their average word count. In the case of dialect and gender, there were no notable differences between tweet lengths. However, for age, we can observe a trend where younger authors write shorter tweets than older authors, as shown in figure 5 below.

In addition to calculating tweet length, we also calculated the number of mentions per tweet. There were no striking differences between the classes of any of the categories, besides the ”above” age class having very slightly more mentions. Both columns, length of tweets and number of mentions were appended to the matrix of features.

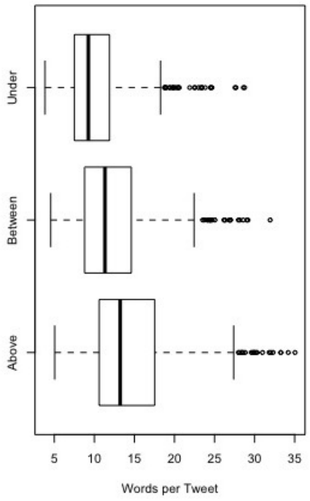


Fig. 5. Tweet length trends by age.

5 Analysis

Overall, we were able to produce good results for the prediction of dialect and gender. However, these high accuracies will be overshadowed in the final combined results due to the low accuracy of the age models.

Classifier	Accuracy (CV)
Dialect	90.22%
Gender	76.89%
Age	54.49%

Table 3. Cross Validation Results of Final Models

It is logical that the best classifier is the dialect classifier. The models are almost entirely based on vocabulary; all the features in the matrix used to train the models are words taken directly from the given corpus. Naturally, the variations of dialect were most easily detected using this method.

Some other vocabulary based experiments that could have been applied to the other categories is a Parts of Speech analysis. It is known that age plays a factor in the way the author conjugates their verbs, young people look to the future and older people are more retrospective. On the other hand, in the case of classifying gender, males use more determiners and women more pronouns. Applying knowledge from these past studies could have had a positive impact on the accuracies of our models [3].

5.1 Future Work

In future projects, it would be imperative to include further experiments in the search for an optimal model. Among these, we suggest n-grams and term frequency-inverse document frequency (TF-IDF).

n-grams n-grams is an additional modification that can be done to the bag of words model. It implies splitting the corpus not strictly into words but into groups of words or characters of varying quantities. We attempted the implementation word 2-grams. That means the corpus was divided into pairs of words and the Bag of Words was created based on the frequency of these word pairs. The computation of this model was highly time consuming, making it too costly to pursue.

TF-IDF ‘Term Frequency-Inverse Document Frequency’ is another technique worth exploring in future developments of the task. It measures the uniqueness of terms within documents compared to others. That is, if a word is highly recurrent with the tweets of one author, and has a low frequency is almost all other documents, then it holds relevant value for the classification of the author in whose tweets this term appears. This would be particularly useful in the development and application of a discriminant vocabulary for each classifier.

6 Conclusion

During the development of this task, the most challenging part of the process was the fact that the tweets were in Arabic. While creating the discriminant vocabularies, we were guided by our biased intuition when selecting some topics for gender and age. We cannot know for sure if they are the most representative words and topics due to our unfamiliarity with Arabic culture. Depending solely on the corpus given to research common topic, we could run the risk of overfitting the models. This is why it would be necessary to dedicate more time to the study of the Arabic world to get a better understanding of the nuances of the cultures in order to improve the results of the task.

The being said, despite the difficulties in programming with an Arabic corpus, the fact that we could not understand the text has given us incentive to approach the task with a more analytic perspective. Just as machines do not understand the content of a corpus until it is processed using NLP techniques and automated learning, we performed this task without using the corpus as a safety net for immediate validation. We were forced to trust the process and the code; our process depended purely on the results of our models and on the understanding of the bag of words technique.

References

1. A. Randa. How different are arabic dialects? 2019.
2. F. Rangel, P. Rosso, A. Charfi, W. Zaghouani, B. Ghanem, and J. Snchez-Junquera. Overview of the track on author profiling and deception detection in arabic. In: Mehta P., Rosso P., Majumder P., Mitra M. (Eds.) Working Notes of the Forum for Information Retrieval Evaluation (FIRE 2019). CEUR Workshop Proceedings. CEUR-WS.org, Kolkata, India, December 12-15.
3. P. Rosso. Author profiling and fake reviews identification. *PRHLT Research Center, Universitat Politècnica de València*, 2019.
4. J. Schler, M. Koppel, S. Argamon, and J. Pennebaker. Effects of age and gender on blogging. *American Association for Artificial Intelligence*, 2005.