Arabic Tweeps Traits Prediction AT2P

Khaled Alrifai [0000-0002-2847-414X], Ghaida Rebdawi [0000-0002-1654-4489] and Nada Ghneim [0000-0003-1167-1718]

Higher Institute for Applied Sciences and Technology, Damascus, Syria
{khaled.alrifai,ghaida.rebdawi,nada.ghneim}@hiast.edu.sy

Abstract. Author profiling is the process of extracting author traits, which constitute the profile of an author, by analyzing his/her writings. Detecting these traits is useful and important process in the domain of social media analysis. In this notebook, we present our approach for author profiling task that is one of the tasks required in "Author Profiling and Deception Detection in Arabic (APDA)" workshop 2019. The focus of this task is to identify the age, gender, and language variety of Arabic Twitter users (tweeps). For this purpose, several feature vectors and classifiers were evaluated to find out the best prediction models for the three traits. *SMO* classifier with the feature vector that consisted of *UniGram* and *Stem* was the best model for each three traits: gender, age and variety[1].

Keywords: Author Profiling, Variety Prediction, Age Prediction, Gender Prediction, Arabic Social Media Processing, Machine Learning.

1 Introduction

Author profiling on social media is a method of analyzing the author writings on social media in order to uncover different traits of the author (e.g. gender, variety and age) based on stylistic or content-based features. This method aims at taking advantage of a huge volume of data generated by a huge number of authors, in order to classify them into predefined classes based on their traits [1].

Author profiling has gained much importance due to its various applications in business, social studies and security areas. From a marketing viewpoint, extracting latent traits of the authors is used for targeting advertisement campaigns, companies also may be interested in knowing, on the basis of the analysis of online product reviews, the demographics of people that like or dislike their products. From social studies viewpoint, having knowledge about the authors that have specific behavior toward new trends is very important issue, these may be used for classifying social accounts for future trends.

With the birth and rise of social media [2], internet users in the Arab world were quick to embrace the new technology, and utilize all what social media has to offer to connect, communicate, and share information with others using Arabic language.

Arabic language used in social media has two forms: the first, is the Modern Standard Arabic (MSA), which is widely used in formal situations like formal speeches, government and official contents; the second, is known as Dialectal Arabic (DA) which is the informal private language, predominantly found as spoken vernaculars with no written standards. Dialects differ in morphologies, grammatical cases, vocabularies and verb conjugations [3]. These differences call for dialect-specific processing and modeling when building Arabic automatic analysis systems [4].

Concerning gender of tweeps, Twitter does not collect users' self-reported gender as do other social media sites (e.g., Facebook and Google+) [5]. Topics and style of writing vary depending on the author gender (males and females). Males could be interested in sports, politics and economy, whereas females could be interested in fashion and celebrity news. In Arabic language, words suffixes and prefixes differ between males and females, for example: "تشاركين" "t$Arkyn" word for second person female vs. "تشارك" "t$Ark" word for second person male. Moreover, females tend more to use emojis, whereas males may tend to write textual tweets. These differences could be used as indicators when developing a gender prediction model for Arab tweeps.

According to age of the tweeps, the writing style and type vary between young and old tweeps. Old tweeps may write long and formal tweets more than young ones. Young tweeps may write more about sport, studying issues and fashion, whereas old tweeps may tweet politic and social topics. These assumptions could be used as indicators to distinguish between various age groups.

In this paper, we summarize our participation in "Author Profiling and Deception Detection in Arabic (APDA)" workshop 2019 [6], in the "author profiling in Arabic tweets" task. We represent our methodology for developing the prediction models for the traits under study in the workshop: age, gender and variety (dialect) of the Arabic authors.

In the rest of this paper, we present in section 2 our methodology that includes: the characteristics of training and testing data, the features used for the developed models, and a step-by-step approach to build the prediction model. In section 3, a brief discussion of the results is addressed. At the end, a short summary and insights for the future are presented.

2 Methodology

In this section, we describe the dataset used in this work, and the features tested for the prediction models. The proposed models are explained in detail hereafter, including: data pre-processing, features extraction, features filtering and the algorithms with their evaluation criteria.

2.1 Dataset

APDA workshop provided the participants in author profiling task with training and testing data from Twitter. Training data consists of 225000 tweets written in Arabic language by 2250 authors equally (100 tweets per each author). Authors were tagged with three traits: variety, age and gender. Variety trait was categorized into fifteen classes: Algeria, Egypt, Iraq, Kuwait, Lebanon-Syria, Libya, Morocco, Oman, Palestine-Jordan, Qatar, Saudi Arabia, Sudan, Tunisia, UAE and Yemen. Age trait was categorized into three age groups: under 25, Between 25 and 34, and Above 35. Gender trait was categorized into two classes: male and female. Authors were divided according to these three traits equally into: 150 authors per each variety, 750 authors per each age group and 1125 authors per each gender.

Testing data was also provided by APDA workshop. It consists of 72000 Arabic tweets, divided into 720 authors equally. Testing data was not tagged corpus, it was blinded. The evaluation process and the accuracy calculation using testing data were achieved by the workshop organizers.

2.2 Studied Features

In our attempt to reach to the best prediction models for age, variety and gender traits, we tried a number of features for each model; some features contributed in improving the accuracy, while others did not. We categorized these features into:

- Content-based features:
  - Uni-gram, bi-gram and tri-gram of words.
  - Stems of words.
  - Lemmas of words.
  - Words Part Of Speech tags (POS), i.e. NOUN_MS_PRON and V_PRON.
  - Character *n*-gram, where *n* ranges from 2 to 7.
- Style-based features:
  - Links to websites ("http").
  - Hashtags to active public trends ("#").
  - Mentions to other authors ("@").
  - Lengthened words, i.e. the intentional repetition of a character in a word to emphasize and to exaggerate in describing something like laughing "هههههههه", magnification "واااااااااااااو", indignation "لاااااااا", etc.
  - Average tweets length, i.e. the average number of words in an author tweets.
  - Tweets punctuation marks, i.e. the summation of punctuation marks used in an author tweets.

2.3 Our Model

In our attempt to find out the best prediction models, we prepared the dataset and extracted the features. These features have been filtered to reduce the size of feature vectors. Depending on reduced feature vectors, we implemented several experiments that differed from each other in feature vector and the algorithm used for training. The resulting models were compared using specific evaluation criteria to select the best one.

Data Pre-processing. Before starting feature extraction stage, we concatenated all the 100 tweets for each author into one long text. This long text was tokenized using Farasa tokenizer [7]. All extracted tokens have been grouped and weighted with their frequency in the dataset (all the tokens from all authors).

Features Extraction. After the tokenization stage, lemmas and stems were extracted from the calculated tokens using Farasa toolbox. Tokens were used also to extract character 2-7 grams.

In all content-based features, the calculated value for each feature was the frequency of use in the dataset. This step produced a huge size of feature vector that should be reduced.

Style-based features were also calculated and extracted for each author. We considered a word is lengthened if it included a character repeated three times at least.

In case of all style-based features -except the average tweets length-, the values that are considered in each feature vector were the frequency of use. In case of average tweets length, the considered value in each author feature vector was the average number of words in a tweet.

Features Filtering. The number of elements of each *content-based* feature vector was very huge, which made the training process very hard and time-consuming. We applied the following steps to reduce the feature vector size:

- Eliminating features with a value less than two (we have two classes at least). The probability that these items contribute in the classification is low.
- Discarding all elements with Information Gain (IG) equals to zero.

Model Training. In our experiments, we trained different models using Weka toolbox [8]. The features mentioned previously have been used separately or jointly to create various feature vectors to be evaluated in several experiments.

According to the classifiers, we used Sequential Minimal Optimization (*SMO*) classifier for each trait. Depending on previous experiments, *SMO* already gave us good results relatively [9].

Evaluation of Models. For the models evaluation, we used training data to find out the best models by calculating the accuracy over 10-folds cross-validation (*AccTrain*).

Concerning the testing data, it was blinded (hidden tags), so the testing accuracy (*AccTest*) for the models are calculated and declared later by APDA organizers.

3 Experiments and Results

In this section, we represent the best six experiments according to *AccTrain* for each trait. The used features in these experiments are abbreviated here as: *UniGram* for word uni-gram, *Stem* for stems, *CNGram* for character *n*-gram and *Ratios* for the six style-based features that previously mentioned. In the following three charts, we ranked the prediction experiments per each trait descending from left to right according to *AccTrain* values.

3.1 Gender

Fig. 1 represents the results of gender prediction. The feature vector that consists of *UniGram* and *Stem* gave the best *AccTrain* (98%). We notice that using *UniGram* alone (97.96%) was better than using *Stem* alone (93.55%). Adding *ratios* to *Stem* improved the model (93.86%), but adding *ratios* to *UniGram* did not. *CNGram* gave less *AccTrain* relatively (87.24%). Fig. 1 shows the results:

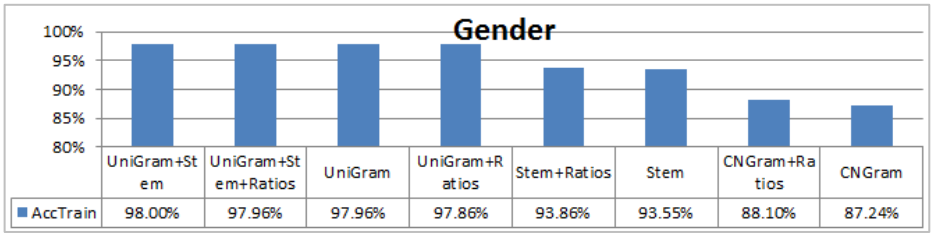

Fig. 1. Gender prediction results

3.2 Age

Fig. 2 represents the results of age prediction. Similarly to the best result for gender prediction, the best feature vector consists of *UniGram* and *Stem* with *AccTrain* equals to (83.24%). We notice that developing age prediction models took the same behavior of gender prediction above. Fig. 2 shows the results:

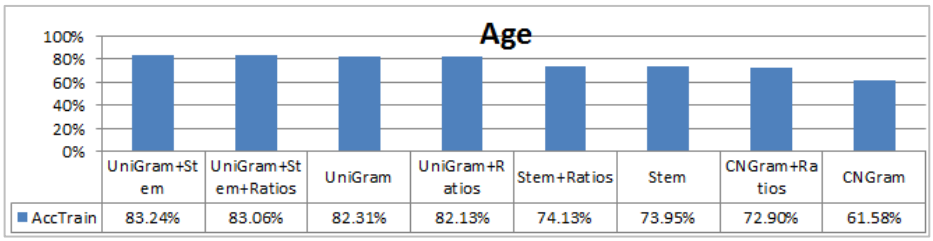

Fig. 2. Age prediction results

3.3 Variety

Fig. 3 represents the results of variety prediction. Here, we notice that adding Ratios to *UniGram* or to *Stem* made a little improvement. Adding Ratios to *UniGram* and *Stem* together gave the best model (92.17%). Using *CNGram* also gave less *AccTrain* relatively. Fig. 3 shows the results:

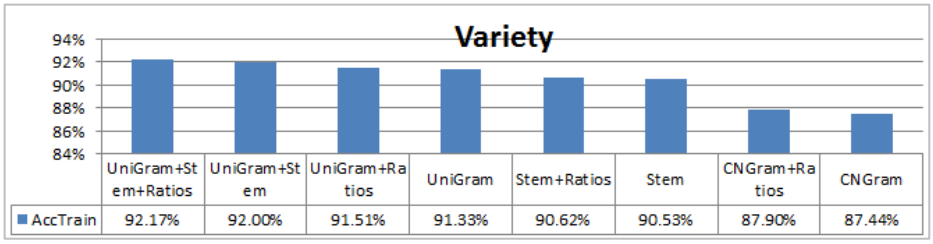

Fig. 3. Variety prediction results

3.4 Discussion

Generally, *UniGram* as *content-based* features gave good results in predicting the three traits comparing with Stems and character *n*-gram. From the results above, the problem of detecting age of author is consider hard problem relatively, the best *AccTrain* for age prediction is (83.24%) comparing to (98%) for gender prediction and (92.17%) for variety prediction. This fact ensures that authors of various ages (young and old) may use the same behavior on social media, making the predicting task harder.

Also, variety prediction considering fifteen classes is harder than gender prediction which considers just two classes; this is axiomatic fact in prediction problems. So, we notice that the highest *AccTrain* for gender prediction (98%) is more than (92.17%) for variety prediction.

3.5 APDA Runs

APDA organizers allow the participants to submit three runs to be compared to other participants. We select the best three experiments per each trait according to *AccTrain* to constitute the three runs: Run1, Run2 and Run3. In Table 1, we represent the three runs with the *AccTrain* and *AccTest* for each run.

Table 1. APDA results

		Run1	Run2	Run3
Gender		**UniGram+Stem**	**UniGram+Stem+Ratios**	**UniGram**
	AccTrain	**98.00%**	97.96%	97.96%
	AccTest	**77.08%**	76.81%	76.67%
Age		**UniGram+Stem**	**UniGram+Stem+Ratios**	**UniGram**
	AccTrain	**83.24%**	83.06%	82.31%
	AccTest	**53.75%**	53.47%	51.39%
Dialect		**UniGram+Stem+Ratios**	**UniGram+Stem**	**UniGram+Ratios**
	AccTrain	**92.17%**	92.00%	91.51%
	AccTest	89.03%	**89.17%**	86.81%
Joint AccTest		**36.39%**	**36.11%**	**34.31%**

According to *AccTest* which calculated by APDA organizers for each trait, Run1 was the best one in case of gender and age traits, Run2 was the best in case of variety. The feature vector that consists of *UniGram* and *Stem* is the best one according to *AccTest* for all traits. We can notice that *AccTest* for age is considered low relatively in comparing with gender and variety (53.75% in best case), this ensures the fact that age prediction is harder than gender and variety prediction as we already discussed.

In the last line of Table 1, joint *AccTest* means the accuracy that is calculated when the three developed models correctly predict the three traits for each author. Best joint *AccTest* is calculated in Run1 (36.39%), this means the three traits of (36.39%) of testing data are correctly predicted by the three models.

4 Conclusion

In this notebook, we summarized our participation in APDA workshop in task of "author profiling in Arabic tweets". For this purpose, we tried several features and classifiers to find out the best prediction models. Depending on testing data, *SMO* classifier with the feature vector that consisted of *UniGram* and *Stem* was the best model for each three traits: gender, age and variety.

Our accomplished ranks were 12th, 21st and 20th for gender, age and variety respectively in comparison with 28 participants. Our rank was 16th in case of joint prediction.

It will be worth investigating more features that could improve the prediction accuracy, especially in case of age prediction, such as: type of written topics, special words and emojis, type of shared links, etc. Using deep learning algorithms may be useful for Arabic author profiling problem in case of availability a huge dataset written by Arabic authors.

References

1. Mart Busger op Vollenbroek, Talvany Carlotto, Tim Kreutz, Maria Medvedeva, Chris Pool, Johannes Bjerva, Hessel Haagsma, and Malvina Nissim. GronUP: Groningen User Profiling. Notebook for PAN at CLEF 2016. University of Groningen, Groningen, The Netherlands (2016).
2. TNS. Arab Social Media Report. First Report (2015).
3. Mohammed Abdulmalik Ali. Artificial intelligence and natural language processing: the Arabic corpora in online translation software. International Journal of Advanced and Applied Sciences (2016).
4. Fei Huang. Improved Arabic Dialect Classification with Social Media Data. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (2015).
5. Thomas Oshiobughie Ugheoke. Detecting the Gender of a Tweet Sender. A Project Report Submitted to the Department of Computer Science In Partial Fulfilment of the Requirements For the Degree of Master of Science In Computer Science, University of Regina, (2014).
6. Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., Snchez-Junquera, J.: 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. In: CEUR-WS.org, Kolkata, India, December 12-15 (2019).
7. Ahmed Abdelali, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. Farasa: A Fast and Furious Segmenter for Arabic. Qatar Computing Research Institute, Hamad Bin Khalifa University, Doha, Qatar. Proceedings of NAACL-HLT 2016 (Demonstrations), pages 11–16, San Diego, California, June 12-17, 2016. Association for Computational Linguistics (2016).
8. Weka official website: www.cs.waikato.ac.nz/ml/weka
9. Khaled Alrifai, Ghaida Rebdawi, and Nada Ghneim. Arabic Tweeps Gender and Dialect Prediction. Notebook for PAN at CLEF (2017).