Arabic Tweeps Gender and Dialect Prediction
Notebook for PAN at CLEF 2017

Khaled ALRIFAI, Ghaida REBDAWI, Nada GHNEIM

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

Abstract. In this paper, we present our approach for author profiling task based on Arabic content (Twitter case), which was one of the tasks required in PAN at CLEF 2017. Author profiling is the process of identifying authors' traits, which constitute the profile of an author, by analysing his/her writings. In our research, we considered the gender and the variety (dialect) of an author as two important traits that have many useful applications in the domain of Arabic social media analysis. For this purpose, several feature vectors and classifiers were tried to reach to the best prediction models for these two traits.

1 Introduction

Author profiling on social media is an attempt to take advantage of a huge volume of data generated every second by a huge number of authors [6]. Author profiling aims at classifying these authors into predefined classes based on their traits. This has obvious advantages in targeted marketing and advertising, as well as in the forensic and security areas.

With the birth and rise of social media [7], 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.

In PAN 2017 [13], dialect and gender of tweeps are the traits under study of author profiling task [12], which are required using several tweets languages: English, Spanish, Portuguese and Arabic. In this research, we focus on the Arabic language for proposing dialect and gender prediction models.

Arabic language has two forms [1][8][10]: the first is the Modern Standard Arabic (MSA), which is widely used in formal situations like formal speeches, government and official operations; the second form 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 [2][5]. These differences call for dialect-specific processing and modeling when building Arabic automatic analysis systems. In PAN 2017, Arabic dialects (or varieties) have been divided into four classes: Levantine, Gulf, Egypt and Maghrebi. Accordingly, the required task was to develop a model that can predict a tweep's dialect based on his/her Arabic tweets.

Concerning gender [3][11], Twitter does not collect users' self-reported gender as do other social media sites (e.g., Facebook and Google+) [3], but such information

could be useful for targeting a specific audience for advertising, personalizing content, and for legal investigation. It is interesting to investigate if a difference in writing patterns is existed between two genders: male and female. Males may use words with prefixes and suffixes different from that been used from females. Also males may be interested in sports, politics and economy, whereas females can be interested in fashion and celebrity news. In addition, writing style may differ between males and females, as females tend to use emojis more frequently than males. These differences could be used as indicators to distinguish between genders.

In the rest of this paper, we represent the characteristics of provided data in section 2. In section 3 we list all used features for the developed models. A step-by-step approach to build the target models is shown in section 4. In section 5, a brief discussion about the results we got is addressed. At the end, insights for the future and a short summary are presented.

2 Data Description and Pre-processing

In this section we describe the provided training data by PAN 2017, how to evaluate the models and what pre-processing was performed.

2.1 Data Description

PAN 2017 provided his participants in author profiling task with training data from Twitter, this data consists of 240000 tweets written in Arabic from 2400 authors equally (100 tweets for each user). Authors were tagged with two traits: variety and gender. Variety trait were divided into four classes: Levantine, Gulf, Egypt and Maghrebi, and gender trait were divided into two classes: male and female. Authors were categorized according to these two traits equally, 600 authors for each variety class and 1200 authors for each gender class. Also gender and variety categories were divided between each other equally, i.e. the 600 Levantine authors was divided into 300 male and 300 female, and so on.

As for the testing dataset, PAN 2017 prepared a platform from self-evaluation called Tira[1] [9]. With Tira we could evaluate our models using blinded test data which consists of tweets for 1600 authors.

2.2 Data Pre-processing

Before starting feature extraction stage, for each user we concatenate all his/her 100 tweets into one long text. Then this long text was tokenised into tokens using Farasa segmenter. Farasa[2] is a fast and accurate text processing toolkit for Arabic [4], it's used repeatedly in this research as we will see.

[1] Tira website: www.tira.io

[2] Farasa official website: www.qatsdemo.cloudapp.net/farasa

3 Studied Features

In our attempt to reach to the best prediction models for variety and gender traits we tried a number of features for each model, some features contributed in improving the accuracy of models, while others non. In this section, we describe all of the features we implemented, while later we will detect the features which contribute to the performance of two models well.

3.1 Character N-gram

Means the most frequent of successive N characters of tokens, where N takes its value from 2 to 7. We considered the frequency of use of a feature that used by a user to consist full feature vector of an author. These features gave the best results for both models as we will see. The secret behind character n-gram features that gather the best features of tokens as full form and their stems with all related prefixes and suffixes. So by using character n-gram, we can dispense with using tokens and stems at the same time without duplication and repetition.

Prefixes and suffixes of stems help effectively in detecting genders especially there are prefixes and suffixes special for each gender. Also every dialect has uses tokens differ from other dialects with same meaning, i.e. Levantine authors use "**كتير**" word a lot which means "much", whereas Gulf authors use "**وايد**" which have the same meaning. These characteristics make using character n-gram effective.

3.2 Links, Hashtags and Mentions Usability Ratios

Authors on Twitter differ from each other in how much use links to another websites in their tweets, also in using hashtags "#" which drive Twitter daily trends and hot discussions happen in worlds, also in using mentions "@" which used to call authors in Twitter within tweets. These differences create motivation to analyse the effect of these Twitter characteristics on models of prediction. These characteristics were counted for each author in training and testing data then normalized into range [0,100]. By experimenting, we noticed that normalization led to better results. Normalization equation where $x = (x_1, \dots, x_n)$ and y_i is normalized value of x_i is:

$$y_i = \frac{x_i - \min(x)}{\max(x) - \min(x)} * 100$$

Formula 1. Normalization equation.

3.3 Lengthened Words Ratio

Lengthened words one of most distinctive characteristics of using social media. It means intentionally repeating of a character in a word for ensuing exaggerating in describing of something like laughing "هههههههه", magnification "واااااااااااااو", indignation "ااااااااا؟" and etc. As we considered, a word is lengthened if a character within it repeated three times and more. Lengthened words was counted for each author then normalized into range [0,100] using Formula 1.

3.4 Others

We mention here several features which tested but didn't contribute well for two models, or gave accuracy less than the best models. We tried uni-gram, bi-gram and tri-gram of tokens, which mean the most frequent of single words, dual words and triple words. These features gave accuracy less than that accuracy given by character n-gram. Also we tried stems of tokens as features which mean the most frequent stems in corpora, these also gave bad accuracy relatively. Instead of using tokens as is, we tried part of speech tagger of tokens, which means the most frequent of combinations of tokens parts, i.e. V_PRON and NOUN_MS_PRON, the results was relatively worst. It is worth mentioning that we used Farasa for extracting POS and stems of tokens.

4 Proposed Models

After discussing data and features that can be extracted from, we can now dive into the proposed models for variety and gender traits. We achieved several experiments and trained a lot of models for each trait to reach to the best ones. Models differ from each other in selected features vector and the algorithm used for training.

4.1 Features Selection

Our methodology to reach to best model for a trait is to start with feature vector and calculate corresponding testing accuracy, then to add new feature, if added feature increases the accuracy then we adopt it, else discard it and move to new features and so on.

For both traits we started with character n-gram which mentioned in 3.1 as feature vector, then tried adding links, hashtags and mentions usability ratios. This addendum was effective for gender but not for variety, so we adopted them just for gender model. Then we tried adding lengthened words ratio which was effective for variety but not for gender, so we adopted them just for variety model.

All these experiments and their training and testing accuracies are shown in Figure 1 and Figure 2.

4.2 Features Filtering

The number of all character n-gram combinations extracted from tokens was very big, and reached about 600,000 features. This made training the model a very hard and time-consuming task. The calculated feature vector size should be reduced by removing unnecessary features. The reduction process had two consecutive steps:

- Neglecting all items with a frequency less than 5. The probability that these items will be useful to the classification is relatively very poor. This filtering step eliminated more than 450,000 feature.
- Neglecting all items with information gain IG less than an experimentally determined threshold.

4.3 Training of Models

In our research, we used Weka[3] toolbox for training models. At a first step, we trained our proposed models (using different feature vectors) using Support Vector Machine SVM algorithm with various kernels, until we reached the best testing accuracy with this algorithm. Then, the best feature vector was used to train a model using Sequential Minimal Optimization SMO algorithm. SMO gave a better accuracy than SVM as we will see later in the results section.

An appropriate choice of the classifier is considered a major step of any machine learning problem, as well also the configuration of the classifier itself plays a crucial role. In our research, especially in use of SVM, we notice that the kernel of SVM is very important parameter we should test accurately, we tired linear and exponential kernel which led to worst results comparing with polynomial one as we will see.

4.4 Evaluation of Models

For the evaluation process, we used both training and testing dataset to get insight about best models. In the training phase, we used the accuracy and the F-Measure (F1) over 10-folds cross-validation for the evaluation. These criteria are mentioned in Formula 2.

$$accuracy = \frac{\#\ of\ sucessfull\ predicted\ samples}{\#\ of\ all\ samples}$$

$$F1 = 2.\frac{precision\ .recall}{precision + recall}$$

Formula 2. F1 and accuracy equation.

On the other hand, for the testing phase, only the accuracy was provided by Tira as an evaluation criteria.

[3] Weka official website: www.cs.waikato.ac.nz/ml/weka

5 Results

In this section, we will present our gender and dialect detection experiments. Here, we will use these abbreviations to show results, CNGram for character n-gram feature vector, LHM_Rs for links, hashtags and mentions usability ratios, and Len_R for lengthened words ratio.

5.1 Features' Comparison

As we already mentioned, several feature vectors have been used to train a number of gender and variety classification models. We compared their results in order to get the best model using SVM classifier as a training algorithm.

At a first step, we trained our model using word n-gram, where n ranging from 1 to 3. We also trained the model using the stems (with stems extracted using Farasa [4]). Although these features did well when used separately or concatenated, but using character n-gram produced more effective results.

Figure 1 shows a comparison between various feature vectors of variety, where training accuracy, training F1 and testing accuracy are calculated. We can notice that adding LHM_Rs to CNGram feature vector have negative effect on accuracies. Whereas, adding Len_R increases them. In conclusion, the best feature vector for variety trait, using SVM algorithm, is consisted of both character n-gram and lengthened words ratio, where training F1, training accuracy and testing accuracy equal to 73.2%, 72.93% and 67.69% respectively.

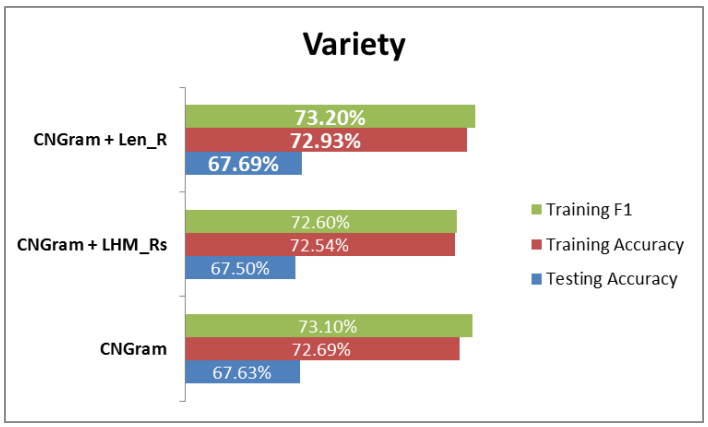

Figure 1. Comparison between feature vectors (variety case).

Figure 2 presents a comparison between gender detection models built using various feature vectors. We can notice that using LMH_Rs with character n-gram increases accuracies, but adding Len_R to both decreases them. Therefore, we conclude that the best feature vector corresponding to the best accuracy is the integration of LMH_Rs and character n-grams. Then, training F1, training accuracy and testing accuracy equal to 74.5%, 70.12% and 68.81% respectively.

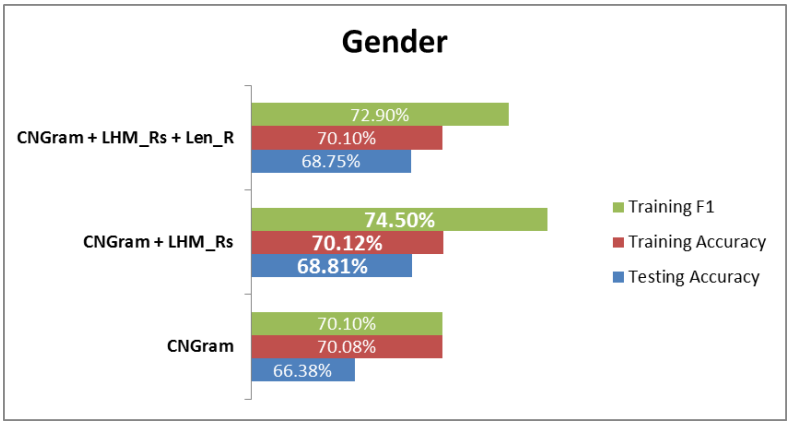

Figure 2. Comparison between feature vectors (gender case).

In Figure 3, we present a summary of best SVM modelling results, which corresponding to best training and testing accuracies for variety and gender traits separately. In addition, we present the accuracies of "both" traits, which is calculated as the percentage of cases where variety and gender of an author are both predicted correctly at the same time.

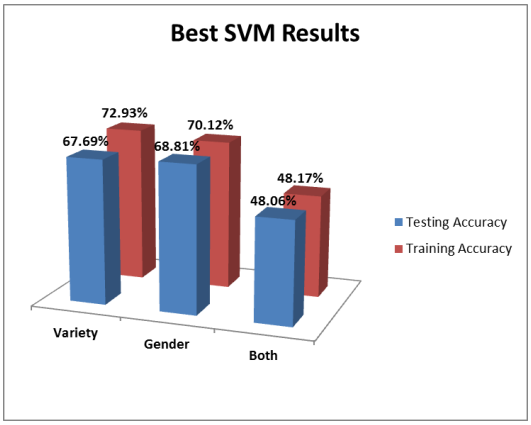

Figure 3. Best SVM models.

5.2 Training Algorithms' Comparison

In this research, we trained using SVM with linear, polynomial and exponential kernels, then using SMO classifiers. The results of using SVM algorithm with the different kernels, on best feature vector of variety for example, showed that the polynomial kernel is the best with F1 equals to 73.2%, compared to 67.1% for the linear and 62.7% for the exponential kernels.

Moreover, we retrained a new model using SMO classifier instead of SVM, and the same best feature vector. The training and testing accuracies for both traits have increased. The results are shown in Figure 4. The testing accuracy has increased about 7% for variety, and 3% for gender. Moreover, the testing accuracy for "both" traits together has increased more than 8%.

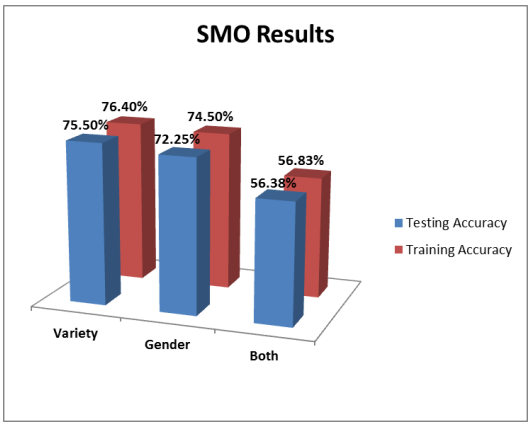

Figure 4. SMO classifier results.

6 Conclusion

In this research, we presented our work in author profiling task PAN 2017 to predict variety and gender of Arabic Twitter authors.

We trained several models using various features and classifiers to find the best models for predicting each trait (variety and gender). We found that character n-gram with SMO classifier led to optimum models for both traits, with testing accuracy equal to 75.5% for variety, 72.25% for gender and 56.38% for both.

It will be worth investigating more classification algorithms, with other stylistic features, that may contribute to enhance the accuracy of variety and gender prediction models.

7 References

[1] Ali M.A.: Artificial intelligence and natural language processing: the Arabic corpora in online translation software. International Journal of Advanced and Applied Sciences (2016).

[2] Huang F.: Improved Arabic Dialect Classification with Social Media Data. Proceedings of the Conference on Empirical Methods in Natural Language Processing (2015).

[3] Ugheoke T.O.: 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, Saskatchewan (2014).

[4] Abdelali A., Darwish K., Durrani N., and Mubarak H.: Farasa: A Fast and Furious Segmenter for Arabic. Proceedings of NAACL-HLT (Demonstrations), San Diego, California (2016).

[5] Nassar N.: Arabic Dialect Identification. Internal report, HIAST, Syria (2017).

[6] Vollenbroek M., Carlotto T., Kreutz T., Medvedeva M., Pool C., Bjerva J., Haagsma H., and Nissim M.: GronUP: Groningen User Profiling. Notebook for PAN at CLEF 2016 (2016).

[7] TNS: Arab Social Media Report. First Report (2015).

[8] Malmasi S., and Zampier M.: Arabic Dialect Identification in Speech Transcripts. In Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects, Osaka, Japan (2016).

[9] Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., and Stein, B.: Improving the Reproducibility of PAN's Shared Tasks: Plagiarism Detection, Author Identification, and Author Profiling. In: Kanoulas, E., Lupu, M., Clough, P., Sanderson, M., Hall, M., Hanbury, A., Toms, E. (eds.) Information Access Evaluation meets Multilinguality, Multimodality, and Visualization. 5th International Conference of the CLEF Initiative (CLEF 14). pp. 268–299. Springer, Berlin Heidelberg New York (2014).

[10] Mubarak H., and Darwish K.: Using Twitter to Collect a Multi-Dialectal Corpus of Arabic. Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), Doha, Qatar (2014).

[11] Al-Ghadir A.I., Alabdullatif .A, and Azmi A.: Gender Inference for Arabic Language in Social Media. International Journal of Knowledge Society (2014).

[12] Rangel, F., Rosso, P., Potthast, M., and Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Cappellato, L., Ferro, N., Goeuriot, L., and Mandl, T. CLEF 2017 Labs Working Notes. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017).

[13] Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., and Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G. J. F., Lawless, S., Gonzalo, J., Kelly, L., Goeuriot, L., Mandl, T., Cappellato, L., Ferro, N. (eds.): Experimental IR Meets Multilinguality, Multimodality, and Interaction. 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11-14, 2017, Proceedings.