Author Profiling in Arabic Tweets：An Approach based on Multi-Classification with Word and Character Features

Yutong Sun[1], Hui Ning[1], Kaisheng Chen[3], Leilei Kong[2,*], Yunpeng Yang[2], Jiexi Wang[2] and Haoliang Qi[2]

[1] Harbin Engineering University, Harbin, China
[2] Heilongjiang Institute of Technology, Harbin, China
[3] East China Normal University, Shanghai, China
Kongleilei1979@gmail.com

Abstract. This paper focuses on the author profiling task published in the FIRE 2019 (Forum for Information Retrieval Evaluation), which includes automatic identification of the age, gender, and language variety of Arabic tweets. We think the author profiling task as a multi-Classification problem. We have used word and character based on TFIDF features, learned the logistic regression classifier to predict the labels. In the final results, our proposed method shows a good performance in terms of age prediction, the accuracy rate is 0.6250. Additionally, we have obtained 0.5111 and 0.9604 accuracy for gender and language variety classifications respectively. In the experiment, We have used the different feature combination and adjusted the feature parameters to test the system. The combination of word and character features can improve the prediction accuracy and enhance the system performance significantly.

Keywords: Author Profiling , Logistic Regression , Word and Characters N-gram.

1 Introduction

With the continuous development of social media, the research of author profiling task has significant progress that has been made [1, 2, 3]. Author profiling task is to identify the user profiling aspects such as age, gender, and language variety ,among others. We formalized the author profiling task into a multi-classification problem. we have used word and character or their combination as features of learning the classifier. In order to extract the effective features, we have exploited TFIDF based method to filter features. In the paper, the model which has proposed is based on the Logistic Regression classifier, using word feature from unigram and character

* Corresponding author

features from bigram to 4-gram and its combinations as a standard of the label predictions. The final evaluation results show that in the model the age prediction accuracy can reach 0.6250. For the language variety, it is 0.9694, and 0.5111 accuracy for the gender.

2 Methods

2.1 Preprocessing

Firstly, we read and parsed all .xml documents, and combined the each author's tweets into a single text. Secondly, in this paper we have proposed the method based on text vocabulary to extract the corresponding features. So we filtered out the non-text content in document, such as @, emoticons and URL. Thirdly, we normalized the text content, removed the unnecessary spaces, tabs and punctuations.

2.2 Experimental Methods

Following the successful of author profiling system[4], we applied a model based on classification to build our system. The gender prediction task is a problem of binary classification, the age and language tasks are the multi-classification problems.

We compared to the Logistic Regression classifier and Linear SVC classifier, found the performance of LR classification is more stable and simple. So we chose LR as the final classifier. In terms of feature selection, we have used the character features from bigram to 4-gram ,word feature from unigram and its combinations of the features, exploited TFIDF method to extract more representative features. Giving a tern, to calculate its TF, IDF and DF values, combine TF and IDF as features and remove features of DF value which is lower than predefined minimum and higher than the predefined maximum.

The process of our method is shown in Fig. 1.

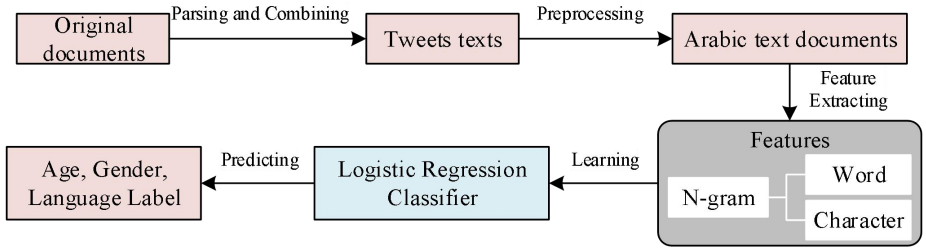

Fig. 1. The process of the proposed approach for Author profiling

Firstly, we preprocessed the training data set. Secondly, based on TF and IDF values, we extracted the classification features. In the experiment, we combined the character features from bigram to 4-gram with the word feature from unigram. Thirdly, we learned the LR classifier (default setting all parameters) based on the filtered features. Finally, we used the learning classifier to predict the test data set.

3 Experiments

3.1 Data Set Description

The corpus[2] of this task is consists of Arabic tweets, sign with age, gender, language variety labels. Data set divided into five groups, each group contains three languages, all of which belong to Arabic. In gender classification, the label including two types: male and female. Age label divided into three types: Under(< 25), Between(25-35) and Above(>=35).

Through by analyzing the corpus, we found the number of types is same in the labels and it is a balance state.

3.2 Evaluation Measures

The performance of author profiling approach is evaluated by the joint accuracy. The accuracy is defined as the ratio of the predicted correct number *Pc* to the total predicted number *Pt*.

3.3 Experimental Results

We split the training data, 80% for training and 20% for testing, to observe the different effect in the feature combinations. The experimental results are shown in Table 1.

Table 1. Experimental results with different feature combinations

Features	Gender	Age	Variety
word-unigram	0.8123	0.5648	0.9236
character-bigram	0.7821	0.5417	0.8823
word+char-bigram	0.8046	0.5872	0.9405
word+char-bigram-trigram	0.8058	0.6235	0.9423
word+char-bigram-trigram-4gram	0.8052	0.6148	0.9542

Table 1 shows that language variety predictions have the highest accuracy , about 95%. In order to identify age, we found that using the combination features is better than word and character alone. In gender, the difference of the experimental results using various features are seldom , the word unigram feature is slightly better than others.

Table 2 describes the final experimental results of the top three teams and our team. In the yutong.2 file, we have used the word unigram feature for gender and language variety ,the word + char-bigram -trigram combination for age. The age classification accuracy rate is shown in Table 3.

[2] https://www.autoritas.net/APDA/corpus/

Table 2. The final evaluation results

Team	Gender	Age	Variety	Joint
DBMS-KU.2(Top 1)	0.7944	0.5861	0.9722	0.4556
Nayel.1(Top 2)	0.8153	0.5708	0.9750	0.4486
Nayel.3(Top 3)	0.8014	0.5792	0.9708	0.4486
Yutong.2(Our team)	0.5111	0.6250	0.9694	0.3125

Table 3. The age accuracy of the top five groups

Age Group Ranking					
Team	Yutong.2	Yutong.3	Yutong.1	DBMS_KU.2	DBMS_KU.3
Accuracy	0.6250	0.6000	0.5875	0.5861	0.5819

Table 4 compares the effects of the parameters min_df and max_df in the TFIDF model .

Table 4. Results of different parameter values

Parameter Combination		Gender	Age	Variety
min_df=4	max_df = 0.7	0.7770	0.6041	0.9310
	max_df = 0.8	0.7772	0.6043	0.9312
	max_df = 0.9	0.7838	0.6154	0.9322
min_df=5	max_df = 0.7	0.7921	0.6126	0.9410
	max_df = 0.8	0.7944	0.6224	0.9412
	max_df = 0.9	0.8058	0.6235	0.9423

4 Conclusions

This paper presents the method based on multi-classification with word and character features for author profiling in Arabic tweets. In our method, we have chose word and character and their combinations as the features and classified the LR classifier. The final evaluation results show that the best performance is the combination features of gender and language (word unigram) + age (word + char-bigram-trigram). We have obtained 0.6250, 0.5111 and 0.9604 accuracy for age, gender and language variety classifications respectively. In the future work, we will consider the feature extraction of non-text content, and further improve the experimental performance.

Acknowledgments

This research was supported by the Social Science Fund of Heilongjiang Province of China (No.18TQB103).

References

1. Marquardt James, et al.: Age and Gender Identification in Social Media. In: CEUR Workshop Proceedings, vol.1180, pp. 1129-1136 (2014).
2. Michał Meina, Karolina Brodzi ń ska, Bartosz Celmer, Maja Czoków, Martyna Patera, Jakub Pezacki, Mateusz Wilk: Ensemble-based Classification for Author Profiling Using Various Features -Notebook for PAN at CLEF 2013. In: CLEF 2013 Evaluation Labs and Workshop-Working Notes Papers. Valencia, Spain (2013).
3. A. Pastor López-Monroy, Manuel Montes-y-Gómez, Hugo Jair Escalante, Luis Villaseñor-Pineda, Esaú Villatoro-Tello: Using Intra-Profile Information for Author Profiling-Notebook for PAN at CLEF 2014. In: CLEF 2014 Evaluation Labs and Workshop-Working Notes Papers. Valencia, Spain (2014).
4. Sharmila Devi V, Kannimuthu S, Ravikumar G, Anand Kumar M: KCE_DAlab @MAPonSMS-FIRE2018: Effective Word and Character-based Features for Multilingual Author Profiling. In: Working Notes for MAPonSMS at FIRE'18 -Workshop Proceedings of the 10th International Forum for Information Retrieval Evaluation, pp. 213-222. Gujarat, India (2018).
5. 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).