Author Profiling for Arabic Tweets based on n-grams

Ayoub Abbassi[1], Seifeddine Mechti [2], Lamia Hadrich Belguith[1], and Rim Faiz[3]
[1]ANLP Group MIRACL Laboratory,
FSEGS, University of Sfax
[2]LARODEC Laboratory, ISG of Tunis,
2000 Le Bardo, Tunisia
[3] LARODEC Laboratory, ISG of Tunis IHEC,
2016 Carthage, Tunisia
ayoub.abess@gmail.com,mechtiseif@gmail.com
l.belguith@fsegs.rnu.tn,Rim.faiz@ihec.rnu.tn

Abstract. This paper presents an approach for author profiling of an unknown users from their texts produced in social media. In particular, we address the identification of two profile dimensions: gender and language variety, of Arabic twitter users based on their tweets. Our approach focused on applying meta-classification technique on features extracted from tweets body. We explored two main sets of features which are character and word n-grams. The proposed approach allowed us to reach promising results for both language variety and gender identification

Keywords: Author profiling, Meta-classifier, N-gram features.

1 Introduction

The rapid growth of internet and computer technology during the last two decades makes humanity in front of an incredible increased amount of online data. According to internet live stats[1], in one second the Internet traffic is about 36,411 GB. This impressive amount of data -mostly of text type- are shared, published, and transit in a free (sometime in anonymous) way. In fact, an important portion of internet users are misrepresenting themselves while surfing in the net, therefore there are a need to deal with the data that come from unknown source.

Two main sectors are interested in knowing the potential source of data. First, the commercial sector where information such as age, gender, nationality, and native language about customers is of higher value for marketing intelligence. Second, the

[1] http://www.internetlivestats.com/

security sector that bear the burden to protect the internet from crime such as plagiarism and identity theft, etc.

Therefore, research community promotes researchers to discover and develop effective methods and techniques in related fields such as plagiarism detection and author profiling.

This work is made in the context of the participation of the Author Profiling task in the PAN17 shared task[2]. In particular, we focus on identifying the gender and Language variety of Arabic users from their twitter tweets.

2 Dataset Description

We used training dataset provided by PAN clef 2017 to train our proposed system. We participated in the author-profiling task for the Arabic subtask. The training dataset is composed of Twitter tweets and annotated with authors' gender and their specific variation of their native language. A detailed statistics of the used dataset is given in Table 1.

Table 1: Distribution of data for Arabic author-profiling task in the PAN17 training corpus

Task		Number of files
language variety	Egypt	600
	Gulf	600
	Levantine	600
	Maghrebi	600
Gender	Male	1200
	Female	1200

As the above table shows, it is clear that the turning dataset is well distributed across classes. However, the analysis reveals that some documents are written in Modern Standard Arabic, not in one of the Arabic varieties [1], which can affect the performance of our system.

3 System Architecture

Our proposed system is divided into three steps: pre-processing, feature extraction and Classification. Firstly, in the pre-processing step, we prepare the input data to be used in the next step. Then, in the feature extraction phase, we extract the set of features that seem to be useful for the task. Finally, we generate the classification model. This model will be used to predict the class of new document.

[2] http://pan.webis.de/clef17/pan17-web/

3.1 Pre-processing

As the input dataset is basically composed of Twitter tweets, these tweets have the nature of being noisy including a lot of useless data such as links, tags, emoticons, etc. Thus they can't be exploited directly. The idea is to remove these noisy data. However, in stand of looking for the variety of noisy, we simply extracted the Arabic text. The example below shows a tweet before and after prepossessing step.
Example:
Input tweet: "#thanx @alaakarmus 😳 كان في تحدي ع سؤال وانا ربحت حصلت شكلاطه 😁 https://t.co/UySVCM1qwm https://t.co/wKBUpGXmZo"

Tweet after extract the Arabic text: "كان في تحدي ع سؤال وانا ربحت حصلت شكلاطه"

3.2 Features extraction

We extracted tow n-gram feature types, namely 'character n-grams' and 'word n-grams'. Accordingly, we generated two sets of features for each input document. For each individual feature, we calculated the Inverse Document Frequency (IDF) with which it appears. The documents are then represented as TF-IDF matrix.

Given a text extracted from tweets, the set of n-grams was extracted by moving a window of n cases across the text body. For example, based on the word as a feature, word n-grams means all the n consecutive words in the text.

For the previous tweet "كان في تحدي ع سؤال وانا ربحت حصلت شكلاطه", the word n-gram model is illustrated in Table 2.

Table 2: Example of word n-gram model

N-gram model	Example
Word-based 1-grams	كان, في, تحدي , ع, سؤال, . . .
Word-based 2-grams	كان في, في تحدي, تحدي ع, . . .
Word-based 3-grams	في تحدي ع, كان في تحدي , . . .

3.3 Classification

Once the documents have been transformed to their new representation, they will be used as input to train the classifier. Training the classifier is the main key of this work, we apply a meta-classifier technique known as 'stacking' [2] to generate the finale module, which will be used to predict the correct class of unlabelled document. Stacking consist in combining several base classifiers of different type, in our case, we use the three most popular machine-learning algorithm (Support vector machines, Decision trees and Naive Bayes) [3].The principle of this technique is illustrated in the following figure:

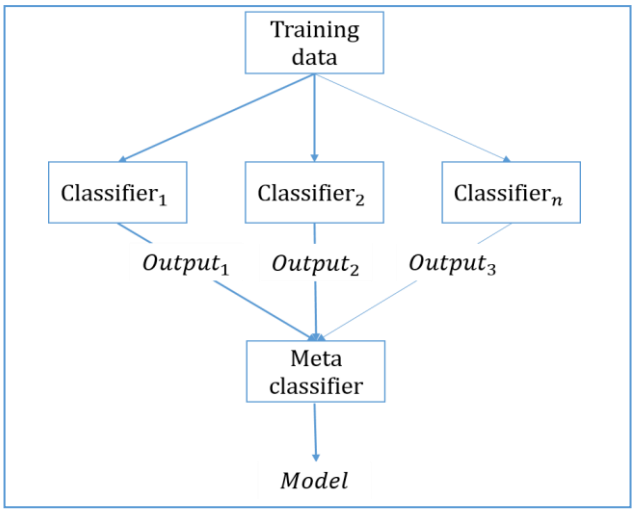

Figure 1: Stacking principle

4 Results

We carried out several series of experiments in order to evaluate the performance of the classifiers mentioned before individually and combined, using different sets of features. Table 3 and Table 4 show the result of our experiments:

Table 3: Language variety results for PAN 17 Training Dataset

Classification technique		Features set		
		Word n-grams	Character n-gram	Combined
individual	Decision trees	27.1	28.0	29.2
	Naive Bayes	26.0	26.5	27.3
	SVM	29.0	28.2	31
combined	Stacking	**34.0**	**31.3**	**33.0**

Table 4: Gender results for PAN 17 Training Dataset

Classification technique		Features set		
		Word n-grams	Character n-grams	Combined
individual	Decision trees	55.0	56.0	57.0
	Naive Bayes	56.2	54.2	56.0
	SVM	58.1	57.0	59.3
Combined	Stacking	**61.1**	**59.0**	**63.2**

For gender dimension, the best accuracy is 59.3 which is obtained using SVM, in the case of individual classifier, and **63.2** using Stacking as classification technique. These results are obtained by combining all features together. Such results confirm our finding [4] of the outperformance of SVM compared with other learning algorithms in author profiling problem.

However, for language variety, the result obtained using word n-grams outperformed those obtained using character n-grams or combination with 34 of accuracy. This is obtained by combining (Stacking) the performance of classifier.

5 Conclusion

In this paper we described our approach of profiling the users of Twitter based on meta-classifier trained on n-grams features. In particular, we focused on the identification of gender and language variety of Arabic users. We found out that combining the n-grams- features in a meta-classification process allowed us to achieve higher results, on the tow tasks. The best result are obtained using word n-grams for language variety detection and using all features combined for gender detection.

Reference

1. A. FARGHALY and K.SHAALAN “Arabic natural language processing: Challenges and solutions”, in proceedings of ACM Transactions on Asian Language Information Processing (TALIP), vol. 8, no 4, 2009.
2. S.B. Kotsiantis., I. Zaharakis, and P. Pintelas. “Supervised machine learning: A review of classification techniques”, p.3-24, 2007.
3. K. Vandana and M. Namrata, “Text classification and classifiers: a survey” Artificial Intelligence & Applications, vol. 3, n. 2, 2012.
4. S., Mechti, A., Abbassi, L. H. Belguith, R., Faiz,, C. “An empirical method using features combination for Arabic native language identification”, in proceedings of the 3th International Conference of Computer Systems and Applications (AICCSA),2016.