Author Profile Prediction Using Trend and Word Frequency Based Analysis in Text

Notebook for PAN at CLEF 2017

Jamal Ahmad Khan

Department of Computer Science and Software Engineering, International Islamic University, Islamabad, Pakistan

J_Ahmadkhan@Yahoo.com

Abstract. PAN 2017 Author Profiling task include two target predictions, one is to predict the gender of text authors and second is to predict the language variety. The presented approach analyzed trends and topics followed in training dataset e.g. Authors discussing Politics, Tech, Religion, Nature etc. in their respective tweets. Along with that single words and word pair frequencies were also taken into account. A cross-lingual, general, simple and flexible approach was created that could be applied over all languages without any changes according to each task language.

1 Introduction

Social media nowadays has become an important indicator for trend based analysis and become a reflection of not only our personality but also a powerful way of expression. We express our feelings about different events or persons through text in the form of short expressive sentences that carry information about a person's likes, dislikes, beliefs and interests.

The benefits of Author profiling tasks include the right idea of author's gender, age, regions and traits, and this possibility of finding people's traits is of growing importance [1] especially for the business organizations that are interested to invest in areas following trend based analysis over different types of social media tools like Twitter, Facebook, WhatsApp, SnapChat etc. where one or different types of socializing tools are more used in different regions of world and marketing organizations may also be interested to find out the reviews submitted by people of different genders and origins in order to make strategic decisions [2].

Also the task of Author profiling deals with the aspects of forensics where the gender and specific regions of countries are important, and this task of author profiling task of PAN@CLEF 2017 [3] deals with this aspect where one can create a prediction model for two aspects of a tweeted text; one is the gender prediction having two classes male or female and second is to predict language variety which in turn can predict the specific region or regional affiliation of the person tweeting the text.

By identifying general topics discussed in social networks can provide us better understanding of collective interests [4] and trends and locations (location in the sense of language variety) are interdependent or correlated [5]. The Submitted system uses the approach of trend/topic based analysis combining with it are the word frequencies and word pair frequencies to classify language varieties and gender.

Five different trends including Politics, Technology, Nature, Travel and Religion were analyzed. Frequent words appearing in texts of different languages were also analyzed, because people use social media, like twitter to follow different topics and issues that are trending in their respective geographical areas and of their specific interests. They also use specific expressive words that are common or more frequent in their regional areas, so term frequencies and word pair frequencies were also taken into account because both are an important indicator for text based analysis [6]. Hence both these aspects can be handy in predicting the language variety and gender of different persons.

2 Dataset

Like previous year, the training dataset of PAN CLEF 2017 for author profiling task contained xml files for each language representation with an exception that demographic traits in current dataset are processed side by side [3]. A total of four languages were presented including Arabic, English, Portuguese and Spanish. All languages have a range of varieties shown in Table.1. Each language variety represents a region in the world where it is used and each variety had 600 xml documents the training dataset. Also an equal gender representation for each language was provided in training dataset as shown in Table. 2.

Table 1. XML Documents Representation for Language variety

LANGUAGE	*VARIETIES*	*Total Examples*
Arabic	Egypt, Gulf, Levantine, Maghrebi	600 each
English	Australia, Canada, Great Britain, Ireland, New Zealand, United States	600 each
Portuguese	Brazil, Portugal	600 each
Spanish	Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela	600 each

Table 2. XML Documents Representation for Task for Gender

LANGUAGE	*MALE*	*FEMALE*	*Total Documents*
Arabic	1200	1200	2400
English	1800	1800	3600
Portuguese	600	600	1200
Spanish	2100	2100	4200

3 System Methodology

In order to fulfill the task of Author Profiling at PAN CLEF 17, a general system was built so that it may be able to perform equally irrespective of language shifts from one to another or having to jump from one basic approach to another. By performance it's meant that the system would predict both gender and language variety for each author in Twitter corpus without using complex linguistic analysis methods like Part of Speech (POS) analysis or stylistic approaches from variety to variety and for gender prediction as well.

A simple scoring approach is used where each document is assigned scores in order to classify it as one of the classes that are created as preprocessing of training dataset. Each class represents exactly one variety in each language representation with two subclasses under each variety representing gender i.e. male and female.

The experimental setup is divided into following steps

- Trend word-lists Preparation
- Dataset Preprocessing
- Language Variety Classification
- Gender Classification

3.1 Trend word-lists Preparation

A total of five general topics were chosen including Politics, Technology, Religion, Nature and Romance to create lists of 500 words in each language i.e. Arabic, English, Portuguese and Spanish and for each chosen topic. These word lists are referred as trend list T_{Li} in presented system, where L is the main language variety and i is variable representing specific trend. Different sources were used in this regard.

3.2 Dataset Preprocessing

The system performs following steps in order to preprocess data for creation of language variety classes and gender subclasses.

1. All xml based twitter text was extracted from each language variety training dataset discarding Hash-tags, html/xml tags, html/web links and extra white

spaces in order to get plain text from each file and all text was combined as a single document D_V representing a language variety. Where v is the represented language variety in document.

2. All extracted text was combined in order to find top 100 most frequent words TF_{V_t} and word pairs PF_{V_t} each, where t is the id for specific term

$$Term\ Frequency\ (TF_{V\_t}) = \frac{Total\ Term\ Occurences}{Text\ Length} \quad (1)$$

$$Word\ Pair\ Frequency\ (PF_{V\_t}) = \frac{Total\ Word\ Pairs}{Text\ Length - 1} \quad (2)$$

3. Each common term occurring in both trend list T_{Li} and document D_V is scored according to term occurrence and Trend score S_{v_i} for each document is calculated.

$$S_{v_i} = \frac{Trend\ Score}{Text\ Length} \quad (3)$$

4. Steps 2 and 3 are repeated in order to find gender based term frequency, word pair frequency and Trend scores calculation under each language variety v.

5. Steps 1 to 4 are repeated for each Language L.

6. Hence we have created different language variety classes C_V and gender subclasses C_{Vg}, where g represents the gender i.e. male or female as shown in figure a; where $TF = TF_{V_{1....100}}$ and $PF = PF_{V_{1....100}}$

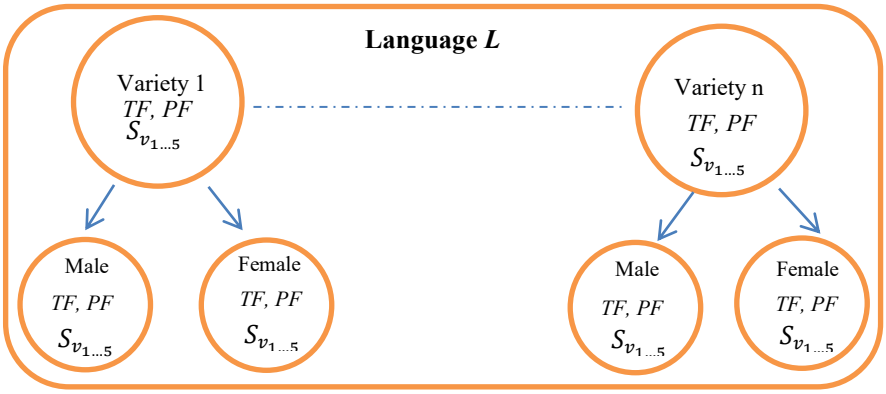

Figure a. Language Variety Classes and Gender Subclasses for a Language L

7. The System loops through each language folder of dataset and every document is processed in the same manner as a separate class C_{u_id} with unknown variety label *u* and Author Identity *id* having its own TF_{u_t}, PF_{u_t} and S_{u_i}

3.3 Language Variety Classification

Following Steps are taken by the system in order to assign the class C_{u_id} with a language variety label while looping through each variety class C_V .

1. For each single word and word pair common in each C_V and C_{u_id} score S_{V_id} is increased for class C_{u_id} as shown in equation 4.

$$S_{V_{id}} = \sum_{t=1}^{100}(TF_{u_t} + TF_{v_t}) + (PF_{u_t} + PF_{v_t}) \quad (4)$$

2. For each trend score S_{v_i} in C_V and S_{u_i} in C_{u_id} absolute difference D_{V_id} between both is calculated as shown in equation 5.

$$D_{V\_id} = \sum_{i=1}^{5} abs(S_{v_i} - S_{u_i}) \quad (5)$$

The least trend difference score D_{V_id} is added to class variety score S_{V_id} and in this way a class C_{u_id} with unknown language variety is assigned a language variety *v* having highest score S_{V_id} .

3.4 Gender Classification

Once the Language variety of Document class C_{v_id} has been decided, the system repeats similar steps as shown in section 3.3 to find out the gender for class C_{v_id} but this time the system has to decide among only two gender subclasses of variety class *v*. The gender subclass having highest score is assigned to class C_{v_id} .

4 Results

A 10-fold cross validation criteria was used during the initial validation of the system and no different approach was used for prediction of language varieties among main languages or gender prediction as shown in above section. Once all parameters were adjusted during system validation over training datasets, the system was ready to run over full training and test datasets provided at TIRA [7] in order to predict both gender and language variety.

Following are the evaluator results shown in table 3.

Table 3. Training and Test Results over Author Profiling Task Dataset

Corpus	*Gender*	*Variety*	*Both*
PAN 17 Training dataset Arabic	0.5942	0.6079	0.3788
PAN 17 Training dataset English	0.6578	0.3017	0.2094
PAN 17 Training dataset Portuguese	0.6392	0.8975	0.5750
PAN 17 Training dataset Spanish	0.6307	0.3519	0.2193
PAN 17 Test dataset Arabic	0.58630	0.58440	0.36500
PAN 17 Test dataset English	0.66920	0.27790	0.19000
PAN 17 Test dataset Portuguese	0.61000	0.90630	0.54880
PAN 17 Test dataset Spanish	0.63540	0.34960	0.21890

The results consistency distributed over different languages in above table for both training and test datasets clearly indicate the generalization of used system; it's because of the designed approach that denied over-fitting while training stages. The performance of model is good where it has to predict among smaller group of variables either it be gender or language variety, a typical example of which is the results of Portuguese language where the system's prediction accuracy is highest for both gender and variety because each has only 2 values to predict from. As the number of values in language variety increases, the performance of system decreases and hence the overall results are affected.

4 Conclusion

In this paper a new and generalized approach is presented which may be able to perform with same attributes over all given languages in dataset without changing its modes or functionality with language change. The results however indicate a major flaw in the system over which more work is required. This flaw is the decrease in prediction when there is an increase in number of language varieties. This can however be improved either by increasing the quantity of single words and word pairs while the creation of variety classes or by increasing the trends.

The future work will certainly emphasize to improve the prediction eminence of the system by following the suggestions discussed above.

References

1. Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Pottast, Benno Stein. Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: (Eds.) CLEF Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 1609, pp. 750-784 (2016)
2. Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstatios Stamatatos, Giacomo Inches. Overview of the Author Profiling Task at PAN 2013. In: (Eds.) Notebook Papers of CLEF LABs and Workshops. CEUR-WS.org, vol. 1179 (2013)
3. Francisco Rangel, Paolo Rosso, Martin Potthast and Benno Stein. Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: (Eds.) CLEF Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings. CEUR-WS.org, vol. 10456 (2017)
4. Ceren Budak, Divyakant Agrawal, Amr El Abbadi. Structural Trend Analysis for Online Social Networks. Proceedings of VLDB Endowment (PVLDB), 4(10):646656 (2011)
5. Ceren Budak, Theodore Georgiou, Divyakant Agrawal, Amr El Abbadi. GeoScope: Online Detection of Geo-Correlated Information Trends in Social Networks. PVLDB 7(4): 229-240 (2013)
6. Yuen-Hsien Tseng, Chi-Jen Lin, Yu-I Lin. Text mining techniques for patent analysis. Information Processing & Management. Volume 43, Issue 5, Pages 1216–1247 (Sep, 2007)
7. [Online] http://www.tira.io/tasks/pan/ (2017)