Language Variety and Gender Classification for Author Profiling in PAN 2017

Notebook for PAN at CLEF 2017

Alexander Ogaltsov and Alexey Romanov

Antiplagiat CJSC,
Higher School of Economics, Moscow Institute of Physics and Technology
ogaltsov@ap-team.ru, avogaltsov@edu.hse.ru, alexey.romanov@phystech.edu

Abstract We describe the method of Author Profiling task. The task deals with study of profile aspects like gender and language variety. We explore an approach of using high-order char n-grams as features and logistic regression as a classifier for all subtasks. This approach appears to be simple and effective for the task. We also investigated feature importances and low-dimensional embeddings of the data.

1 Introduction

Author profiling task considers different profile dimensions of the author of the text. This year shared task [12][11] is focusing on gender and language variety. Previous competitions explored properties like gender, age group [13] and personal traits [8]. This task is interesting from both industrial and scientific points of view. Applications like accurate advertising targeting, security and forensic fields make this task highly relevant for practice. Also, the task can be considered as a tool for filling missing information about a person in some political or demographic research. Research community also pays attention to the task special track of PAN [7] shared task is held since 2013. Each year contributed a new language or new profile dimension to classify. The common part of all years was gender identification. The first task was on blog data in Spanish and English [10]. Competition in 2014 concentrated on different sources like reviews, tweets etc. [9]. The task of 2015 extended by additional languages and real-valued personal traits [8]. The main characteristic of the most recent shared task was cross-genre. The target was to develop a model such that it will be robust to the domain of data [13]. Since gender identification was presented in all previous competitions, there were many tested approaches. The main features were n-grams and various text statistics [4].

Language variety task was first to appear at PAN 2017, but there were language variety detection competitions like *Discriminating between similar languages and national language varieties* (DSL) 2016 [1]. Winning approach of this contest used char n-grams in wide range (1-7) with a linear classifier [3]. We used this method not only for language variety task but also for gender classification. A new feature of the current shared task is language variety. Each language has several variants. For instance, we have two several Portuguese: Brazil variant and European one. The task is to distinguish one

from another. Languages and their varieties can be found in Table 1. Our approach tries to automatically extract features for each of variant Portuguese, English, Spanish and Arabic without any linguistic knowledge. We use char n-grams as features and logistic regression as a classifier. Evaluation metric is accuracy for both subtasks.

Language	Variety
Portuguese	Portugal, Brazil
English	Australia, Canada, Great Britain, Ireland, New Zealand, United States
Spanish	Argentina, Chile, Colombia, Mexico, Peru, Spain, Venezuela
Arabic	Gulf, Levantine, Maghrebi, Egypt

Table 1. Languages and Varieties

2 Methodology

This section is about our approach to current PAN Author Profiling task. First, we briefly discuss preprocessing steps. Then, we describe how we construct the feature space. Finally, we explain our choice of logistic regression as our classifier.

2.1 Preprocessing

We did not perform any preprocessing like removing hashtags, HTML tags and urls, because we considered it as potentially informative features.

2.2 Classification

Our main assumption was to consider all short texts written by a single author as an object in machine learning task formulation. We formulated the problem as classification task with two or more classes depending on language (Table 1). If language has more than two varieties we used "one versus other" scheme.
Let dataset

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, \quad i = 1, \ldots, m,$$

to be consisted of pairs "object-class", $\mathbf{x}_i \in \mathbb{R}^n$. Each object $\mathbf{x}_i$ has one of Z class labels $y_i \in \mathbb{Y} = \{1, \ldots, Z\}$. We have to find mapping $\hat{f} \in \mathfrak{F} : \mathbb{R}^d \to \mathbb{Y}$, which minimizes empirical risk on dataset $\mathfrak{D}$:

$$\hat{f} = \arg\min_{f \in \mathfrak{F}} \sum_{\mathbf{x}_i, y_i \in \mathfrak{D}} [f(\mathbf{x}_i) \neq y_i],$$

where $\mathfrak{F}$ – family of models.
Feature space was constructed such that for each language corpus we performed counting of character level n-gram in some range. This counts were used as features. The

number of authors and features for different tasks can be founded in Table 2. One can see that the data is quite sparse. Density distribution of non-zero n-grams for Portuguese is shown in Figure 1. We did not used higher-order n-grams because of RAM restric-

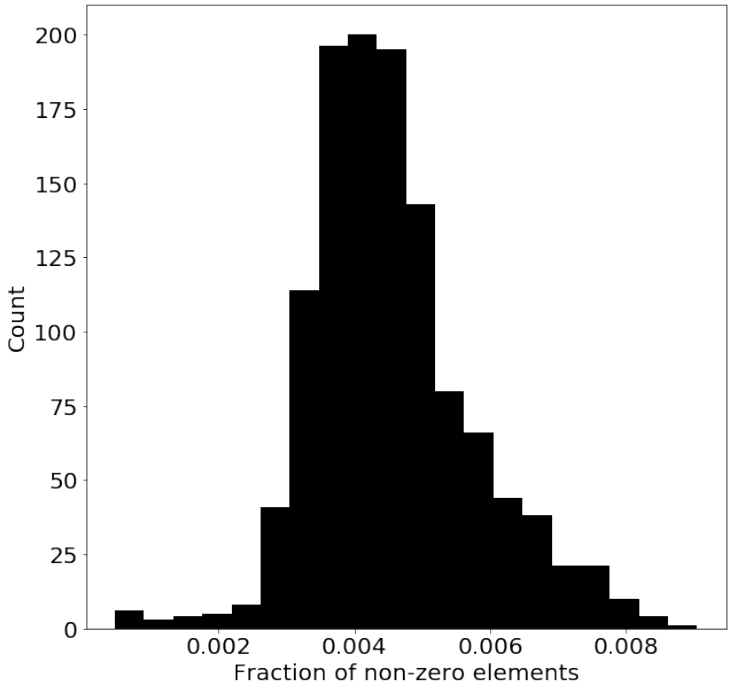

Figure 1. Density distribution for Portuguese.

Language	n-gram range	# of objects	# of features
Portuguese	2-6	1200	3379797
English	2-5	3598	5922462
Spanish	2-5	4198	6030424
Arabic	2-6	2375	6655335

Table 2. Languages and Varieties

tions, although [3] reported quality to increase up to 7 char n-gram level.
We performed classification by means of logistic regression model with regularization parameter $C = 1$. Our choice was justified by the fact that logistic regression has high bias and low variance.

3 Evaluation

In this section we describe our results during cross-validation and on the test set. Next we present embedding of the data in low-dimensional space. Finally, we discuss about feature importances of our classifier.

3.1 Results and Data Visualization

Evaluation metric this task is accuracy:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + TF}$$

We evaluated quality of gender and language variety subtasks separately by using cross-validation scheme with five folds. Results can be found in Table 3.
Example ROC-curve for language variety classification of Portuguese is shown at Figure 2. FPR and TPR are false positive rate and true positive rate respectively with various classification threshold. We evaluated test scores via TIRA. [6]

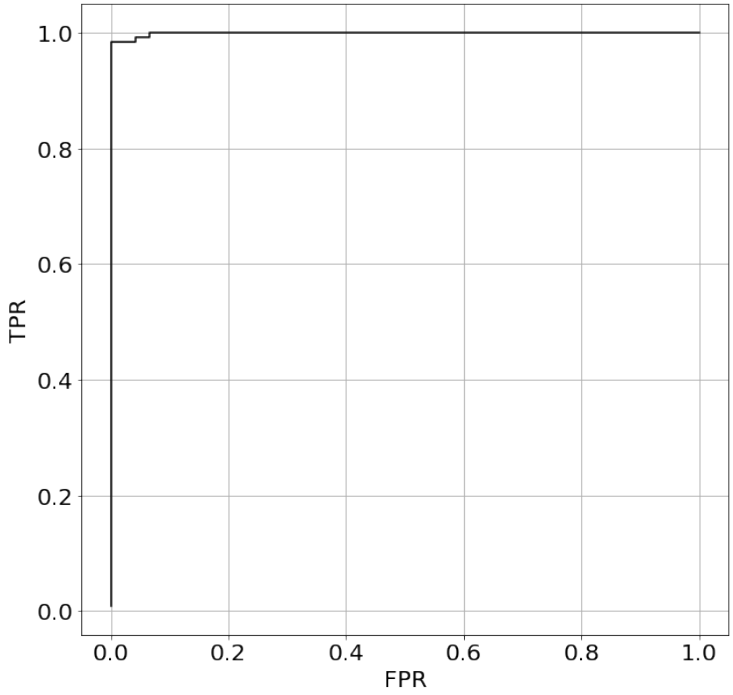

Figure 2. ROC-curve for Portuguese language variety.

Language	CV gender acc.	CV variety acc.	Test gender acc.	Test variety acc.
Portuguese	0.8025	0.9850	0.7988	0.9725
English	0.7918	0.7913	0.7875	0.8092
Spanish	0.7456	0.8892	0.7600	0.8989
Arabic	0.7263	0.7739	0.7213	0.7556

Table 3. Evaluation

It was interesting to see how data is located in a feature space. To do so we exploited modern dimensionality reduction and data visualization techniques. Our choice of algorithm was t-SNE [2] since it reported to be fast when the number of objects is small and tends to efficiently preserve local structure of the data. Also, Python scikit-learn [5] implementation of the algorithm supports sparse matrices as an input. Example for Portuguese authors is at Figure 3. Unfortunately, axes of this algorithm have no clear interpretation.

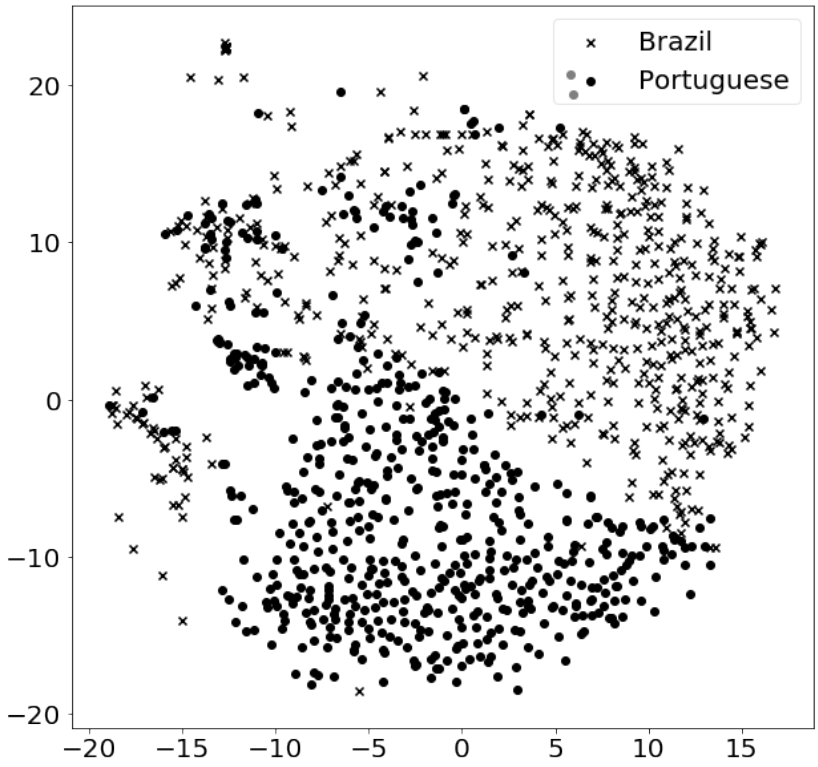

Figure 3. t-SNE data visualization for Portuguese.

3.2 Feature importances

We investigated absolute values of coefficients of our model for Portuguese language variety. This values can be considered as feature importances (Figure 4). Axis x means position in array of linear regression coefficients sorted in descending order. Axis y is absolute value of the coefficient. One can see that on the one hand feature coefficients have pretty low magnitude, but on the other hand there is group of features with relatively high importance.

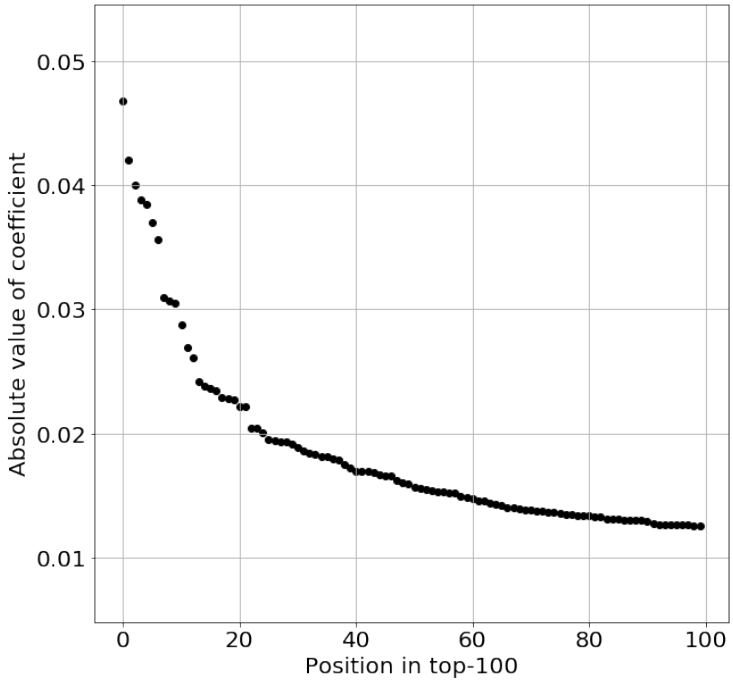

Figure 4. Feature coefficients for Portuguese language variety.

4 Conclusion and Future Work

We explored a simple and robust method for gender and language variety classification for PAN17 Author Profiling task. It turned out that high-order char n-grams are good features that are easy to generate with no need of handcrafting or expert linguistics knowledge. The main disadvantage of such features is that this is almost impossible to perform error analysis. We trained logistic regression classifier for both subtasks and evaluated accuracy measure. We will explore effects on quality measure due to adding even more n-grams.

References

1. Dsl shared task 2016 (2016), http://ttg.uni-saarland.de/vardial2016/dsl2016.html
2. van der Maaten, L., Hinton, G.: Visualizing high-dimensional data using t-sne. Journal of Machine Learning Research 9: 2579-2605 (Nov 2008)
3. Malmasi, S., Zampieri, M., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In: Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3). pp. 1–14. The COLING 2016 Organizing Committee, Osaka, Japan (December 2016), http://aclweb.org/anthology/W16-4801
4. Modaresi, P., Liebeck, M., Conrad, S.: Exploring the Effects of Cross-Genre Machine Learning for Author Profiling in PAN 2016—Notebook for PAN at CLEF 2016. In: Balog, K., Cappellato, L., Ferro, N., Macdonald, C. (eds.) CLEF 2016 Evaluation Labs and Workshop – Working Notes Papers, 5-8 September, Évora, Portugal. CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
5. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
6. Potthast, M., Gollub, T., Rangel, F., Rosso, P., Stamatatos, E., 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 (Sep 2014)
7. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Jones, G., 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 Initiative (CLEF 17). Springer, Berlin Heidelberg New York (Sep 2017)
8. Rangel, F., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd Author Profiling Task at PAN 2015. In: Cappellato, L., Ferro, N., Jones, G., San Juan, E. (eds.) CLEF 2015 Evaluation Labs and Workshop – Working Notes Papers, 8-11 September, Toulouse, France. CEUR-WS.org (Sep 2015)
9. Rangel, F., Rosso, P., Chugur, I., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2nd Author Profiling Task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers, 15-18 September, Sheffield, UK. CEUR-WS.org (Sep 2014)
10. Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G.: Overview of the Author Profiling Task at PAN 2013. In: Forner, P., Navigli, R., Tufis, D. (eds.) CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, 23-26 September, Valencia, Spain (Sep 2013)
11. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs
12. Rangel F., Rosso P., P.M.S.B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: CLEF 2017 Labs and Workshops, Notebook Papers. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)

13. Rangel Pardo, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th Author Profiling Task at PAN 2016: Cross-Genre Evaluations. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/