Language- and Subtask-Dependent Feature Selection and Classifier Parameter Tuning for Author Profiling

Notebook for PAN at CLEF 2017

Ilia Markov, Helena Gómez-Adorno, and Grigori Sidorov

Instituto Politécnico Nacional,
Center for Computing Research,
Mexico City, Mexico
imarkov@nlp.cic.ipn.mx, helena.adorno@gmail.com, sidorov@cic.ipn.mx

Abstract We present the CIC's approach to the Author Profiling (AP) task at PAN 2017. This year task consists of two subtasks: gender and language variety identification in English, Spanish, Portuguese, and Arabic. We use typed and untyped character n-grams, word n-grams, and non-textual features (domain names). We experimented with various feature representations (binary, raw frequency, normalized frequency, log-entropy weighting, tf-idf), machine-learning algorithms (liblinear and libSVM implementations of Support Vector Machines (SVM), multinomial naive Bayes, ensemble classifier, meta-classifiers), and frequency threshold values. We adjusted system configurations for each of the languages and subtasks.

1 Introduction

Author Profiling (AP) is the task that aims at identifying author demographics basing on the analysis of text samples. The AP methods contribute to marketing, security, and forensic applications, among other. From the machine-learning perspective, the task is viewed as a multi-class, single-label classification problem, when the automatic methods have to assign class labels (e.g., male, female) to objects (text samples). The Author Profiling task at PAN 2017 [10,13] consists in predicting gender and language variety on a corpus composed of Twitter messages in English, Spanish, Portuguese, and Arabic.

According to the AP task literature, combinations of character n-grams with word n-gram features have proved to be highly discriminative for both gender and language variety identification, regardless of the language the texts are written in or the genre of the texts [12,11,14,16]. In this study, we use combinations of typed (introduced in [15]) and untyped character n-grams with word n-gram features, and exploit domain names as non-textual features.

We examine various feature representations (binary, raw frequency, normalized frequency, log-entropy weighting, tf-idf), machine-learning algorithms (liblinear and libSVM implementations of Support Vector Machines (SVM), multinomial naive Bayes, ensemble classifier, meta-classifiers), and fine-tune the feature set for each of the targeted languages and subtasks.

2 Experimental Settings

The Author Profiling task at PAN 2017 [13] consisted in predicting gender and language variety in Twitter. The training corpus covers the following languages and their varieties:[1]

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

In order to determine the best system configurations for each of the considered languages, we conducted experiments on the provided PAN AP 2017 training dataset under 10-fold cross-validation.

The examined features, machine learning algorithms, feature representations, and threshold values are shown in Table 1.

Table 1. Examined system configurations.

Features	**ML algorithm**	**Feature representation**	**Frequency threshold**
Typed char. n-grams	Liblinear	Binary	1
Untyped char. n-grams	LibSVM	Raw freq.	2
Word n-grams	Multinomial naive Bayes	Normalized freq.	3
Lemmas	Ensemble	Log entropy	5
Domain names	Meta-classifiers	Tf-idf	10
			20
			30

Typed character n-grams, that is, character n-grams classified into 10 categories based on affixes, words, and punctuation were introduced by Sapkota *et al.* [15]. In our approach, we used the modified version of typed character n-grams as proposed in [8]. We examined typed character n-grams with n varying between 3 and 4. These features have shown to be predictive for both gender [7] and language variety identification [2]. Untyped character n-grams correspond to the more common approach of extracting n-grams without dividing them into different categories. In this work, we examined untyped character n-grams with n varying between 3 and 7.

We evaluated the performance of word unigrams (henceforward, words) when including and excluding punctuation marks and several implementations of word 2- and 3-grams: including and excluding punctuation marks, with and without splitting by a full stop.

The performance of each of the feature sets described above was evaluated separately and in combinations.

[1] Detailed description of the PAN Author Profiling 2017 corpus can be found in [13].

We applied several pre-processing steps: removed @mention instances, picture links, and URL mentions. We used the information regarding the particular domain name in order to form our feature set of domain names (e.g., https://www.instagram.com $\rightarrow$ instagram $\rightarrow$ feature set of domain names).

We examined the performance of the machine learning classifiers, shown in Table 1, using their scikit-learn [1] implementation. These classification algorithms are considered among the best for text classification tasks [14,5,16,6]. We evaluated the performance of each of the classifiers separately, as well as examined several ensemble setups and meta-classifiers as described in [4].

The most appropriate frequency threshold values were selected for each of the languages based on grid search. The following frequency threshold values were examined: 1, 2, 3, 5, 10, 20, 30, that is, we considered only those features whose frequency in the entire corpus is higher than the examined threshold value.

Table 2 shows the early bird system configurations. Here, word features contain punctuation marks; word 2-grams are splitted by a full stop and punctuation marks are excluded. 30 most frequent domain names were used for English and Spanish, 16 for Portuguese, and 7 for Arabic. As machine-learning algorithm, we used liblinear classifier with 'ovr' multi-class strategy and default parameters, which showed high results across all the targeted languages. Ensemble and meta-classifiers showed similar results; however, were discarded due to their high computational costs. For our early bird submission, we adjusted system configurations for each of the languages and used the highest average results for the both subtasks.

Table 2. Early bird system configurations.

Language	**Subtask**	**Features**	**Feature representation**	**Frequency threshold**
English	Gender and Variety	Untyped char. 5-grams, words, word 2-grams, domain names	Binary	10
Spanish	Gender and Variety	Untyped char. 4-grams, words, word 2-grams, domain names	Binary	10
Portuguese	Gender and Variety	Typed char. 4-grams, untyped char. 5-grams, words, word 2-grams, domain names	Binary	10
Arabic	Gender and Variety	Untyped char. 6-grams, words, word 2-grams, domain names	Binary	10

For our final submission, we adjusted system configurations for each of the subtasks within each language. First, we selected the most predictive feature combination and the best performing feature representation for each of the subtasks. Word features included punctuation marks, while word 2- and 3-gram implementations varied depending on the language and subtask. Then, we selected the optimal threshold values that were the same for the both subtasks within each language. We also filtered out the features that occurred in only one document in the corpus. Finally, we selected the optimal liblinear classifier parameters: penalty parameter (C), loss function (loss), and tolerance for stopping criteria (tol) based on grid search. The best final system 10-fold cross-validation results were obtained with the configurations shown in Table 3.

Table 3. Final system configurations. Thr. corresponds to frequency threshold; typed and untyped n-grams – typed and untyped character n-grams.

Language	**Subtask**	**Features**	**Feature representation**	**Thr.**	**Liblinear classifier parameters**
English	Gender	Typed 3-grams, untyped 3- and 7-grams, words, word 3-grams	Binary	5	C: 0.01 loss: squared_hinge tol: 0.0001
	Varierty	Typed 3-grams, untyped 4- and 7-grams, words, word 3-grams	Log entropy		C: 10.0 loss: hinge tol: 0.0001
Spanish	Gender	Untyped 3- and 5-grams, words, word 3-grams	Binary	3	C: 0.01 loss: squared_hinge tol: 0.0001
	Varierty	Typed 4-grams, untyped 3- and 5-grams, words, word 3-grams			C: 0.01 loss: hinge tol: 0.0001
Portuguese	Gender and Variety	Typed char. 4-grams, untyped char. 5-grams, words, word 2-grams, domain names	Binary	10	C: 1.0 loss: squared_hinge tol: 0.0001
Arabic	Gender	Typed 3-grams, untyped 6-grams, words, word 2- and 3-grams, domain names	Binary	3	C: 1.0 loss: squared_hinge tol: 0.0001
	Varierty	Untyped 4- and 6-grams. words, word 2- and 3-grams, domain names	Log entropy		C: 0.1 loss: hinge tol: 0.0001

3 Results

The early bird 10-fold cross-validation (10FCV) results in terms of classification accuracy on the PAN Author Profiling 2017 training corpus and the number of features (N) for each language are shown in Table 4. Table 5 presents the results obtained on the PAN Author Profiling 2017 test dataset evaluated using TIRA evaluation platform [9].

Table 4. Early bird 10FCV accuracy on the PAN AP 2017 training corpus.

Language	Gender	Variety	N
English	0.8047	0.8203	265,495
Spanish	0.7933	0.9517	181,997
Portuguese	0.8425	0.9875	119,382
Arabic	0.7817	0.8012	200,478

Table 5. Early bird accuracy on the PAN AP 2017 test set.

Language	Gender	Variety	Joint
English	0.7929	0.8225	0.6504
Spanish	0.7986	0.9511	0.7621
Portuguese	0.8125	0.9825	0.7963
Arabic	0.7625	0.7900	0.6256

The final system results on the PAN Author Profiling 2017 training corpus under 10-fold cross-validation and on the PAN Author Profiling 2017 test dataset are shown in Tables 6 and 7, respectively. N corresponds to the number of features.

Table 6. Final system 10FCV accuracy on the PAN AP 2017 training corpus.

Language	Gender	N	Variety	N
English	0.8211	734,457	0.8719	837,437
Spanish	0.8000	658,337	0.9531	771,224
Portuguese	0.8400	118,311	0.9875	118,311
Arabic	0.7975	706,527	0.8271	831,073

Table 7. Final system accuracy on the PAN AP 2017 test set.

Language	Gender	Variety	Joint
English	0.8133	0.8767	0.7125
Spanish	0.8114	0.9439	0.7704
Portuguese	0.7863	0.9850	0.7750
Arabic	0.7719	0.8169	0.6525

As one can see comparing Tables 4 and 6, the 10-fold cross-validation results of our final system are higher than of the early bird submission for all the languages and subtasks, except for Portuguese gender identification. This decrease in accuracy is caused by mistakenly using not optimal classifier parameters and filtering out the features that occurred in only one document in the corpus. The highest 10-fold cross-validation improvement, more than 5%, was achieved for the English language variety classification. Overall, the results were improved by approximately 1% for gender and 2% for variety identification.

Similarly to the 10-fold cross-validation results, our final system showed higher accuracy than the early bird submission when evaluated on the test set (see Tables 5 and 7) for all the languages, except for Portuguese (a drop of 2.1%). The highest improvements were achieved for the two languages that showed the lowest early bird evaluation results: English and Arabic (improvements of 6.2% and 2.7%, respectively). On average, our final system outperformed the early bird submission by 1.9% (72.76% vs. 70.86%) on the PAN AP 2017 test set.

4 Conclusions

We described our system for gender and language variety identification that took part in the Author Profiling task at PAN 2017. The system configurations are adjusted for each of the languages and subtasks within the competition. The system uses combinations of typed and untyped character n-grams with word n-grams and non-textual features. Feature representations, classifier parameters, and threshold values vary depending on the targeted language and subtask.

One of the directions for future work would be to examine the contribution of other pre-processing steps, such as replacing digits, splitting punctuation marks, and replacing highly frequent words as described in [8], as well as of standardizing non-standard language expressions: slang words, contractions, and abbreviations, as proposed in [3].

Acknowledgments

This work was partially supported by the Mexican Government (CONACYT projects 240844, SNI, COFAA-IPN, SIP-IPN 20162204, 20162064, 20171813, 20171344, and 20172008).

References

1. Buitinck, L., Louppe, G., Blondel, M., Pedregosa, F., Mueller, A., Grisel, O., Niculae, V., Prettenhofer, P., Gramfort, A., Grobler, J., Layton, R., VanderPlas, J., Joly, A., Holt, B., Varoquaux, G.: API design for machine learning software: experiences from the scikit-learn project. In: ECML PKDD Workshop: Languages for Data Mining and Machine Learning. pp. 108–122 (2013)
2. Gómez-Adorno, H., Markov, I., Baptista, J., Sidorov, G., Pinto, D.: Discriminating between similar languages using a combination of typed and untyped character n-grams and words. In: Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects. pp. 137–145. VarDial 2017, ACL (2017)
3. Gómez-Adorno, H., Markov, I., Sidorov, G., Posadas-Durán, J., Sanchez-Perez, M.A., Chanona-Hernandez, L.: Improving feature representation based on a neural network for author profiling in social media texts. Computational Intelligence and Neuroscience 2016 (2016)
4. Malmasi, S., Dras, M.: Native language identification using stacked generalization. CoRR abs/1703.06541 (2017)
5. 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 3rd Workshop on Language Technology for Closely Related Languages, Varieties and Dialects. pp. 1–14. VarDial 2016 (2016)
6. Markov, I., Gómez-Adorno, H., Posadas-Durán, J., Sidorov, G., Gelbukh, A.: Author profiling with doc2vec neural network-based document embeddings. In: Proceedings of the 15th Mexican International Conference on Artificial Intelligence. MICAI 2016, vol. 10062. LNAI, Springer (2017)
7. Markov, I., Gómez-Adorno, H., Sidorov, G., Gelbukh, A.: Adapting cross-genre author profiling to language and corpus. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, vol. 1609, pp. 947–955. CLEF and CEUR-WS.org (2016)

8. Markov, I., Stamatatos, E., Sidorov, G.: Improving cross-topic authorship attribution: The role of pre-processing. In: Proceedings of the 18th International Conference on Computational Linguistics and Intelligent Text Processing. CICLing 2017, Springer (2017)
9. 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. pp. 268–299. CLEF 2014, Springer, Berlin Heidelberg New York (2014)
10. 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 2017, Springer, Berlin Heidelberg New York (2017)
11. Rangel, F., Celli, F., Rosso, P., Pottast, M., Stein, B., Daelemans, W.: Overview of the 3^{rd} author profiling task at PAN 2015. In: Cappelato, L., Ferro, N., Jones, G., Juan, E.S. (eds.) CLEF 2015 Labs and Workshops, Notebook Papers. vol. 1391. CEUR (2015)
12. Rangel, F., Rosso, P., Chugur, I., Pottast, M., Trenkmann, M., Stein, B., Verhoeven, B., Daelemans, W.: Overview of the 2^{nd} author profiling task at PAN 2014. In: Cappellato, L., Ferro, N., Halvey, M., Kraaij, W. (eds.) CLEF 2014 Labs and Workshops, Notebook Papers. vol. 1180, pp. 898–827. CEUR (2014)
13. Rangel, F., Rosso, P., Potthast, M., 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., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2017)
14. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4^{th} 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 (2016)
15. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL: Human Language Technologies. pp. 93–102. NAACL-HLT '15, Association for Computational Linguistics (2015)
16. Zampieri, M., Malmasi, S., Ljubešić, N., Nakov, P., Ali, A., Tiedemann, J., Scherrer, Y., Aepli, N.: Findings of the vardial evaluation campaign 2017. In: Proceedings of the 4th Workshop on NLP for Similar Languages, Varieties and Dialects. VarDial 2017 (2017)