Using character n-grams and style features for gender and language variety classification

Notebook for PAN at CLEF 2017

Rodrigo Ribeiro Oliveira and Rosalvo Ferreira de Oliveira Neto

Universidade Federal do Vale do São Francisco
rodrigo18br@hotmail.com; rosalvo.oliveira@univasf.edu.br

Abstract Author profiling is the problem of determining the characteristics of an author of an anonymous text. In this paper, we detail a method to determine the language variety and the gender of the authors of tweets, as a submission for the Author Profiling Task at PAN 2017. This method seeks to select the most significant character n-grams for each class considered, combining them with style features for gender identification. The experimental evaluation shows that the proposed method gives good performance to determine the language variety and the gender of authors of tweets.

1 Introduction

As computational power grows and artificial intelligence techniques evolve, more problems come to the hands of machine learning researchers. One such problem is author profiling. Different from the traditional authorship identification, in which a closed set of possible authors is known, author profiling aims to determine what are the characteristics of the authors: their age, gender, native language among others. Interest in this field has been growing. One of the reasons for this is how much text is produced in the internet, and a considerable number without a defined author behind. There are multiple practical applications on this field: forensic investigation of criminal messages; linking of certain opinions to profiles of people; better targeting in advertisement.

Since 2013, PAN organizes various tasks in author profiling, with variations on the characteristics of the authors to be determined and the source of the texts used. Each year, the task receives many submissions. In 2017, the texts came from Twitter in four languages: English, Spanish, Arabic and Portuguese. The task was classifying authors in regard to their gender and variety of the language used. This last characteristic was included in PAN for the first time, although it already has been the focus of other tasks, specifically the DSL (Discriminating Similar Languages). However, while this last one presented only two or three variations for each language, the language varieties in this year PAN author profiling task can reach seven (in the specific case of Spanish), offering a more challenging scenario.

In this paper, our approach to the problem will be described. We will introduce the data and how we preprocessed it, our classification method, present the features built from the texts, and the results of our experiments. Based on these, the features we submitted in our solution to the Author Profiling task in PAN will be described.

2 Data and Preprocessing

The data consists of a series of XML files, each one corresponding to an author, containing 100 tweets. These are in raw format, containing links, mentions to other users and hashtags. There are data to four different languages: Portuguese, English, Spanish and Arabic. These languages are divided into varieties. The varieties for each language are:

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

Table 1 contains the number of files present in the training data belonging to each language. The number of authors is divided equally for all varieties in each language. In each variety, the authors are distributed evenly regarding gender, 300 of them being male and 300 female.

Table 1. Number of files for all languages

	English	Spanish	Arabic	Portuguese
Number of files	3600	4200	2400	1200

In each language, the number of files is divided equally by gender and language variety. The split for gender is always of 50%, including for the varieties. After extraction of all tweets from the XML, html tags, links, punctuation and mentions of users are removed. All tweets are lowercased and treated as one single text. This text is tokenized into words through the Natural Language Toolkit (NLTK) [1].

3 Classifier

In most research on author profiling, Support Vector Machines (SVM) has been used with satisfactory results. Comparisons with methods supposed to be more effective, especially deep learning, revealed SVMs to be more discriminatory. Therefore, the implementation of linear SVM in the library scikit-python [7] was chosen as the classification method. In order to prevent overfitting, the value of C was fixed in 1.0, as done in [11].

In order to validate the approach, the data for each language was split in 66% for training and 33% for test. For the experiments on gender, this split is done in order to maintain the number of authors from each gender equal. Although the experiments are made from a subset, the classification in the final task will be made using all the training data.

3.1 Language variety

Most papers use character n-grams as features in the identification of the variety of the language of the text. Approaches diverge in the n considered, and way the grams are used. In [5], the winner submission to DSL 2015, the n-grams are weighted using tf-idf, and the n-grams from the entire corpus considered are used to classification, resulting in more than 13 million features. In [3] and [4], the authors combine character n-grams and bag of words in a method called "Token-Based Backoff", with high dimensionality too. Identification between Brazilian and European Portuguese was done by [13] using n-grams weighted through tf-idf in a Bayesian framework. The authors in [8] discriminates between seven Spanish varieties (Argentina, Chile, Mexico, Peru and Spain) using a method that reduces the dimensionality of the data calculating five metrics from tf-idf 3 and 4-grams.

Table 2. Variation in the amount of unique n-grams for each language as n grows

	English	Spanish	Arabic	Portuguese
2-gram	14676	24050	21929	6653
3-gram	69019	127911	88423	54441
4-gram	756533	576010	780980	220287
5-gram	1175057	747372	1197949	268689

From that, n-grams were decided as the feature to be used in classification of language variety. But the high dimensionality of this sort of feature is a problem, especially due to our limited time to prepare for the task. Table 2 shows the number of distinct n-grams for all languages in the corpus, with n going from 2 to 5. For 4 and 5 grams, most results are around half a million and some more than one million. Therefore, a method was used to select a subset of the n-grams for each language, in such a way that this subset is able to differentiate between their varieties.

A language L contains j language varieties, indicated by L_p, with 0<p<j-1. Existing k n-grams in L, each one of them expressed by g_i, in which 0<i<k-1, $L_j(g_i)$ is the number of times g_i appears in a particular language variety. Each gram receives, for a particular variety L_p, the following ranking:

$$r_p = \sum_{m=0}^{j-1} \frac{L_p(g_i)}{L_m(g_i) + 1}, m \neq p$$

So, the more gi appears in other language varieties, lower will be its ranking relating to L_p. If is more frequent in L_p than other varieties, however, its ranking in this variety will be higher. Therefore, selecting the first N grams with higher values of r_p will give features characteristic of L_p.

3.2 Gender

There are basically two approaches found in former works to finding the gender of an author:

– Using content-based features, that reflect the particular topic the text is about, albeit powerful, they are susceptible to overfitting, if the classification is too dependent of the subject treated in the training texts;
– Style-based features, reflecting stylistic structures in the writing of individual authors. They have the advantage of being less dependent of the particular conditions of a given text (topic or size, e.g.).

Usual content features used are individual words, word n-grams and character n-grams. A great number of style features appears in literature: function words, use of punctuation signs, orthographic errors, vocabulary richness, parts of speech, sentiment analysis. Some of these features, however, require specific tools to be used in a language.

For example, sentiment analysis require a Dictionary of Emotions, to find parts of speech in a text a tagger is needed. The lack of familiarity with one of the languages treated here, Arabic, and the low amount of work on another, Portuguese, makes hard to apply many style-based features present in past research, specially one of the most promising, function words.

Given that, the features used to classify texts by gender will be the same selection of character n-grams used in language variety, but considering only the amount of grams relative to if the author is male or female, as content features. As style-based features, were chosen:

– number of repeated vowels;
– number of hashtags used;
– number of mentions to other users;
– vocabulary variety, i.e., the ratio between the number of unique words and the number of total words in the text;
– cleanliness of text, i.e., the ratio between number of characters it had after and before the preprocessing step.

4 Experiments

This section will describe the experiments done, their methodology, the results achieved and the decisions taken for the evaluation of the Author Profiling task.

4.1 Optimal number of characters in n-grams

There is support in research that if the number of characters of the n-grams grows too much, performance will fall, because the terms computed will start to coincide too much with individual words. Some works suggest the average length of words in a corpus to be the maximum value of n considered to build n-grams. Therefore, the average length of words for all languages was calculated, the results presented in Table 3. The average length oscillates between 4 and 5 for all languages. So, 3 to 6-grams were built for all English and Spanish, and 3 to 5-grams for Arabic and Portuguese, in order to check their individual accuracy in classification.

Table 3. Average word length for each language

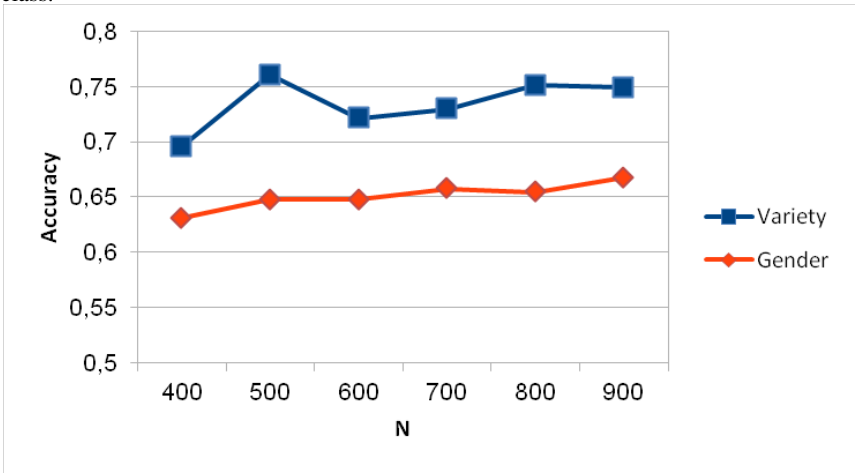
	Spanish	Portuguese	Arabic	English
Average word length	5.17	4.58	4.59	5.15

4.2 Choosing the number of features

An optimal number N of n-grams must be chosen for be used as features to classification, after their ranking has been calculated. Choosing the first N with higher ranking for each class is the obvious approach, in which N must not be too low, so the features will not be powerful, nor too high, making the selection of a subset of n-grams useless. Figure 1 show the results of the gradual raise of N for classification on English using 3-grams. As can be seen from them, the accuracy increases along with N and determining a optimal point would take doing experiments on much higher values of N, which was not possible due to the deadline.

The number of first N n-grams to be used was fixed by class used, resulting in a total of features N*number of classes for language. Each language has a different number of classes, therefore, different values of N were chosen for each one, seeking both good accuracy and a system not too heavy.

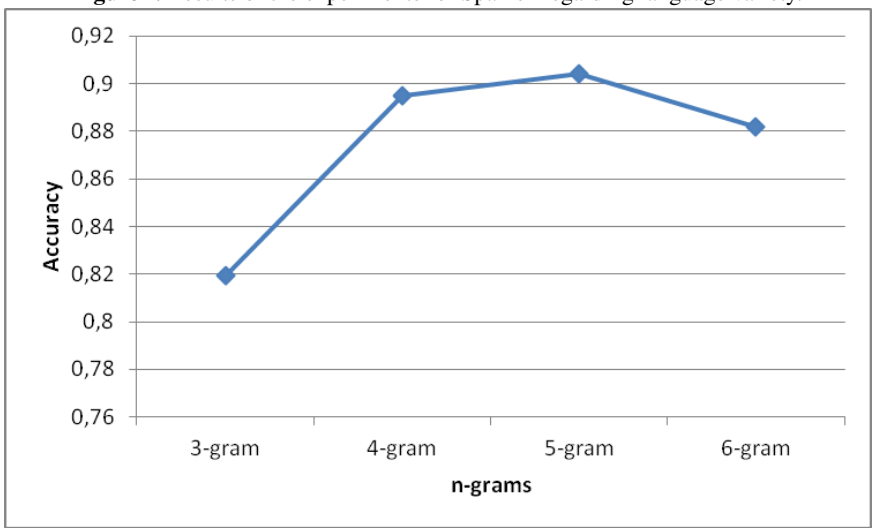
Figure 1. Accuracy for 3-grams for language variety in English varying the number of features for class.



Considering the corpus size, experiments were realized using the following values to select the best ranked character n-grams. For English and Spanish, languages with larger corpus, we used 2000 features per class to model variety, and 6000 in gender. For Arabic, were used 4000 features per class in variety, and 8000 in gender. And in Portuguese, 10000 features per class to classify both variety and gender. Moreover, the gender classification used the style features described in section 3.

Defined the number of features for each language, the experiments were ran varying the size of the n-grams. Figures 2 to 5 show the results regarding language variety for each language. There is a peak in accuracy in 4 or 5 grams, with values stabilizing or falling in the next value of n. This is consonant to what was predicted in the preceding section. Given that 6-grams tend to present a fall in accuracy, 3, 4 and 5 were chosen as the size of the n-grams used in the submission.

Figure 2. Results of the experiments for Spanish regarding language variety.



Due to lack of time, the full tests regarding gender were made only for Portuguese and Arabic. The results are in the Figures 6 and 7. The accuracy tends to grow with the raise of n, but the increase between 4 and 5-grams is smaller. This suggests that in 5-grams accuracy for gender is stabilizing.

From the experiments, the value of N for variety and gender chose for each language is in the Table 4. Due to good results using the 2000 best ranked grams in English and Spanish, we raised the number of features for the 3000 best ranked n-grams. In Portuguese, due to the poor performance of 5-grams, we reduced the number of these features to 2500 per class.

Each text in the corpus was represented as a binary vector marking the presence or absence of a specific n-gram previously selected as a feature.

4.3 Using tf-idf weighted character n-grams as supplementary features

An issue that was noted when obtaining the vectors for each text: some texts would have too few of the features selected previously (some less than 100). In order to supplement such cases, 3 and 4-grams weighted through tf-idf were chosen as additional features. For Portuguese, English and Arabic, the 15000 most frequent 3-grams and 4-grams

Figure 3. Results of the experiments for English regarding language variety.

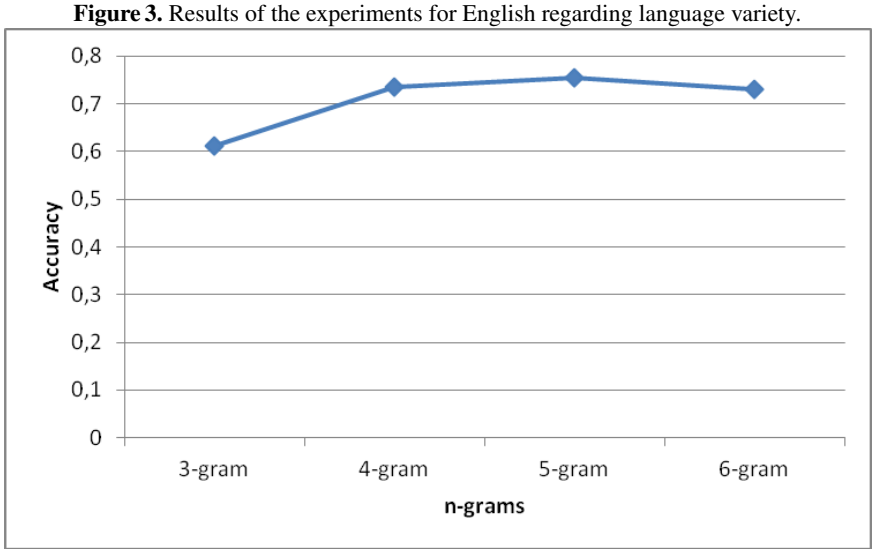


Figure 4. Results of the experiments for Portuguese regarding language variety.

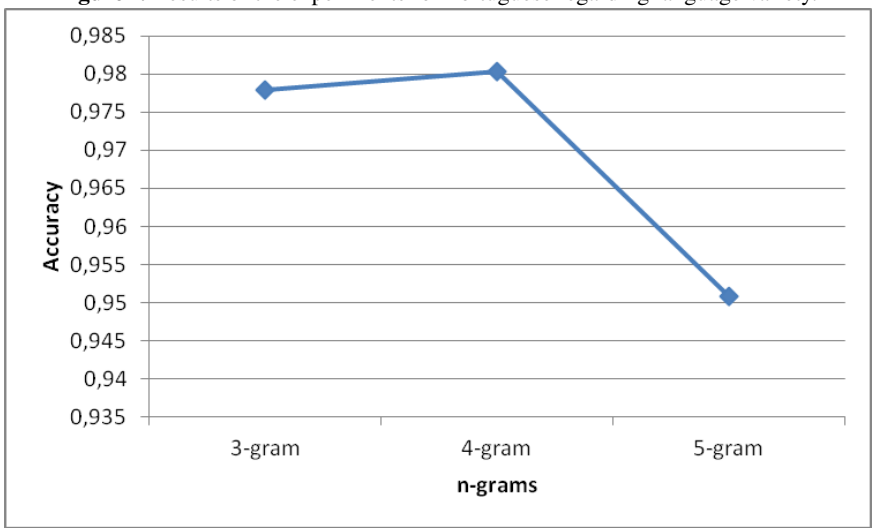


Figure 5. Results of the experiments for Arabic regarding language variety.

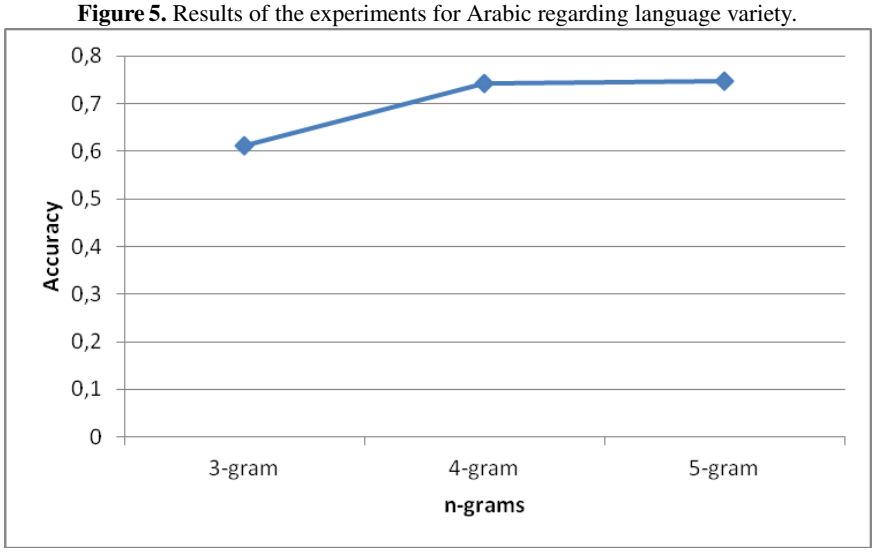


Figure 6. Results in gender for Portuguese.

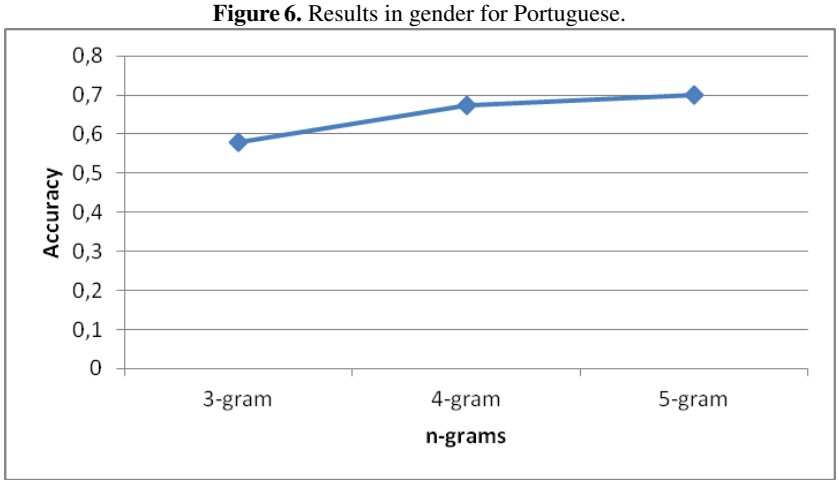
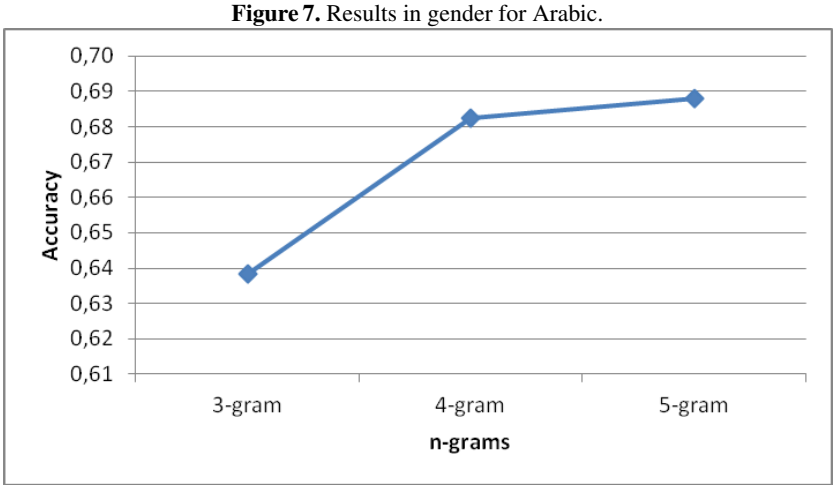


Table 4. Value of N for variety and gender for each language, in which N is the number of features more relevant per class, and the respective totals.

		Spanish	Portuguese	Arabic	English
N Variety(per class)	3-gram	3000	10000	4000	3000
	4-gram	3000	10000	4000	3000
	5-gram	2000	2500	4000	3000
N Gender (per class)	3-gram	6000	10000	8000	6000
	4-gram	6000	10000	8000	6000
	5-gram	6000	10000	8000	6000

Figure 7. Results in gender for Arabic.



were selected, for Spanish, given the amount of texts, the 7500 first. No experiments could be conducted on these features due to time constraints.

5 Conclusion

The approach here defined performs well when compared with the ones in the state of the art. In [12] and [6], accuracy of over 90% was achieved in the DSL 2015 and 2016 Shared Task, respectively. But these tasks deal with two or three classes at most, and only would compare to our work in Portuguese, that does reach that level of accuracy. In [8], the authors achieved accuracy of over 71% considering five varieties of Spanish, which some of our results surpass, achieving more than 90% of accuracy. In DSL 2016 [2], the winner of spoken Arabic dialect task achieved accuracy of 51.36%. Although the solution presented here performs better, the comparison is somewhat inadequate because that work deals with spoken language and ours with written.

Regarding gender, the experiment that we managed deliver a performance equivalent or superior to the two most recent PAN tasks in Author Profiling. In [10], the authors report the best results ranging from 70% to 90% in the 2015 task. In [9], the authors reported a fall in performance in the 2016 task, with results in gender varying between 50% and 80%.

On our submission to the 2017 Task, although we applied for all languages, a execution issue in our software involving memory resulted in only the output for Portuguese and Arabic being generated. The results are in the Table 5. The accuracy for variety in Portuguese is the best result achieved by us, and the best in all results along with three other teams. The detection of variety in Arabic was rather poor, being the third worst result. The results for gender were in the average range for both languages. These results follow closely the ones obtained in the experiments, with exception of gender for Portuguese, that exceeded in more than 6% the experimental results.

Table 5. Results for Portuguese and Arabic

	Variety	Gender
Portuguese	0.9850	0.7650
Arabic	0.6713	0.7013

Portuguese was the language with the least number of varieties, so would naturally be the one with best results regarding this task. It was the one with more number of character n-grams by variety selected, so the model is a better classifier. Although the total number of n-grams used for training the classifiers for Portuguese and Arabic was similar, the amount by specific variety in this latter language is much lower, presumably reducing the discriminsating power of the model trained.

Moreover, the rankings for the n-grams for each class are obtained comparing their frequencies in that class with the ones from all the other classes. In Portuguese, with only two varieties, this is not a problem, but in Arabic, there is the possibility that information from some of the other classes was ignored or neglected, and only some relations between classes considered. Therefore, for future works, we propose investigation of selection of the character n-grams combining varieties in pairs in languages with more than two varieties.

References

1. Bird, S., E.L., Klein, E.: Natural Language Processing with Python. O'Reilly Media Inc. (2009)
2. Eldesouki, M., Dalvi, F., Sajjad, H., Darwish, K.: Qcri@ dsl 2016: Spoken arabic dialect identification using textual features. In: VarDial 3. pp. 221–226 (December 2016)
3. Jauhiainen, T., Jauhiainen, H., Lindén, K.: Discriminating similar languages with token-based backoff. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (2015)
4. Jauhiainen, T., Lindén, K., Jauhiainen, H.: Heli, a word-based backoff method for language identification. In: VarDial 3. p. 153 (2016)
5. Malmasi, S., Dras, M.: Language identification using classifier ensembles. In: Proceedings of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial). pp. 35–43 (September 2015)
6. Malmasi, S., Zampieri, M., Ljubešic, N., Nakov, P., Ali, A., Tiedemann, J.: Discriminating between similar languages and arabic dialect identification: A report on the third dsl shared task. In: VarDial 3. pp. 1–14 (December 2016)
7. 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)
8. Rangel, F., Franco-Salvador, M., Rosso, P.: A low dimensionality representation for language variety identification. In: Postproc. 17th Int. Conf. on Comput. Linguistics and Intelligent Text Processing, CICLing-2016, Springer-Verlag, LNCS: arXiv:1705.10754 (2017)
9. Rangel, F., Rosso, P., Verhoeven, B., Daelemans, W., Potthast, M., Stein, B.: Overview of the 4th author profiling task at pan 2016: cross-genre evaluations. Working Notes Papers of the CLEF. (2016)

10. Rangel, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF (2015)
11. op Vollenbroek, M.B., Carlotto, T., Kreutz, T., Medvedeva, M., Pool, C., Bjerva, J., Haagsma, H., Nissim, M.: Gronup: Groningen user profiling (2016)
12. Zampieri, M., Tan, L., Ljubešic, N., Tiedemann, J., Nakov, P.: Overview of the dsl shared task 2015. In: Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects (LT4VarDial). pp. 1–9 (September 2015)
13. Zampieri, M., Gebre, B.G.: Automatic identification of language varieties: The case of portuguese. In: KONVENS2012-The 11th Conference on Natural Language Processing. pp. 233–237. Österreichischen Gesellschaft für Artificial Intelligende (ÖGAI) (2012)