UniNE at CLEF 2017: TF-IDF and Deep-Learning for Author Profiling

Notebook for PAN at CLEF 2017

Nils Schaetti

University of Neuchâtel
rue Emile Argand 11
2000 Neuchâtel, Switzerland
nils.schaetti@unine.ch

Abstract. This paper describes and evaluates a strategy for author profiling using TF-IDF and a Deep-Learning model based on Convolutional Neural Networks. We applied this strategy to the author profiling task of the PAN17 challenge and show that it can be applied to different languages (English, Spanish, Portuguese and Arabic). As features, we suggest using a simple cleaning method for both models, and for the Deep-Learning model, a matrix of 2-grams of letters with punctuation marks, beginning and ending 2-grams, as features. Applying this strategy, we determine that the TFIDF-based model is the best one for language variety classification and that the Deep-Learning model achieve the highest accuracy on gender classification. The evaluations are based on four tweet collections (PAN AUTHOR PROFILING task at CLEF 2017).

1 Introduction

Today, a large amount of data is produced by web applications based on social contexts, like social networks and blogs, where a variety of contents (e.g. pictures, videos, articles, links, texts) are shared directly from web sites and smartphones. Social networks like Facebook and Twitter allow a new kind of communication based on fast interactions which generate multimedia contents with their own characteristics, that are difficult to compare with traditional texts like essays and articles.

This rises new questions : can we detect differences in writing styles between men and women, language varieties, age groups or psychological profiles? Theses questions are appealing as answers to new problems created by the age of social networks and blogs, such as fake news, plagiarism and identity theft. The problem of author profiling is therefore of particular interest.

Moreover, author profiling is becoming more and more important for applications in marketing, security and forensics. For example, in forensic linguistics, one would like to know certain characteristics (gender, age group, socio-cultural background) of an author of harassing messages from its linguistic profile. In marketing, companies and resellers would like to know the characteristics of people liking or disliking their products based on the analysis of blogs and product reviews.

This paper is organised as follow. Section 2 introduces the dataset used for training

and testing, as well as the methodology used to evaluate our approach. Section 3 describes the cleaning and tokenization process. Section 4 explains the proposed TFIDF-based model. Section 5 describes the Deep-Learning based classifier. In section 6, we evaluate the strategy we created and compare results on the four different test collections. In the last section, we draw conclusions on the main findings and possible future improvements.

2 Tweet collections and methodology

To carry out experiments on the author profiling task with different algorithms, we need a common ground composed of the same datasets and evaluation measures. In order to create this common ground, and to allow the large study in the domain of author profiling, the PAN CLEF evaluation campaign was launched ([7]). Multiple research groups with different backgrounds from around the world have proposed a profiling algorithm to be evaluated in the PAN CLEF 2017 campaign with the same methodology [5].

All teams have used the *TIRA* platform to evaluate their strategy. This platform can be used to automatically deploy and evaluate a software [4]. The algorithms are evaluated on a common test dataset and with the same measures, but also on the base of the time need to produce the response. The access to this test dataset is restricted so that there is no data leakage to the participants during a software run.

For the PAN CLEF 2017 evaluation campaign, four test collections of tweets were created, one for each of the following languages : English, Spanish, Portuguese and Arabic. Based on these collections, the problem to address was to predict the author's *language variety* (varieties of the main language) and its *gender* [6].

The training data was collected from Twitter. For each tweet collections, the texts come from the same language and are composed of tweets from authors, 100 tweets per authors. For each author, there is two labels we can predict :

1. The author's gender (*male*, *female*);
2. The author's language variety, specific to the language;

The test sets are also texts collected from Twitter and the task is therefore to predict the *gender* and *language variety* for each Twitter author in the test data. There is one collection per language (English, Spanish, Portuguese and Arabic). The English collection is composed of 3600 authors, coming from six different countries, United-States, Great Britain, Ireland, New Zealand, Australia and Canada, 600 for each variety, and 1800 for each gender, for a total of 360'000 tweets.

The Spanish collection is composed of 4200 authors, coming from seven different countries, Colombia, Argentina, Spain, Venezuela, Peru, Child and Mexico, 2100 for each gender, for a total of 420'000 tweets.

The Portuguese collection is composed of 1200 authors coming from Brazil and Portugal, 600 for each gender, for a total of 120'000 tweets.

Finally, the Arabic collection is composed of 2400 authors from Gulf, Levantine, Maghrebi and Egypt, 1200 for each gender, for a total of 240'000 tweets.

An overview of these collections is depicted in table 1. The number of authors from

	Training sets			
Corpus	Authors	Tweets	Language varieties	Genders
English	3600	360k	US, GB, Ireland, New Zealand, Australia, Canada	1800; 1800
Spanish	4200	420k	Columba, Argentina, Spain, Venezuela, Peru, Chile, Mexico	2100; 2100
Portuguese	1200	120k	Portugal, Brazil	600; 600
Arabic	2400	240k	Gulf, Levantine, Maghrebi, Egypt	1200; 1200

Table 1: PAN CLEF 2017 *training* corpora statistics

the training set is given under the label ”Authors” and the total number of tweets in the collection is indicated under the label ”Tweets”. The label ”Languages varieties” shows the varieties for each collections, and the label ”Genders” indicates the number of authors for each gender.

The training data set is well balanced as for each collection, there is the same number of authors for each language variety and gender. The Spanish collection is the biggest with 4’200 authors, and the smallest is the Portuguese collection with 1’200 authors. All the collections have the same number of authors for each language varieties (600), but varies for the genders.

A similar test set will be used to compare the participants’ strategies of the PAN CLEF 2017 campaign, and we don’t have information about its size due to the *TIRA* system.

For the PAN CLEF 2017 campaign, the software must provide its answer to each problem as an XML data. The response for the gender is a binary choice (*male* / *female*), and the language variety is one of the possible outputs for the language of the collection.

The overall performance of the system is the joint accuracy of the gender and language variety. The accuracy is the number of authors where both the gender and language variety is correctly predicted for the same author divided by the number of authors in the collection.

The accuracy for language varieties and genders are also computed as the number of correct answers divided by the total number of authors.

3 Cleaning and Tokenization

Before selecting the features from the texts, we need to clean the text and extract tokens. This section aims to explain these two steps.

To carry out these two steps, we apply a serie of rules to the tweet’s text in the following order :

1. Remove URL (*http://.../*);
2. Remove Twitter quotes (*@username*);
3. Remove special characters #, $ and §;
4. Clean numbers (100.00 → 100 § 00, 100,00 → 100 $ 00, 100'000 → 100 # 000);
5. Tokenize punctuation (???? → " ? ? ? ? ");
6. Remove new line character;
7. Replace useless characters by space (-, ..., *, /, +, \);
8. Remove multiple spaces;

The step 4 aims to introduce three tokens (§, $ and #) that indicate the way the user represents numbers (points or comma for float, comma or apostrophe for thousands).

Each collection are from a different language and, therefore, may use a different alphabet. For each collection, we keep in the text only the letters and punctuation corresponding to the language's alphabet. The alphabets for each language are depicted in the table 2.

We keep letters with accents that are not often used in the language to represent their usage by the authors. For example, we keep accents in the English texts to represent the usage of French words by English authors that could help as information for profiling the author. We make the hypothesis that authors use more or less French or Spanish words depending on the country from which they're coming.

Concerning the hashtags, we keep them and compute them as normal words. At the end of the cleaning process, the words are splitted by space and used as tokens for the feature selection.

4 TFIDF-based model

The TF-IDF (*Term-Frequency-Inverse Document Frequency*) is a weighting method often used in *information retrieval* and mostly in *text mining*. This statistical measure

Language	**Alphabet**	**Punctuation**
English	AaÀàÁáÂâÃãBbCcÇçDdEeÉéÈèÊêFfGgHhIiÍíÏïÎîJjKkLlMmNn OoÓóÔôÕõPpQqRrSsTtUuÚúÜüÛûVvWwXxYyZz	?.!,:;
Spanish	AaÀàÁáÂâÃãBbCcÇçDdEeÉéÈèÊêFfGgHhIiÍíÏïÎîJjKkLlMmNn ÑñOoÓóÔôÕõPpQqRrSsTtUuÚúÜüÛûVvWwXxYyZz	?¿.!¡,:;
Portuguese	AaÀàÁáÂâÃãBbCcÇçDdEeÉéÈèÊêFfGgHhIiÍíÏïÎîJjKkLlMmNn OoÓóÔôÕõPpQqRrSsTtUuÚúÜüÛûVvWwXxYyZz	?.!,:;
Arabic	ثشنتجججححخدذرزسشصضطظعغفقكلمنهوىيءئإؤأـنيتث. عَحَشصفعسمكحوهدجبا؛زطبلنقرتثذضظب	؟،؛,!.؟،

Table 2: Language alphabet for each language collection

makes it possible to assess the importance of a term in a document, in relation to a collection or corpus. The raw frequency of a term in simply the number of occurrences of this term in a specific document, this frequency is also called *term frequency*.

The *inverse document frequency* is a measure of the importance of a term in the whole collection. In the TF-IDF model, it gives more weight to less frequent terms, considered to be more discriminatory. It consists in calculating the base-10 logarithm of the inverse of the proportion of documents in the corpus that contain the term.

To describe our problem more formally, a document d in our collection is the set of all tweets belonging to a class to predict (*gender* or *language variety*). D is the set of all documents in the collection and $|D|$ is the number of documents in the collection ($|D| = 2$ if the problem is to predict gender).

The *term frequency* for a term t and a document d is therefore defined by

$$tf_{d,t} = \frac{n_{d,t}}{|d|}$$

where $n_{d,t}$ is the number of occurrences of the term t in the document d. The *term frequency* $tf_{d,t}$ is then the number of occurrence of the term t in document d divided by the total number of tokens in the document.

The *inverse document frequency* of a term t in the whole collection is,

$$idf_t = log\frac{|D|}{|\{d : t \in d\}|}$$

where $|D|$ is the number of classes in the classification problem and $|\{d : t \in d\}|$ is the number of document(s) where the term t appears. The final *tfidf* value for a document d and term t is defined by

$$tfidf_{d,t} = tf_{d,t} * idf_t = \frac{n_{d,t}}{|d|} * log\frac{|D|}{|\{d_j : t \in d_j\}|}$$

For each document d, we compute a vector $tfidf_d$ with the $tdidf$ values for each

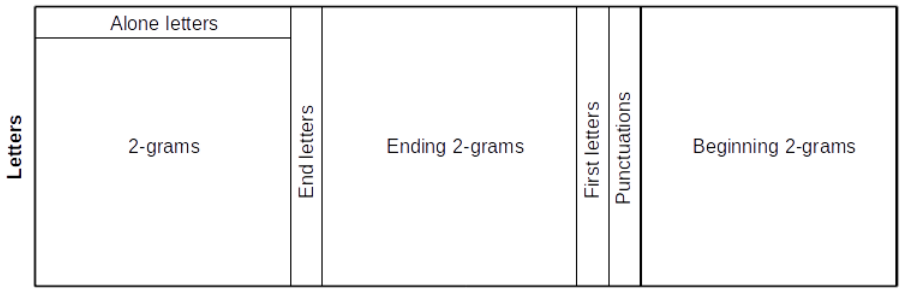

Fig. 1: Structure of the input features matrix. From left to right, the matrix represents the ratios of 2-grams of letters and of single-letter tokens, the ratio of word ending 2-grams, the ratio of word beginning letters, the ratio of punctuation marks, and the ratio of word beginning 2-grams.

terms in the collection. If a term does not appear in the document, its value is set to zero. When we want to predict the class (*gender* or *language variety*) of a previously unseen author from the collection, we consider it as query q and compute the cosinus similarity between the $tfidf_d$ vector and the vector tf_q of term frequencies in q :

$$sim(d,q) = \frac{tfidf_d \cdot tf_q}{\|tfidf_d\| * \|tf_q\|}$$

where

$$tf_{q,t} = \frac{n_{q,t}}{|d|}$$

and finally, we choose the predicted class $\hat{c}_q$ for the query q with the biggest similarity. For example, in the case of gender, we choose $\hat{c}_q$ as

$$\hat{c}_q = \max_{d \in \{male, female\}} sim(d,q)$$

5 Two-grams of letters-based Convolutional Neural Networks

In machine learning, a *Convolutional Neural Network* (or CNN) is a kind of feed-forward artificial neural network [1] [2] [3], in which the patterns of connection between the neurons are inspired from the visual cortex.

In our system, we applied a CNN to a matrix representing the 2-grams of letters for an author in a collection. The figure 1 shows the structure of the 2-gram matrix.

There is a row for each letter in the alphabet. From left to right, the matrix is composed of a first part where ratio of each 2-gram of letters in the author's tweets, the upper line representing the ratio of alone letters. The second part is composed of ratio

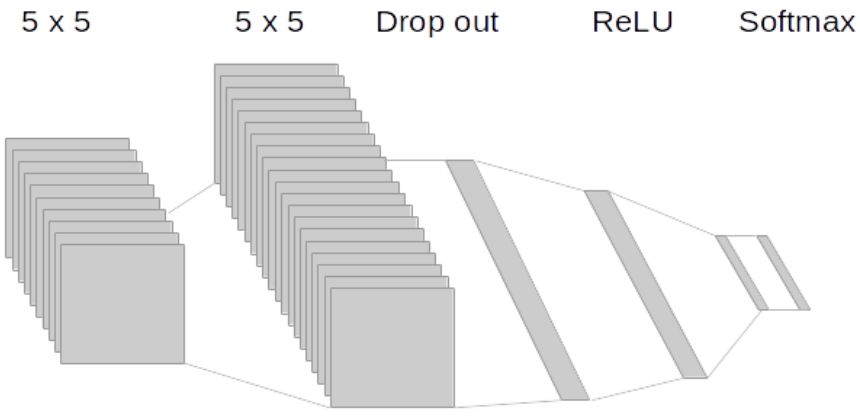

Fig. 2: Structure of the Convolutional Neural Network on 2-grams of letters with the following layers : 10x5x5 kernel, 20x5x5 kernel, drop-out, two linear layer (ReLU), softmax.

of ending letters. The next part is the ratio of ending 2-gram of letters computed from the tokens obtained after the cleaning process of section 3.

The fourth part is the ratio of first letters in tokens, and the fifth the ratio of punctuation marks. Finally, the last part is the ratio of 2-grams found at the beginning of each tokens. This matrix representing an author is the input for the *Convolutional Neural Network* shown in figure 2.

The first layer is a *convolution layer* of 10 kernels of size 5×5. This layer is the input for a second *convolution layer* of 20 kernels of size 5×5 followed by a *drop-out layer*. This serves as an input for two *linear layers* based on *ReLU*. Finally, the outputs are obtained from a *softmax* function and give the predicted class of the author. The predicted class is then the class with the highest corresponding output of the *softmax* function.

The training phase consists of using 90% of the training set for training and 10% to evaluate the performances at each iteration. The figure 3 shows the evolution of training and test losses after each iteration of the training phase for each language collection. Vertical lines show the lowest test loss for each collection.

For English, the lower loss is attained after 64 iterations and 66 for Spanish. For Portuguese and Arabic, the lower losses are respectively attained after 87 and 38 iterations. We can see that our CNN model quickly overfit, especially for the Arabic language collection. The main challenge with this model is then to fight effectively overfitting.

At the end of the training phase, we choose the CNN obtained at the iteration with the lowest loss.

6 Evaluation

To evaluate our two models we tested their accuracy on each language training collections. The table 3 shows the results of 10-fold cross validation for each combination

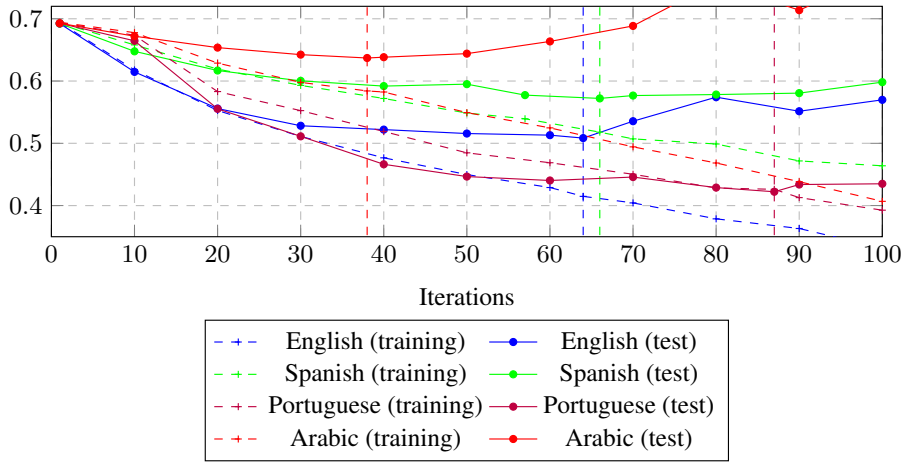

Fig. 3: Training and test loss after n iterations for the gender task on each collection

of model and collection with the random baseline as comparison.

For English, the *TF-IDF* model attains 83% and 68% accuracy on language variety and gender respectively against 16% and 50% for the random classifier. The CNN model got respectively 65% and 78% accuracy for language variety and gender. The combination of the two models achieve a final accuracy of 65%.

For Spanish, the *TF-IDF* model got 93% and 64% accuracy on language variety and gender respectively against 14% and 50% for the random classifier. The CNN model achieves respectively 78% and 72% accuracy for language variety and gender. The combination of the two models got a final accuracy of 67% against 7% for the random classifier.

For Portuguese, the *TF-IDF* model achieved an accuracy of 99% and 73% on the language variety and gender classification task respectively. The CNN model got an accuracy of respectively 98% and 85% for the language variety and gender tasks against 50% for a random classifier.

For Arabic, the *TF-IDF* model achieved an accuracy of 86% and 68% for the language variety and gender classification problem compared to 25% and 50% for a random classifier. The *CNN* model got an accuracy of respectively 67.5% and 75% for language variety and gender. The combination of both models gives a final accuracy of 64%.

We can see that the *no free-lunch principle* applies here as the *TF-IDF* achieved better results on the language variety profiling while the *CNN* model achieves better results when it comes to profiling gender.

The *TF-IDF* model achieves an impressive results for the language variety task on the Spanish collection with 93% accuracy while there is 7 different possible varieties (Argentina, Spain, Chile, Mexico, Colombia, Venezuela, Peru). Its best accuracy is at-

Corpus	TF-IDF	CNN	**Final**	*Random*
English varieties	**0.8333**	0.6563		0.1666
English genders	0.6805	**0.7803**		0.5000
	0.4724	0.5228	**0.6502**	0.0833
Spanish varieties	**0.9323**	0.7804		0.1428
Spanish genders	0.6491	**0.7238**		0.5000
	0.6051	0.5648	**0.6747**	0.0714
Portuguese varieties	**0.9925**	0.9833		0.5000
Portuguese genders	0.7317	**0.8500**		0.5000
	0.5313	0.8358	**0.8436**	0.2500
Arabic varieties	**0.8609**	0.6750		0.2500
Arabic genders	0.6888	**0.7500**		0.5000
	0.5929	0.5028	**0.6456**	0.1250
Overall	0.6228		0.7035	0.3824

Table 3: 10-fold cross validation on the four *training* collections

tained on the Portuguese collection with 99% accuracy and its lowest on the gender classification on the Spanish collection. The *CNN* model achieves its best performance on the Portuguese collection for language variety classification with 98.3% accuracy and its lowest on the English collection for language variety classification with 65.6%.

The table 4 shows the results on the six training collections obtained on the *TIRA* platform.

The results shows the same pattern as the previous 10-fold cross validation. Portuguese-

Corpus	Variety	Gender	**Both**	*Random*
English	0.9717	0.7903	**0.7681**	0.0833
Spanish	0.9895	0.7674	**0.7595**	0.0714
Portuguese	0.9992	0.8367	**0.8358**	0.2500
Arabic	0.8967	0.7521	**0.6904**	0.1250
Overall	0.9642	0.7866	**0.7634**	0.3824

Table 4: Evaluation for the four *training* collections

speaking authors are the easier to profile with an overall accuracy of 83%, a little bit lower than with the 10-fold CV (84%). The Arabic are the hardest to profile with an overall accuracy of 69% against 64% on the 10-fold CV. The language variety accuracy is much higher than the gender accuracy, as for the 10-fold CV.

The difference between the training results and the previous 10-fold cross validation on the gender problem is 11.8 for English, 8.4 for Spanish, -0.78 for Portuguese and 4.5 for Arabic. We have clear differences between the test and training accuracy except for Portuguese.

The table 5 shows the results obtained on the four test collections, thanks to the *TIRA* platform. For the English language collection, the accuracy goes from 83% with the 10-Fold CV to 81.5% (-1.83) for language variety, and from 78% to 75% (-2.86) for gender classification.

The accuracy on the Spanish language collection goes from 93.23% to 93.36% (-0.13) and from 72.3% to 71.07% (-1.31) for language variety and gender respectively. For the Portuguese collection, the accuracy goes from 99.25% to 98.38% (-0.87) and from 85% to 72% (-12.5) for language variety and gender respectively. Finally, for the Arabic collection, the accuracy goes respectively from 86% to 81% (-4.78) and from 75% to 71% (-3.56) for language variety and gender.

7 Conclusion

This paper proposes a combination of TFIDF-based model and Deep-Learning *Convolutional Neural Network* to predict the language variety and gender of Twitter authors. Based on the hypothesis that an author's writing style can be used to extract its country of origin and its gender, we introduced classifiers that can effectively predict these two characteristics. The TFIDF-based model shows a good performance on language variety classification, on the other hand, the CNN model is effective to classify authors

Corpus	Variety	Gender	**Both**	*Random*
English	0.8150	0.7517	**0.6133**	0.0833
Spanish	0.9336	0.7107	**0.6657**	0.0714
Portuguese	0.9838	0.7250	**0.7138**	0.2500
Arabic	0.8131	0.7144	**0.5863**	0.1250
Overall	0.8863	0.7254	**0.6447**	0.3824

Table 5: Evaluation for the four *test* collections

in gender classes. For both model and for all language, we proposed a simple cleaning process used by both classifiers, and we selected features as a matrix of ratio of various 2-grams for the CNN classifier.

The TFIDF-based model performs well on language variety classification and achieves its best performance, on the test dataset, on the Portuguese collection with 98% accuracy, and an interesting accuracy of 93% on the Spanish collection. The performances on the English and Arabic collections stay behind with respectively 81.5% and 81.3%. The CNN model achieves its best performance on the English collection with 75.1% accuracy. Furthermore, we see two ways to improve this strategy in the future. First, the CNN classifier shows signs of overfitting and a great difference appears between the 10-fold CV and the final test results. Some improvements could probably be done on the training phase. Secondly, the matrix of 2-grams could be improved by determining which features are useful or not, this could significantly lower the computational complexity of the model. The biggest challenge for the CNN model is the small size of the training collections and more work could be done on this point to improve the overall performances.

The biggest challenge of this year's PAN author profiling task were the gender classification problem were our model achieve an average of 72.5% accuracy compare with 88.6% for language variety classification.

References

1. Ciresan, D.C., Meier, U., Schmidhuber, J.: Multi-column deep neural networks for image classification. CoRR abs/1202.2745 (2012), http://arxiv.org/abs/1202.2745
2. Fukushima, K.: Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. Biological cybernetics 36(4), 193–202 (1980)
3. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11), 2278–2324 (1998)
4. 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)

5. 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 Association, CLEF 2017, Dublin, Ireland, Septembre 11-14, 2017, Proceedings. Springer, Berlin Heidelberg New York (Sep 2017)
6. Rangel, F., Rosso, P., Potthast, M., Stein, B.: In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) CLEF 2017 Labs Working Notes
7. 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)