Multi-channel Open-set Cross-domain Authorship Attribution
Notebook for PAN at CLEF 2019

José Eleandro Custódio and Ivandré Paraboni

School of Arts, Sciences and Humanities (EACH)
University of São Paulo (USP)
São Paulo, Brazil
{eleandro,ivandre}@usp.br

Abstract This paper describes a multi-channel approach to open-set cross-domain authorship attribution (AA) for the PAN-CLEF 2019 AA shared task. The present work adapts the EACH-USP ensemble method presented at PAN-CLEF 2018 to an open-set scenario by defining a threshold value for unknown authors, and extends the previous architecture with an additional character ranking model built with the aid of the PageRank algorithm. Results are superior to a number of baseline systems, and remain generally comparable to those in the original closed-set ensemble approach.

1 Introduction

Authorship attribution (AA) is the computational task of identifying the author of a given text by examining samples of texts written by a number of candidate authors [8]. Practical applications include, for instance, the detection of internet misuse, text forensics for copyright protection, and many others [5].

AA may be based on single- or cross-domain settings. In this paper we discuss the latter, that is, situations in which we would like to identify the author of a text in a certain genre based on samples of text written in another genre.

From a computational perspective, we may distinguish two AA problem definitions: closed- and open-set AA. Closed-set AA assumes that the author of a disputed text necessarily belongs to a pre-defined set of possible candidates. This subtask was the theme of the PAN-CLEF 2018 shared task in [7]. Open-set AA, by contrast, assumes that the disputed text may not necessarily belong to any known candidate [18]. This subtask was the theme of the PAN-CLEF 2019 shared task, and it is also the focus of the present work.

In the context of closed-set AA, the work in [2,3] presented an ensemble approach that combines predictions made by three knowledge channels, namely, standard character n-grams, character n-grams with non-diacritic distortion and word n-grams. In the present work, this method is adapted to an open-set scenario by defining a threshold value for unknown authors, and further extended with the inclusion of a fourth channel based on a character ranking model built with the aid of the PageRank algorithm [10,15].

2 Related Work

The present work consists of an extension of the ensemble AA approach in [2]. This, and a number of related studies, are briefly discussed below.

The work in [2] presented an ensemble approach to cross-domain AA called EACH-USP, which combines predictions made by three independent classifiers based on word n-grams (*Std.wordN*), standard character n-grams (*Std.charN*), and character n-grams with non-diacritic distortion (*Dist.charN*). The method relies on variable-length n-gram models and multinomial logistic regression, and selects the prediction of highest probability among the three models as the output for the task by soft voting.

The word-based *Std.wordN* model in [2] is intended to help distinguish an author from another based on word usage. However, given that a single author may favour different words across domains (e.g., fictional versus dialogue text), and that word-based models will usually discard punctuation and blank spaces thyat may represent a valuable knowledge source for AA [13], the character-based models *Std.charN* and *Dist.charN* were added as a means to capture time and gender inflection, punctuation and spacing.

Both *Std.charN* and *Dist.charN* models in [2] are intended to capture language-independent syntactic and morphological clues for AA. In the latter, all characters that do not represent diacritics are removed from the text beforehand, therefore focusing on the effects of punctuation, spacing and the use of diacritics, numbers and other non-alphabetical symbols.

For further details regarding the ensemble method, we report to [2]. Character models are extensively discussed in [14], with details regarding the role of affixes and prefixes in the task. Function words and word n-gram models are discussed in [6]. Text distortion models for removing noise information from text are discussed in [17].

Finally, the work in [19] creates word-adjacency graphs and extracts weighted clustering coefficients and weighted degrees from certain nodes in the word-adjacency network. An AA knowledge channel along these lines will be addressed in our own work as discussed in Section 4.

3 Corpus and Baseline Analysis

We started our investigation by examining the PAN-CLEF 2019 cross-domain AA dataset[1], and by comparing the results obtained by the baseline systems provided. This analysis is described as follows.

The PAN-CLEF 2019 AA development dataset conveys 20 problems written four languages (English, French, Italian and Spanish), with nine candidate authors per problem, seven documents per candidate and an average of 4500 characters per document.

[1] https://pan.webis.de/clef19/pan19-web/author-identification.html

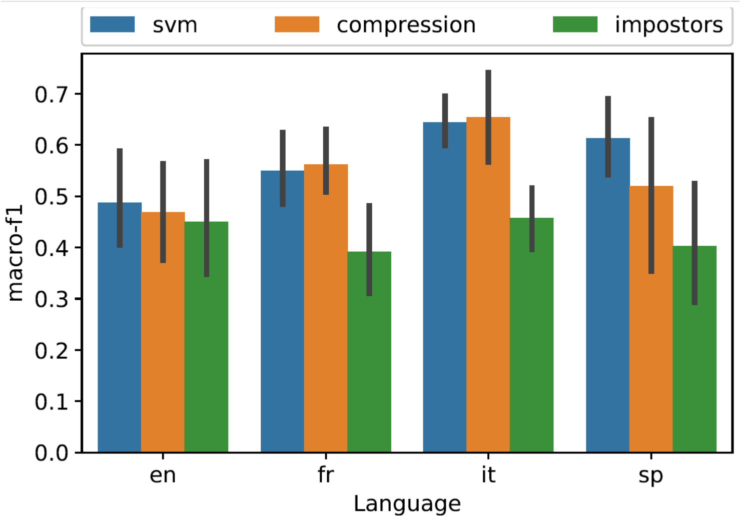

Figure 1. PAN-CLEF 2019 baselines performance according to target language.

The shared task organisers also provided three baseline systems, namely, compression models [20,11], the Impostors method [9], and a SVM classifier based on character trigrams. Further details are provided in [13]. Figure 1 presents a comparison between macro F1 scores obtained from the three baseline systems for each target language.

From Figure 1 we notice that the SVM classifier has the best overall performance among the three baseline systems. Moreover, we notice that the three systems obtained similar results in the case of the English dataset.

Figure 2 presents a comparison among the same baseline methods according to the number of unknown documents under consideration.

From Figure 2 we notice that the proportion of unknown texts in each dataset, or openness of the AA task, has a considerable impact on the performance of all models. This confirms the general intuition that open-set AA is more challenging than closed-set AA.

4 Current Work

As in [2], our current approach to AA assumes that evidence of an author's identity may be found in multiple layers of morphological, syntactic and semantic knowledge. These layers may be modelled as knowledge channels that use character- and word-based n-grams as their main source for feature extraction [4]. Channels of this kind tend to be relatively independent from each other, that is, the information captured by one channel may not necessarily be captured by another.

Based on these observations, we follow the work in [2] and address the AA task by making use of multiple models combined as an ensemble of classifiers. More specifically, our current approach extends the ensemble method in [2] by adding a fourth

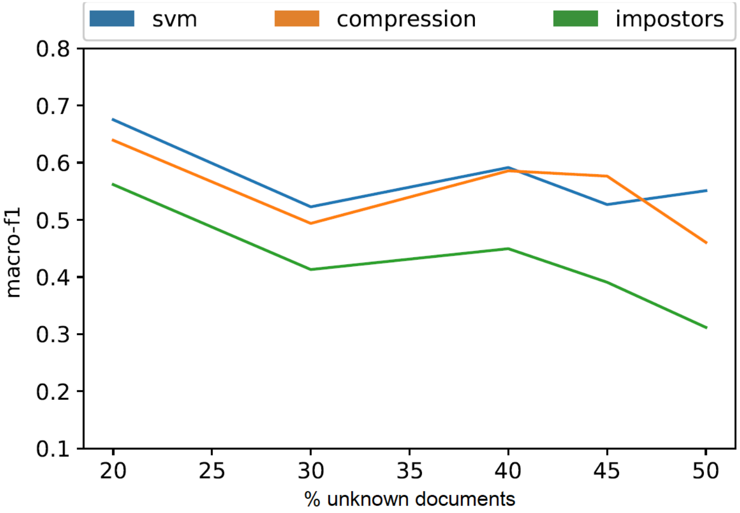

Figure 2. PAN-CLEF 2019 baselines performance according to the number of unknown documents.

module to the existing set of channels (*Std.wordN*), *Std.charN*), and *Dist.charN*, cf. previous section) and proposes further adjustments for the open-set AA setting.

4.1 A Character Ranking Model for AA

Language models are central to a wide range of natural language processing tasks. Accordingly, many studies have attempted to estimate the probability of a word (or character) appearing after a given symbol [4]. N-grams and recurrent neural networks [1,16] are the most well-known methods of this kind.

Of particular interest to the present work, language models may be represented as a character adjacency graph, in which the degree of influence of each node may help capture the (most influential) character sequences that denote a particular author. Influence may be measured, for instance, by using the PageRank algorithm [10,15]. In this case, the influence of a node is defined by the equation 1, in which N is the number of nodes, α is the original alpha factor, and M is the set toward p_i points to.

$$PR(p_i) = \frac{1-\alpha}{N} + \alpha \sum_{p_j \in M(p_i)} \frac{PR(p_j)}{L(p_j)} \tag{1}$$

Using this method as a basis, we envisaged a character ranking model for AA, hereby called *Rank.char*, that computes character adjacency graphs and uses PageRank to select the most influential characters of a set of documents of a given author. For instance, the word ‘the’ gives rise to three nodes t, h and e, and two edges $t \rightarrow h$ e $h \rightarrow e$.

Once the adjacency graph is computed, symbols of frequency lower than five are removed, and the resulting structure is submitted to the PageRank algorithm to determine

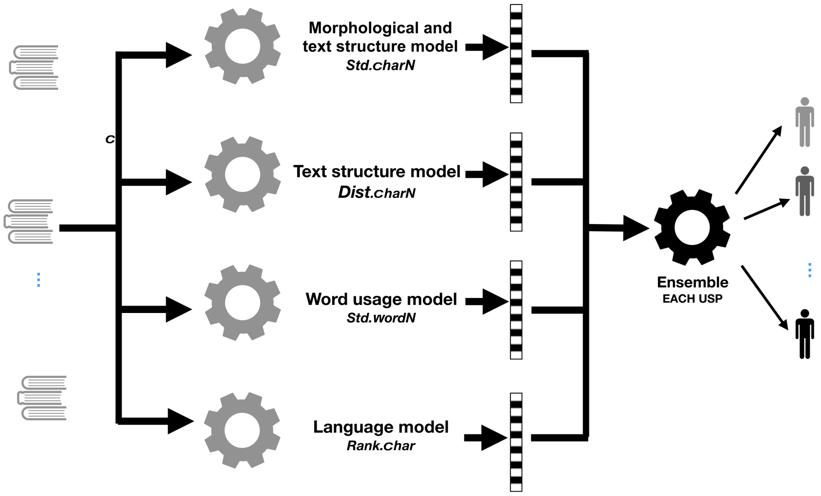

Figure 3. Ensemble cross-domain AA architecture

its most influential nodes. The algorithm is executed with a maximum of 500 iterations, and an alpha value set to 0.85. The output - a matrix of size $|d|, |v|$ where d is the set of documents and v is the corpus vocabulary - is then fed into the AA pipeline.

4.2 A Multi-channel AA Model

There are many possible strategies for combining the outputs of a set of classifiers. Among these, the most common are averaging, soft voting and hard voting. Averaging simply averages the predictions made by each classifier and chooses the class with higher probability. In hard voting, the majority vote is used as the final decision and, in soft voting, a weighted vote is considered.

In the present work we follow [2] and consider the use of a soft voting method in which the probabilities produced by a set of classifiers are concatenated and taken as an input to a softmax logistic regression model. This strategy is motivated by similar methods commonly applied in convolution neural network learning, in which multiples filters are applied to a stream of text, and subsequently combined by using a softmax layer. In the present AA setting, this method allows full filter (or channel) optimisation with the benefits of soft voting, which may be particularly suitable to scenarios with restricted number of text samples per author.

Our resulting architecture is illustrated in Figure 3. The first three channels are similar to those in [2], whereas the last channel (*Rank.char*) represents our current extension.

The output of the ensemble method is a matrix of probabilities conveying d rows representing documents and a columns representing authors, in which d_ij is the probability of a document d_i belong to an author a_j. The openness aspect of the AA task

Module	**Parameters**	**Possible values**
Rare symbols	Min corpus frequency	[0.01, 0.05, 0.1, 0.5]
	Max corpus frequency	[0.25, 0.50, 0.90, 1.0]
TF-IDF	Analyser	Std.charN, Dist.charN, Std.wordN, Rank.char
	N-gram	(1,1) to (2,5)
	IDF	Normal, smoothed, none
PageRank	Alpha	[0.1, 0.5, 0.85, 0.90, 0.99]
Transformation	Document normalisation	L1, L2
	Scaling	MaxAbsScaler, StandardScaler
	PCA percentage of explained variance	0.99 (Fixed)
Classifier	Logistic regression	Softmax with Newton-cg

Figure 4. Model pipeline parameters

Module	Parameters	Optimal values
Extraction	N-gram	Std.charN (2,5) Dist.charN (2,5) Std.wordN (1,3) Rank.char (2,2)
	Alpha	0.9
	Min corpus frequency	0.01
	Max corpus frequency	1.0
	IDF	Smoothed
	Document normalisation	L2
Transformation	Scaler	Standard Scaler

Figure 5. Model pipeline parameters optimal values

at PAN-2019 (i.e., the fact that an input text may not belong to any of the candidate authors) is dealt with by assigning the unknown author (<*UNK*>) label to the input text when the standard deviation of the corresponding row is below a 0.05 threshold.

5 Evaluation

Model parameters were set by using the PAN-CLEF 2019 development dataset as follows. Features were scaled using Python MaxAbsScaler transformer, and dimensionality reduction was performed by using a standard PCA implementation. PCA also helps remove correlated features, which is particularly useful in the present case because our models make use of variable length feature concatenation. The resulting feature sets were submitted to multinomial logistic regression by considering a range of possible alternative values as summarised in Figure 4.

Optimal values for each pipeline were determined by making use of grid search and 3-fold cross validation using an ensemble method. The optimal values that were selected for training of our actual models are summarised in Figure 5. In this summary, a sequence as in, e.g., Start=2 and End=5 is intended to represent the concatenation of subsequences [(2, 2),(2, 3),$\cdots$,(4, 3),(4, 5)], assuming that Start is not greater than End.

In addition to the main experiments presently reported, a large number of alternatives were considered as well. These included the use of BM25 and one-hot represen-

Table 1. Macro F1 results based on the PAN-CLEF 2019 AA development corpus

	Baselines			Channels				Ensembles	
#	Comp	Imp	SVM	Std.charN	Std.wordN	Dist.charN.	Rank.char	EACH-USP	current
01	0.66	0.66	0.67	0.69	0.61	0.65	0.57	**0.86**	0.83
02	0.34	0.40	0.44	0.47	0.45	0.39	0.19	**0.51**	0.49
03	0.49	0.42	0.50	0.62	0.47	0.46	0.37	0.63	**0.67**
04	0.51	0.49	0.35	0.40	0.31	0.24	0.20	**0.48**	0.46
05	0.36	0.27	0.49	0.42	0.46	0.39	0.23	**0.61**	0.53
06	0.67	0.56	0.67	0.65	0.45	0.61	0.47	0.74	**0.74**
07	0.52	0.43	0.50	0.57	0.46	0.41	0.52	0.56	**0.60**
08	0.49	0.36	0.51	0.56	0.37	0.46	0.25	**0.66**	0.61
09	0.61	0.26	0.61	0.67	0.44	0.52	0.13	0.70	**0.71**
10	0.52	0.35	0.46	0.53	0.28	0.44	0.41	0.59	**0.61**
11	0.56	0.42	0.62	0.62	0.37	0.64	0.48	**0.75**	0.73
12	0.50	0.51	0.58	0.62	0.37	0.47	0.46	0.61	**0.65**
13	0.72	0.49	0.67	0.73	0.44	0.53	0.43	0.76	**0.76**
14	0.78	0.53	0.60	0.81	0.42	0.57	0.57	0.70	**0.85**
15	0.72	0.33	0.74	0.71	0.38	0.42	0.34	**0.88**	0.77
16	0.67	0.60	0.74	0.81	0.66	0.53	0.48	0.72	**0.74**
17	0.62	0.32	0.57	0.62	0.53	0.45	0.18	**0.68**	0.65
18	0.65	0.53	0.69	0.80	0.61	0.54	0.28	**0.82**	0.79
19	0.41	0.28	0.55	0.52	0.52	0.41	0.28	**0.62**	0.60
20	0.24	0.29	0.52	0.44	0.31	0.22	0.12	**0.47**	0.44
Overall	0.55	0.42	0.57	0.61	0.44	0.47	0.35	0.67	0.67

tation for feature extraction, and the use of bagging, boosting, multi-layer perceptron, decision tree induction and other learning methods. All these results were however below those obtained by the present approach, and were therefore discarded.

6 Results

Table 1 presents macro F1 results based on the PAN-CLEF 2019 test dataset and evaluation software [12] as obtained by the original baseline systems, our four individual classifiers, the ensemble approach *EACH-USP* taken from [2], and by the current method. Baseline systems were trained with their default parameters, and all models were individually optimised by using the parameters described in Table 5. Best results for each problem are highlighted.

From these results we notice that the current approach keeps a relatively good performance overall. Figure 6 presents a comparison between macro F1 scores obtained from the SVM baseline, the char n-gram model with variable range *Std.charN*, and the EACH-USP and current ensemble methods for each target language.

From these results we notice that the use of *Rank.char* was more effective for the Italian language dataset. Moreover, the task seems to be more challenging in the case of the English dataset than for the other languages.

Finally, Figure 7 presents a comparison among the same methods according to the number of unknown documents under consideration.

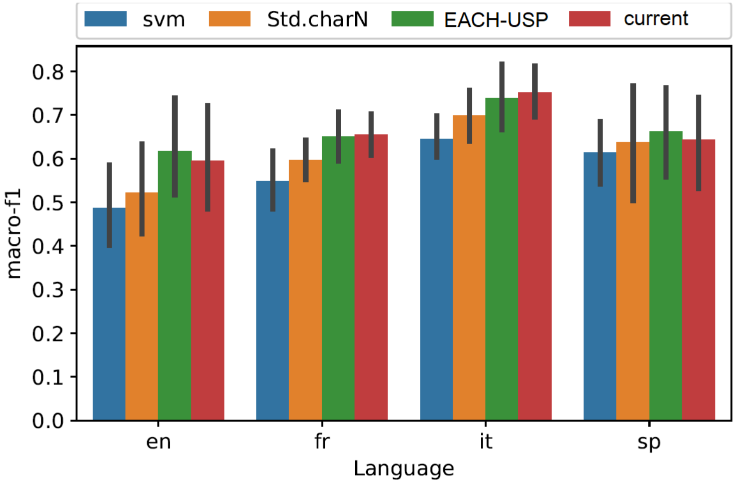

Figure 6. F1 Results according to target language.

Once again, we notice that the percentage of documents of unknown authors had a great impact over all system under evaluation regardless of other factors.

7 Final Remarks

This paper has proposed an extension to the work in [2] by presenting an approach for open-set cross-domain authorship attribution that relies on fully optimised char n-grams, word n-grams and char-ranking models. To this end, results obtained from the individual models as probability vectors were combined by making use of a soft voting ensemble method, and unknown authors were classified by considering the standard deviation of the final probability vector.

Our current results are generally superior to those obtained by the PAN-CLEF 2019 baseline systems, but were not generally superior to the work in [2]. Although the compact text representation provided by the current *Rank.char* model does help improve some of our results, the *Dist.charN* model from [2] remains the most useful knowledge source within this ensemble approach even in the present open AA setting.

As future work, we intend to experiment with other kinds of network influence methods, and further customise the PageRank algorithm [10,15] for the AA problem. The use of part-of-speech and embedding channels for AA is also to be investigated.

Acknowledgements

The second author received support by FAPESP grant nro. 2016/14223-0.

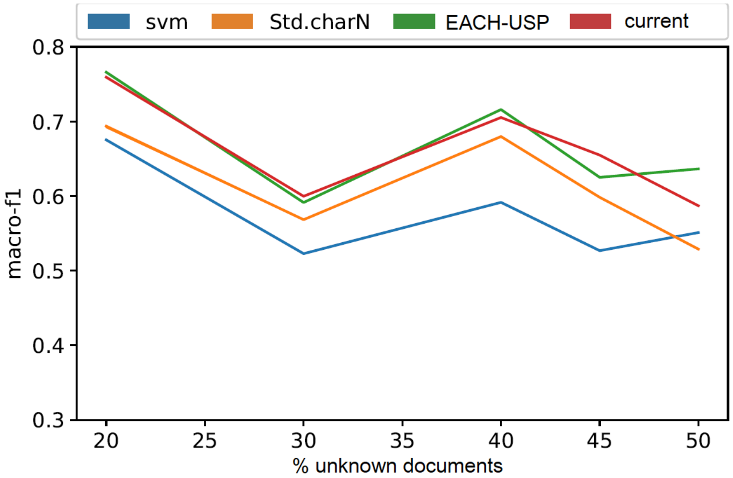

Figure 7. F1 results according to the number of unknown documents.

References

1. Bagnall, D.: Author identification using multi-headed recurrent neural networks. In: Jones G.J.F. Cappellato L., F.N.S.J.E. (ed.) CEUR Workshop Proceedings. vol. 1391, pp. 1–9. CEUR-WS (2015)
2. Custódio, J.E., Paraboni, I.: EACH-USP Ensemble Cross-domain Authorship Attribution: Notebook for PAN at CLEF 2018. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
3. Custódio, J.E., Paraboni, I.: Multi-channel Open-set Cross-domain Authorship Attribution . In: Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF-2019) (to appear). Lugano, Switzerland (2019)
4. Goldberg, Y.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers (2017)
5. Juola, P.: An overview of the traditional authorship attribution subtask. In: CLEF 2012 Evaluation Labs and Workshop, Online Working Notes, Rome, Italy, September 17-20, 2012 (2012), http://ceur-ws.org/Vol-1178/CLEF2012wn-PAN-Juola2012.pdf
6. Kestemont, M.: Function Words in Authorship Attribution From Black Magic to Theory? In: 3rd Workshop on Computational Linguistics for Literature (CLfL 2014). pp. 59–66 (2014)
7. Kestemont, M., Tschugnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2018)
8. Koppel, M., Schler, J., Argamon, S.: Computational Methods in Authorship Attribution. Journal of the Association for Information Science and Technology 60(1), 9—-26 (2009)
9. Koppel, M., Seidman, S.: Detecting pseudepigraphic texts using novel similarity measures. Digital Scholarship in the Humanities 33(1), 72–81 (2018)

10. Page, L., Brin, S., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab (November 1999), http://ilpubs.stanford.edu:8090/422/, previous number = SIDL-WP-1999-0120
11. Potthast, M., Braun, S., Buz, T., Duffhauss, F., Friedrich, F., Gülzow, J.M., Köhler, J., Lötzsch, W., Müller, F., Müller, M.E., Paßmann, R., Reinke, B., Rettenmeier, L., Rometsch, T., Sommer, T., Träger, M., Wilhelm, S., Stein, B., Stamatatos, E., Hagen, M.: Who wrote the web? revisiting influential author identification research applicable to information retrieval. In: Ferro, N., Crestani, F., Moens, M.F., Mothe, J., Silvestri, F., Di Nunzio, G.M., Hauff, C., Silvello, G. (eds.) Advances in Information Retrieval. pp. 393–407. Springer International Publishing, Cham (2016)
12. Potthast, M., Gollub, T., Wiegmann, M., Stein, B.: TIRA Integrated Research Architecture. In: Ferro, N., Peters, C. (eds.) Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF. Springer (2019)
13. Rocha, A., Scheirer, W.J., Forstall, C.W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A.R.B., Stamatatos, E.: Authorship Attribution for Social Media Forensics. IEEE Transactions on Information Forensics and Security 12(1), 5–33 (2017)
14. Sapkota, U., Bethard, S., Montes-y-Gómez, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA. pp. 93–102 (2015)
15. Schult, D.A.: Exploring network structure, dynamics, and function using networkx. In: In Proceedings of the 7th Python in Science Conference (SciPy. pp. 11–15 (2008)
16. Shrestha, P., Sierra, S., Gonzalez, F., Rosso, P., Montes-Y-Gomez, M., Solorio, T.: Convolutional Neural Networks for Authorship Attribution of Short Texts. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. vol. 2, pp. 669–674. Association for Computational Linguistics (ACL) (2017)
17. Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL-2017). Association for Computational Linguistics, Valencia, Spain (2017)
18. Stamatatos, E., Daelemans, W., Verhoeven, B., Stein, B., Potthast, M., Juola, P., Sánchez-Pérez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at PAN 2014. In: Working Notes for CLEF 2014 Conference, Sheffield, UK, September 15-18, 2014. pp. 877–897 (2014)
19. Stanisz, T., Kwapien, J., Drozdz, S.: Linguistic data mining with complex networks: A stylometric-oriented approach. Inf. Sci. 482, 301–320 (2019), https://doi.org/10.1016/j.ins.2019.01.040
20. Teahan, W.J., Harper, D.J.: Using compression-based language models for text categorization. In: Language modeling for information retrieval, pp. 141–165. Springer (2003)