UniNE at PAN-CLEF 2021: Authorship Verification

(Notebook for PAN at CLEF 2021)

Catherine Ikae[1]

[1]University of Neuchâtel, Switzerland, Avenue du 1er-Mars 26, 2000 Neuchâtel, Switzerland

Abstract

This work proposes to solve the Open-set author verification problem using a Term Frequency Inverse Document Frequency (TF—IDF) model with a majority-voting ensemble that incorporates five component models (machine-learning classifiers). The task is to verify if a given pair of text is written by the same or different authors. The training sample contains verification cases from previously unseen authors and topics. Transforming this question into a similarity problem, we can determine whether one or two authors have written a given text pair. Evaluation with 800 unigram features shows an overall performance of AUC = 0.9041, c@1 = 0.7586, F1—score = 0.8145, F_0.5u = 0.7233, Brier =0.8247, leading to an overall score = 0.8050.

Keywords

Author verification, Ensemble classifier, TF—IDF, Open-set author verification

1. Introduction

The increase in the volume of online text in communication, blogging, messaging, commentaries and entertainment content has generated the need for verification and authentication of authorship of the corresponding message. This is crucial in application areas such as analysis of anonymous emails for forensic investigations [1], verification of historical literature [2] continuous authentication used in cybersecurity [3], detection of changes in writing styles with Alzheimer patients [4].

Authorship verification is the application of linguistic style learning to detect whether two or more texts have been written by the same person or by more than one person [5]. By using prior information from the training dataset, we model the style representing the same author text as well as different author text used to construct a classifier that can be used to classify previously unseen text.

In open-set verification, the true author could be absent from the training set. Thus the system cannot generate a stylistic representation for each distinct author. So, the main question to be solved is to determine the level of similarity between two stylistic representations to reach the decision that this pair of texts has been written by the same author.

As the decision must be based on the author style, one can consider extracting stylistic features from each text. To achieve this, linguistic features reflecting the style must be extracted from the training dataset. By applying these selected features to the test dataset, the representation

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

catherine.ikae@unine.ch (C. Ikae)

of each pair of text is possible. In a second step, a classifier must compute a degree of similarity upon which the final decision can be taken.
In this work, we use the term frequency–inverse document frequency (TF−IDF) to determine useful features to discriminate between distinct authors [6]. For this purpose, we create a model using an ensemble of five machine-learning classifiers. By using this method, we take into account all vocabulary from all texts extracted by n-grams (of words or letters) weighted by TF−IDF, and using only a small fraction of them to perform the classification, we determine the optimal performance of the classifier.
The rest of this paper is organized as follows. Section 2 describes the text datasets while Section 3 describes the features used for the classification. Section 4 explains the similarity measure and Section 5 depicts some of our evaluation results. A conclusion draws the main findings of our experiments.

2. Corpus

The corpus consists of data obtained from fanfiction.net, a sharing platform for fanfiction that comes from various topical domains (or 'fandoms') [7] [8]. The contents are mainly fictional texts produced by non-professional authors in the tradition of a specific cultural domain (or 'fandom'), such as a famous author or a specific influential work. Fanfiction is now abundantly available on the internet, as the fastest growing form of online writing providing a platform for data collection. This corpus contains 52,590 text pairs (denoted problems) from which 27,823 pairs correspond to the same author and 24,767 are pairs written by two distinct persons. Each text excerpt contains, in mean, 2,200 word-tokens.
Based on the training sample of the entire corpus, Figure 1 depicts the top 25 most frequently used words.
To quantify the differences and similarities that occur when considering same author text pairs and different author pairs we use the technique of shift graphs. In shift graphs, words are sorted by their absolute contribution to the difference between text pairs. Word shifts quantify how each word contributes to the difference between two text pairs [9].
Figure 2 shows the relative occurrence frequency difference between tokens occurring in the same author pairs. In this graph, the words appear in decreasing order of their occurrence frequency. As one can see, there are only two tokens with a large difference in this text namely the two pronouns I, and she. Otherwise the rest of the tokens appear with small differences.
Figure 3 represents the same information as Figure 2 but with a pair of messages written by two distinct authors. In this case, one can observe that several tokens present large frequency differences (e.g., lola, joseph, said, tone, with, hikaru, normal). The presence of such numerous large differences must be interpreted as evidence of the presence of more than one author. For this reason, we use a difference vector to encode the data.

3. Feature Selection

To determine whether two text chunks have been written by the same author, we need to determine a text representation that can characterize the stylistic idiosyncrasies of each possible

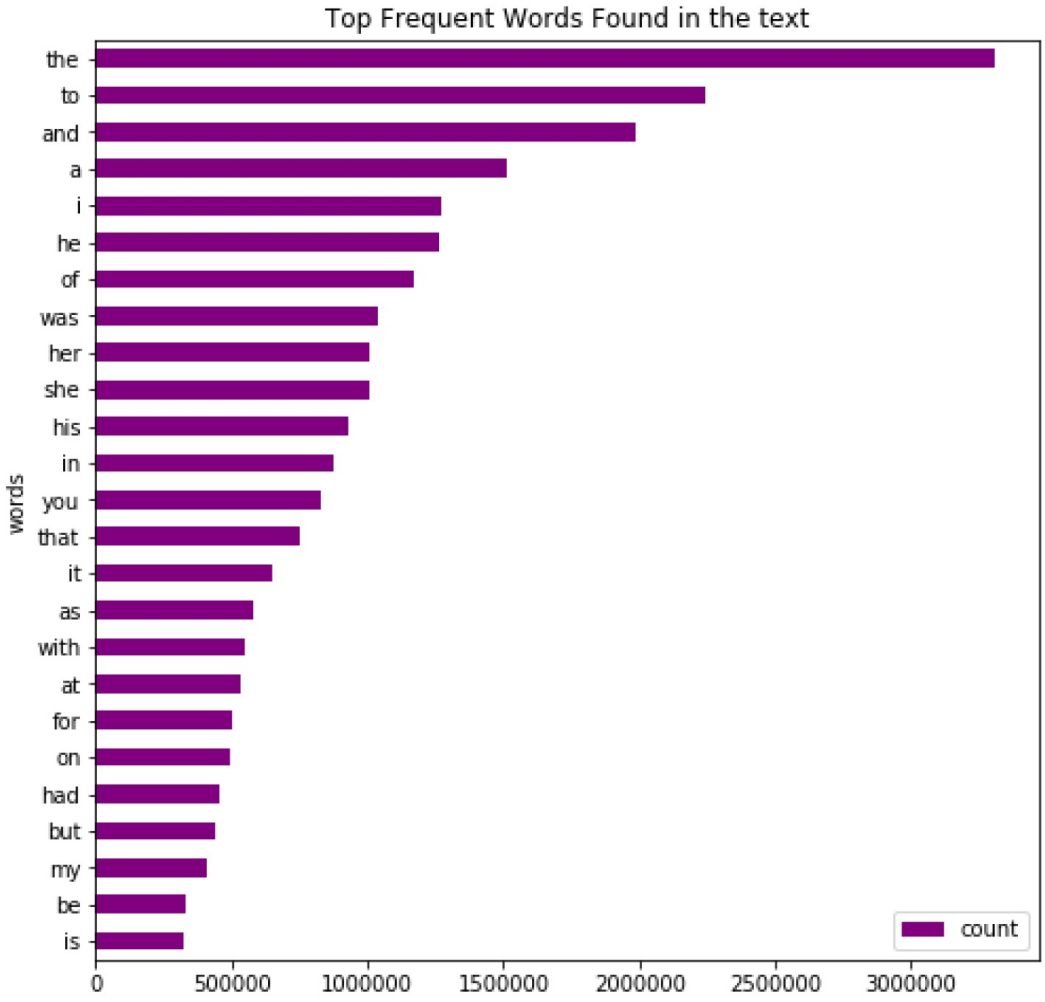

Figure 1: Word frequency distribution in the corpus Sample

author. Various text surrogates have been suggested, some focusing more on stylistic aspects, other on semantics (text vectorization).

As a simple and fast solution, and knowing that we are working with 52,590 text pairs, we will focus on the word uni-gram. In addition, each of them must have a weight computed according to the frequency (TF) which measures how frequently a term occurs in a document and the inverse document frequency (IDF) reflecting how important a term is compared to the entire corpus [10] [11].

TF−IDF is a statistical measure used in information retrieval and text mining that quantifies the importance of a word in a document by evaluating how relevant a word is to a document in

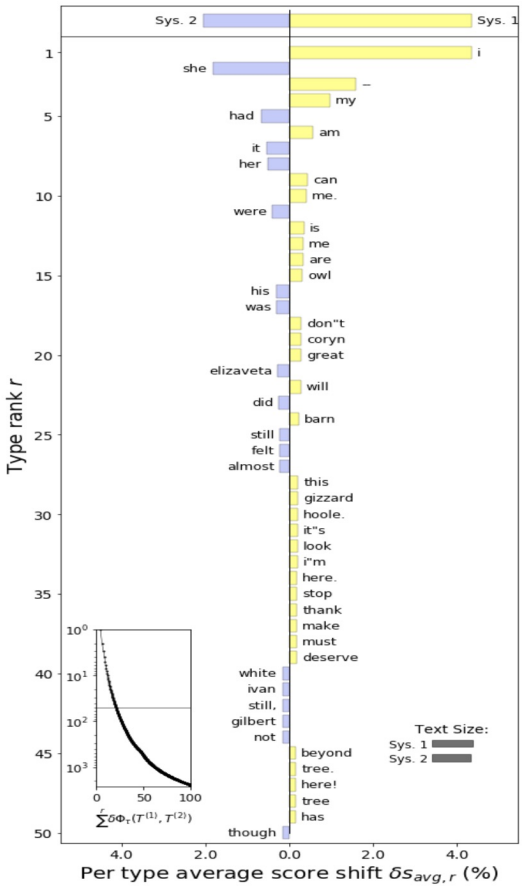

Figure 2: Same Author Pairs

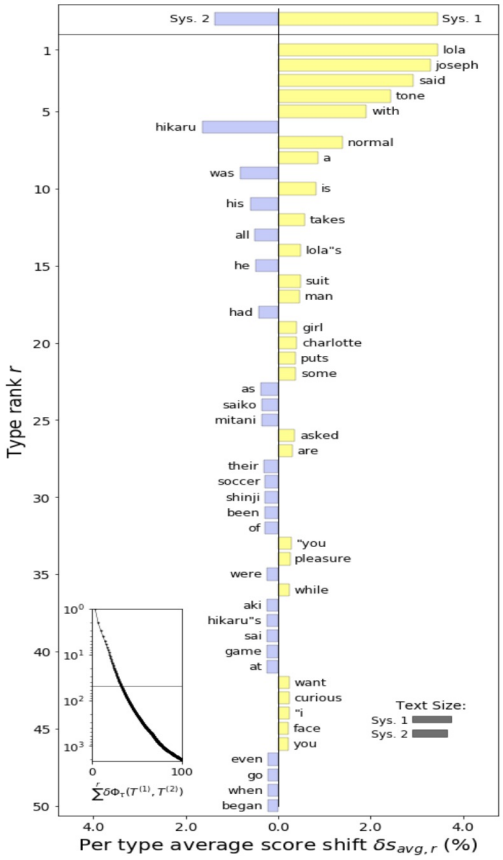

Figure 3: Different Author Pairs

a collection of documents [10] [12]. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents.

It works by increasing proportionally to the number of times a word appears in a document, but is offset by the number of documents that contain the word. So, words that are common in every document, such as “this”, “what”, and “if”, rank low even though they may appear many times, since they don’t mean much to that document in particular.

Applying our mathematical notation, the TF-IDF score for the word t in the document d from the document set (corpus) is calculated as follows:

tf(t,d) = number of occurrences of t in d/ number of tokens in d

df(t) = number of documents in which t occurs

D = Number of documents in the corpus

$$idf(t) = log(D/(df(t)+1))$$
$$tf-idf(t,d) = tf(t,d) * log(D/(df(t)+1))$$
$$TF-IDF = tf(t,d) * log(D/(df(t)+1))$$

An n-gram is a sequence of n-words in a sentence. Here, n is an integer which stands for the number of words in the sequence. For example, if we put n=1, then it is referred to as a uni-gram. For our vectorization we apply the uni-gram of TF−IDF for term weighting. Then, based on the weight associated with each term, one can apply a feature extraction by selecting the top k words having the largest TF−IDF value.

4. Ensemble Classifier

Ensemble learning could improve the effectiveness of isolated machine learning systems by combining several models. Such a combined approach should produce better predictive performance compared to a single model. In this view, democracy is viewed as a better system than the tyranny of a single classifier [13].
Our Ensemble model trains different classifiers including:

1. Linear Discriminant Analysis (LDA) finds a linear combination of features that separates two or more classes of objects in order to classify them [14];
2. Gradient Boosting (GB) which modifies weak learners to propose a strong learner [15];
3. Extra Trees (EF) is an ensemble learning technique which aggregates the results of multiple de-correlated decision trees constructed from the original training sample to obtain its classification result [16];
4. Support Vector Machine (SVM) determines the best decision boundary between vectors that belong to a given group and vectors that do not belong to it dividing the space into two subspaces [2];
5. Stochastic gradient descent (SGD) optimises an objective function equipped with the parameters of a model and updates parameters for each training sample [17];

Finally, we integrate these classifiers into an ensemble predictor to leverage complementary information of the feature representation method encoded by TF−IDF and classifiers. We used the scikit-learn Python machine learning library that provides an implementation of stacking for machine learning [6].

5. Evaluation

To conduct experiments with our approach to distinguish between same author pairs and different author pairs, we used a sample of the provided small training data set split as follows: 10,000 for training and 4,000 for testing. Each of the partitions used is balanced with an equal number of same author pairs and different author pairs.

As a performance measure, five evaluation indicators have been used. The area under the curve (AUC) which measures the ability of systems to assign higher scores to positive cases in comparison to negative cases. F1−score combines precision and recall into a unique value. c@1 measures the accuracy of binary predictions but also the ability of systems to leave difficult cases unanswered [1]. F_0.5u a measure that puts more emphasis on deciding same−author cases correctly [18]. Brier a score used for evaluating the goodness of (binary) probabilistic classifiers.

The proposed method is validated by comparing the AUC values of 15 classifiers. Based on the generated output, an ensemble is made by combining classifiers with consistent high AUC values. As seen on Table 1 increasing the unigram TF IDF values from k = 100 to 1000, we see consistent good performance in Linear Discriminant Analysis (LDA), Gradient Boosting (GB), Extra Trees (EF), Support Vector Machine (SVM), Stochastic gradient descent (SGD).

These chosen classifiers are combined using the hard voting (majority voting), every individual classifier votes for a class, and the majority determines the predicted class. With k=800, this is the point where most of the classifiers are at their maximum AUC values. The chosen classifiers had at least an AUC value of 0.87.

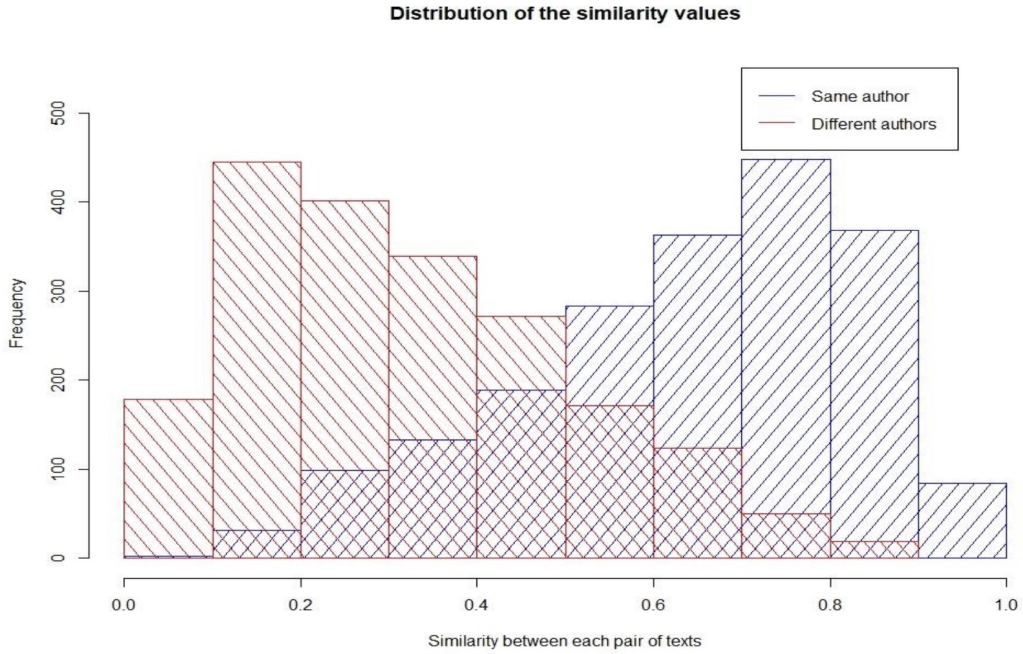

Figure 4: Distribution of AUC values for the two classes, same or distinct authors (k = 800)

With the results obtained from the Ensemble classifier, we view the distribution of AUC results for the two classes, namely “same author” and “different authors”, As one can see in Figure 4, “same author” distribution presents a higher similarity mean (mean: 0.64, sd: 0.19)

Table 1
Evaluation based on different feature sizes

	Number of TF−IDF unigram features									
Classifiers	100	200	300	400	500	600	700	800	900	1000
LDA	0.85	0.87	0.88	0.89	0.89	0.88	0.88	0.89	0.88	0.87
GradientBoost	0.84	0.86	0.87	0.87	0.87	0.88	0.87	0.88	0.88	0.88
ExtraTrees	0.84	0.85	0.86	0.86	0.86	0.86	0.86	0.87	0.86	0.86
KNN	0.76	0.76	0.77	0.76	0.74	0.73	0.73	0.70	0.70	0.68
GaussianNB	0.82	0.82	0.84	0.83	0.82	0.82	0.81	0.81	0.79	0.77
MultinomialNB	0.67	0.71	0.73	0.74	0.74	0.73	0.73	0.74	0.72	0.72
BernoulliNB	0.54	0.66	0.72	0.74	0.76	0.75	0.76	0.76	0.77	0.78
DecisionTree	0.63	0.64	0.64	0.64	0.63	0.63	0.64	0.64	0.64	0.63
RandomForest	0.83	0.85	0.86	0.86	0.87	0.87	0.87	0.87	0.87	0.87
LogisticReg	0.84	0.85	0.87	0.87	0.87	0.87	0.87	0.87	0.86	0.86
AdaBoost	0.82	0.83	0.85	0.85	0.85	0.85	0.85	0.84	0.85	0.84
Bagging	0.78	0.78	0.78	0.79	0.78	0.78	0.77	0.78	0.78	0.76
SGD	0.84	0.85	0.87	0.87	0.87	0.87	0.87	0.87	0.86	0.86
XGB	0.84	0.86	0.87	0.88	0.87	0.87	0.88	0.88	0.87	0.87
SVM	0.85	0.88	0.89	0.89	0.89	0.89	0.89	0.89	0.88	0.88
Ensemble	0.85	0.87	0.89	0.89	0.90	0.90	0.90	0.90	0.90	0.89

Table 2
Official Evaluation with (k = 800)

	800 TF−IDF unigram features					
Classifiers	AUC	c@1	F1−score	F_0.5u	Brier	overall
Ensemble (Early Bird)	0.904	0.71	0.769	0.685	0.821	0.778
Ensemble (Final)	0.9041	0.7586	0.8145	0.7233	0.8247	0.8050

representing mainly the higher values while the “different authors” distribution (mean: 0.32, sd: 0.18) mainly contains the lower values.

The final evaluation result is obtained on the TIRA platform [19] is exposed in Table 2. These results were obtained with a model trained on the 52,590 text pairs (small training set) and tested on 19,999 text pairs (official test set). By considering all results as computed by the proposed system, we achieve an overall score of 0.778 in the early bird submission without mapping any score to 0.5. The second run was made by adjusting some score to 0.5 based on the analysis of the similarity distribution. The values greater than 0.4 but less than 0.6 ($0.4 < x < 0.6$) where equated to 0.5 leading to an improved overall score of 0.8050.

6. Conclusion

This report has presented the proposed solution for open-set author verification at PAN 2021. Our approach is based on modeling the fandom pairs using word unigram TF−IDF features with a majority-voting ensemble that incorporates five machine-learning classifiers. With the ensemble classifier, we achieved an overall score of 0.8050. This simple approach proves to be

effective in distinguishing text written by the same author and text written by different authors. For future work, the idea is to include longer word n-gram models to enrich the current feature and to hopefully boost performance of the current technique.

References

[1] F. Iqbal, L. A. Khan, B. Fung, M. Debbabi, E-mail authorship verification for forensic investigation, 2010, pp. 1591–1598. doi:10.1145/1774088.1774428.

[2] C. Cortes, V. Vapnik, Support vector networks, Machine Learning 20 (1995) 273–297.

[3] T. Neal, K. Sundararajan, D. Woodard, Exploiting linguistic style as a cognitive biometric for continuous verification, in: 2018 International Conference on Biometrics (ICB), 2018, pp. 270–276. doi:10.1109/ICB2018.2018.00048.

[4] G. Hirst, V. Feng, Changes in style in authors with alzheimer's disease, English Studies 93 (2012) 357 – 370.

[5] M. Koppel, J. Schler, S. Argamon, Y. Winter, The "fundamental problem" of authorship attribution, English Studies 93 (2012) 284 – 291.

[6] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[7] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection, in: T. T. S. V. H. J. C. L. C. E. A. N. L. C. N. F. Avi Arampatzis, Evangelos Kanoulas (Ed.), 11th International Conference of the CLEF Association (CLEF 2020), Springer, 2020. URL: http://ceur-ws.org/Vol-2696/.

[8] M. Kestemont, I. Markov, E. Stamatatos, E. Manjavacas, J. Bevendorff, M. Potthast, B. Stein, Overview of the Authorship Verification Task at PAN 2021, in: CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[9] R. J. Gallagher, M. Frank, L. Mitchell, A. J. Schwartz, A. J. Reagan, C. Danforth, P. Dodds, Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts, EPJ Data Science 10 (2021) 1–29.

[10] G. Salton, C. Buckley, Term-weighting approaches in automatic text retrieval, Inf. Process. Manag. 24 (1988) 513–523.

[11] S. Qaiser, R. Ali, Text mining: Use of tf-idf to examine the relevance of words to documents, International Journal of Computer Applications 181 (2018) 25–29.

[12] M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, W. Daelemans, Authenticating the writings of julius caesar, Expert Syst. Appl. 63 (2016) 86–96.

[13] M. Bramer, Ensemble Classification, Springer London, London, 2013, pp. 209–220. URL: https://doi.org/10.1007/978-1-4471-4884-5_14. doi:10.1007/978-1-4471-4884-5_14.

[14] P. Xanthopoulos, P. M. Pardalos, T. B. Trafalis, Linear Discriminant Analysis, Springer New York, New York, NY, 2013, pp. 27–33. URL: https://doi.org/10.1007/978-1-4419-9878-1_4. doi:10.1007/978-1-4419-9878-1_4.

[15] R. E. Schapire, The strength of weak learnability, in: Machine Learning, 1990.
[16] P. Geurts, Extremely randomized trees, in: MACHINE LEARNING, 2003, p. 2006.
[17] L. Bottou, F. E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, SIAM Review 60 (2018) 223–311. URL: https://doi.org/10.1137/16M1080173. doi:10.1137/16M1080173. arXiv:https://doi.org/10.1137/16M1080173.
[18] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 654–659. URL: https://www.aclweb.org/anthology/N19-1068. doi:10.18653/v1/N19-1068.
[19] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, 2019, pp. 123–160. doi:10.1007/978-3-030-22948-1_5.
[20] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.
[21] F. Rangel, G. L. D. L. P. Sarracén, B. Chulvi, E. Fersini, P. Rosso, Profiling Hate Speech Spreaders on Twitter Task at PAN 2021, in: A. J. M. M. F. P. Guglielmo Faggioli, Nicola Ferro (Ed.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[22] F. Rangel, M. Franco-Salvador, P. Rosso, A Low Dimensionality Representation for Language Variety Identification, in: International Conference on Intelligent Text Processing and Computational Linguistics, Springer, 2016, pp. 156–169.