How Important are Linguistic Features for Style Change Detection?

Mahsa S. Shahshahani, Jaap Kamps, and Maarten Marx

University of Amsterdam, Amsterdam, The Netherlands
{m.shahshahani,kamps,maartenmarx}@uva.nl

Abstract. Author identification is a task defined to find out the author(s) of a piece of text. In order to achieve this goal, the system needs to diagnose if the text is written by a single author or it is collaboratively written by multiple authors. This task has been defined in PAN-2018 as a sub-task of Author Identification track. In this study we investigate models and the importance of different features in solving the mentioned problem. Ablation study is performed to analyze the impact of noun phrases, verbs, and punctuation marks on the performance of classification of documents to single- or multi-authored categories. We use three datasets for feature analysis: PAN-2018, PAN-2019, and a dataset we constructed using Canadian Parliamentary debates.

Keywords: Authorship Attribution · Style Change Detection · CLEF PAN

1 Introduction

A lot of our work is on political data, such as the proceedings of national parliaments and provincial and local government. In such a setting, it is not only important *what* is said, but also *who* says it, and *to whom*, and in what role and what context [3]. More broadly, one of the greatest challenges to our current information society is the problem of provenance and attribution of information, as "fake news" and "post-truth" news are a major threat to the fundamental principles underlying our society. Authorship attribution tools can make a great potential contribution to help professionals and civilians assess the value of information.

Author identification is a task defined to find the author of a piece of text and is a major technique in plagiarism detection. Sometimes the document does not have a single author and is written by a collaboration of two or more authors. In this case, the model should be able to identify all of the authors and ideally reveal the exact parts of text that is written by each specific author. Reaching this goal requires diagnosing if the document is a single- or multi-authored one beforehand. Style Change Detection task has been defined as a sub-task of Author Identification track at PAN-2018 with this goal. An example for a multi-authored document is shown in Figure 1. In this case the document has two authors and each of them write parts of the texts only once. But, in some documents authors write multiple parts of a document. In both of this cases, the document should be classified as 'multi-authored'.

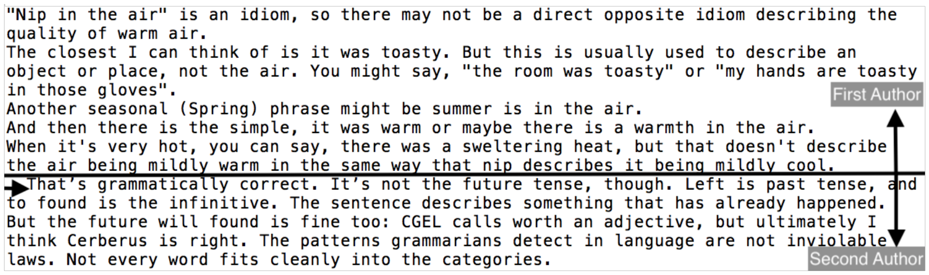

Fig. 1. An example of a multi-authored document from PAN-2018 (data from https://stackexchange.com/)

Authorship attribution is a task for which linguistic features are widely used, and seem of key importance for building effective classifiers [10]. Although user studies performed, there are related observations on the importance of different word categories in studies asking participants to guess the author or topic of a word cloud summaries [8, 9]. Yet the exact role of each feature is poorly understood. In this paper we will investigate the importance of a range of linguistic features commonly used for this style change detection task, and especially focusing on those used in best-performing system at PAN-2018. We also investigate their performance on Canadian Parliamentary debates [2] to see how it works when used for a totally different set of documents. Figure 2 shows an example of the Canadian Hansards data contained an extract transcript of parliamentary speech.

As a first step to try to understand the value of linguistic features, we use a simple classification model and look at the resulting feature weights. In addition, we do different ablation tests to study the effectiveness of document-, sentence-, word-, and character-level features. Our secondary aim to understand whether these results generalize to a wide range of authorship attribution tasks, rather than exploit unique aspects of the document style and genre or depends on particular choices in the test collection construction. For this purpose, we use both data from PAN-2018 [16] competitions for analysis, and construct a new dataset using Canadian Parliamentary debates [2] with a totally different documents and authors to see how currently using features perform in a different set of documents.[1]

The rest of this paper is structured as follows. You are now reading the introduction in Section 1. Next, in Section 2, we briefly discuss related work. Section 3 details the used linguistic features. This is followed by Section 4 detailing the experimental setup, and Section 5 summarizing the initial results. We end with discussion of the results in Section 6.

[1] We will make the parliamentary data available in PAN formats, compatible with the different authorship attribution tasks over the years, from www.politicalmashup.nl/. We are preparing a comprehensive data set covering multiple languages, and including standardized bibliographies of all speakers/authors, including age/gender/constituency.

Fig. 2. An example from a Parliamentary proceeding taken from: https://search.politicalmashup.nl.

2 Related Work

This section briefly discusses related work. Author identification is an old problem studied for many years. Stamatatos [22]'s survey explains many stylometric features including lexical and syntactic features widely used in literature for author identification subtasks.

The main thrust in recent years is the CLEF PAN lab. PAN-2011 [1] studied the author identification problem defined as mapping the uncertain texts onto their true authors, among the candidates having texts of uncertain authorship and texts from a set of candidate authors. This is a problem with a long history starting from 19th century [14]. PAN-2012 [6] added one other subtask to the 2011 setup: to cluster paragraphs of a document to two sets, i) paragraphs writtem by the main author, or ii) paragraphs written by any other authors.

Pan-2013 [7] and PAN-2014[23] defined the task as a binary decision: whether a questioned document is written by a specific author, having a set of document written by him/her. PAN-2015 [19] continued the same the task but the focus was on cross-genre which means that the questioned document may differ in genre or topic with the known set of documents.

PAN-2016 [18] and PAN-2017 [17] redefined the task as clustering documents regarding their authors. There was also a new subtask which asked participants to train models which could be able to recognize the exact positions in a text that authorship has

Table 1. Official Submissions to the PAN-2018 Style Change Detection Task

Model	**Accuracy**
Zlatkova, Kopev, Mitov, Atanasov, Hardalov, Koychev, and Nakov [24]	0.893
Hosseinia and Mukherjee [5]	0.825
Safin and Ogaltsov [20]	0.803
Khan [12]	0.643
Schaetti [21]	0.621

changed. At this proved very hard to predict exactly, PAN-2018 [16] continued this task in a simplified form: decide if a document is written by a single or multiple authors.

As the main experimental data in this paper is based on PAN-2018, we discuss the approaches for this version of the task in more detail. To put this into perspective, the performance of participant systems in PAN-2018 is reproduced in Table 1. Related work targeting multi-author documents is not frequent[16]. In [4] the author uses bayesian probability rule to propose a segmentation algorithm to parts may be written by different authors. Segments are clustered using some simple stylometric features. In [24] the winner system proposed for PAN-2018 competition is described. They use many stylometric features and combine many models including Support Vector Machine (SVM), Multi-linear perceptron, and random forests. Logistic regression model is applied on top of all these models to make the final model. In [5] hierarchical structure of a sentence is captured by using a only parse tree features and without using any lexical features. In [20] they use Random Forest classifier and Logistic Regression with 19 statistical features (ex. number of sentences). In [12] the document is splitted to two or more segments with the same number of sentences. Consecutive segments share one sentence in the middle. Each segment is compared to next segment using a number of frequency-based features. In [21] a character-based convolutional neural network (CNN) is applied. Each document is seen as consecutive characters. This would be the in put to an embedding layerthat captures context similarities of occurring characters. The second layer includes a couple of different convolutional networks with more than 20 filter to capture patterns of consecutive bigrams.

3 Linguistic Features

This section introduces the various linguistic features we will use in this paper.

Every author has its own specific writing style. Identifying the change between two writing style which means changing in authorship can be performed using linguistic features. We use frequently used features [e.g., 11, 13, 15, 25] for this task, and especially the ones applied in the best scoring PAN-2018 system [24]. We roughly grouped all features into four classes: character-level features, word-level features, sentence-level features, and document-level features. This categorization is not strict (e.g., is a question mark at the end of a sentence a character, word, or sentence level feature?) and we did it to be able to analyze features by group and refer them as a member of a category.

Character-level features

- Number of special characters including colons, semicolons, apostrophes, parenthesis, quotes, spaces, digits, and punctuation marks.

Word-level features

- Number of all-caps words
- Number of capitalized words
- Average word length

Sentence-level features

- Number of declarative sentences (end with '.')
- Number of question sentences (end with '?')
- Number of exclamatory sentences (end with '!')
- Number of short sentences (less than 100 characters)
- Number of long sentences (more than 200 characters)

Document-level features

- N-grams (unigrams, bigrams,..., and 5-grams) of words
- N-grams of part of speech tags
- Number of words per each part of speech tag (ex. number of pronouns)
- Frequent words: the occurrence frequency of some frequent words (including stop words and function words) has been used in [24].
- Vocabulary richness: computed as $\log_2 \frac{f(X)}{f(x)}$ in which f(x) is the frequency of word x in the whole document f(X) is the frequency of most-frequent word in the whole corpus [25]. The average vocabulary richness of all words in a document is used as the feature.
- Readability: the ease with which a reader can understand a written text. There are some tests and formulas proposed to compute readability of texts. Flesch reading ease, SMOG grade, Flesch-Kincaid grade, Coleman-Liau index, automated readability index, Dale-Chall readability score, difficult words, Linsear write formula, and Gunning fog are the readability metrics used in [24].

4 Experimental Setup

In this section we will discuss all the experiments and datasets we used to perform tests.

4.1 Datasets

We used two different series of datasets to carry out experiments. Information about datasets is shown in Table 2.

Table 2. Dataset Statistics

	Canadian Parliament	PAN-2018	PAN-2019
#Documents	5,148	4,472	3,054
#Authors	1–138	1–3	1–5
Average of #authors (multi-authored)	15	–	–
Length of documents (tokens)	1000–90,000	300–1,000	300–2,000

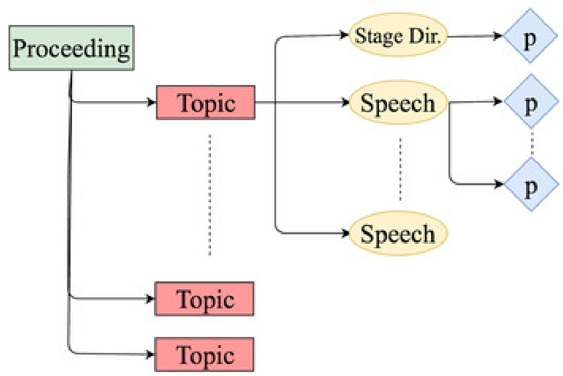

Fig. 3. Structure of the Political Mashup schema (reproduced from [2]).

PAN Datasets The style breach detection task has been defined in PAN-2017[2] as part of the competition. The dataset released for that year was created by using documents on 150 topics used at the TREC WEB Tracks from 2009 to 2011. TREC WEB Tracks has been constructed by professional writers searching for many predefined queries and writing a single document regarding to search results.

As style breach detection is a hard task to solve, in PAN-2018 the task was relaxed by skipping need for prediction the exact shift positions among witers of a multi-authored document. PAN-2018 dataset is constructed using user posts from 15 different question and answering sites in StackExchange. PAN-2019 dataset is created in the same way as PAN-2018.

In this study we used both PAN-2018 and PAN-2019 datasets for evaluation.[3] The PAN-2017 dataset is not used because of differences in its construction.

Canadian Parliamentary debates In order to do more evaluation we created a new dataset using minutes of Canadian Parliament [2] including meetings from 1901 to 1993. The structure of Parliamentary proceedings has been depicted in Figure 3. Each proceeding contains all of discussions in a single parliamentary meeting which may contain one or more topics. Discussions regarding each topic consist of many speeches by each parliamentary member and some non-speeches like votes from stage. Speeches

[2] https://pan.webis.de/clef17/pan17-web/author-identification.html

[3] We will mainly report the results for PAN-2018, as only the final evaluation set of 2019 is not available at the time of writing, and the development set showed broad agreement between the years.

and Stage directions may include one or more paragraphs. A short part of a proceeding document was shown in Figure 2 above. As it is obvious in the figure, this short part has multiple (three) authors.

We will now explain the way we constructed our dataset using the Canadian parliamentary proceedings:

Positive samples Multi-authored documents have been created by concatenating all paragraphs of speeches of all topics in each proceeding. Stage directions (everything outside of the spoken words, such as votes) have been removed. Moreover, data has been cleaned by removing any indication of addressing others (such as calling their names or calling them as "Mr./Mrs. Speaker"). We randomly chose a subset of positive samples in away we have the same number of single and multi-authored documents.

Negative samples Single-author documents has been constructed by concatenating all speeches of a specific author from different topics and proceedings.

4.2 Approach

We now explain the setup, parameters and tools we used in this study.

In all experiments an SVM classifier is used to classify the documents to two classes (single- or multi-authored). We chose SVM because it is possible to use feature weights in final trained model to perform feature analysis and discuss about the importance of features which is the goal of this study. All features are extracted using NLTK (Natural Language Toolkit)[4] developed in Python to work with textual data.

In order to extract features we use the same settings as [24]: we equally divide each document to three segments, calculate features for each part, calculate the difference between feature values for each two segments, and use the largest difference as the final value for that feature in the document. Our readability features are extracted using textstat package[5] distributed for Python. Stop words (using in frequent words feature) are extracted from NLTK.

5 Experimental Results

In this section, we explain the results of experiments we performed to investigate the importance of features mostly used for style change detection task. All analysis are reported on the PAN-2018 dataset, as we observed very similar results for PAN-2019.

Table 3 reports the effectiveness of the SVM model with all linguistic features. As should come as no surprize, the simple SVM classifier scores less than the complex system's of the PAN-2018 participants. It is still reasonably competitive, and more effective than the PAN-2018 baseline systems provided by the organizers [10]. We performed the same experiment for Canadian Parliamentary debates dataset. As it is shown in Table 3, the SVM model could reach a very high accuracy on this data.

Recall that our aim is not to achieve the highest possible score, but to get some initial insight in the relative importance of particular linguistic features. For this purpose, we

[4] https://www.NLTK.org/
[5] https://pypi.org/project/textstat/

Table 3. SVM Accuracy on Different Datasets + All features. (*Preliminary PAN-2019 scores are on the validation data.)

Dataset	**Accuracy**
PAN-2018	0.6268
PAN-2019*	0.7350
Parliamentary debates	0.9456

use SVM as our main classifier in this paper because we can directly interpret feature weights in SVM as indicators of importance of features.

As it is shown in Table 4 the most important features with highest weights in trained SVM model for both datasets are as follows (Number of quotes is a high weight feature only in PAN-2018 dataset):

- Vocabulary richness (document-level)
  Vocabulary richness shows how well one author writes in English. This is obviously an important indication an author writing style.
- Number of question sentences (sentence-level)
  As the PAN data comes from StackOverflow, a leading collaborative QA sity, in question-answer dialogues settings the number of questions in each segment can be an obvious indicator in authorship change.
- Number of quotes (character-level)
  Some authors may use more quotes in writing than others. This feature captures this writing style difference among different authors.
- Readability (difficult words metric) (document-level)
  Readability measure assess the understandability of a text. This feature may be widely different among authors regarding their writing abilities.

We also investigated the importance of features for the Parliamentary debates dataset, where the SVM model worked particularly well. We see that feature importance ranks based on their weight in trained model are roughly the same which shows that used features might be useful in other set of documents as well. The only major difference was in some character-level features such as number of quotations. To explain this, we should think about the way Parliamentary minutes are created. The Parliamentary minutes are written from all the discussions taken place in parliamentary meetings. So, they are not actual written documents by authors: Parliamentary members talk and another person writes down whatever he/she says. So it might explain why some character-level features such as number of quotations or spaces are not indicative features in Parliamentary debates dataset.

Our general conclusion is two fold. First, we see that the weight distribution over features is relatively flat: rather than a small number of features being crucial, almost all features seem to matter to a greater or lesser degree. Second, we see great similarity between the PAN and Parliamentary data, which suggests that authorship attribution experiments on PAN data generalize to other document genres and use cases.

Ablation tests Table 4 shows that all categories of features are somehow helping and important for the task, which raises the question on how independent the different fea-

Table 4. Feature Weights on PAN-2018 (on a scale of [0,3], Category is **D**ocument-, **S**entence-, **W**ord-, and **C**haracter-level)

Feature	Category	Weight	Feature	Category	Weight
vocabulary richness	D	2.2200	is	D	0.3379
question_sentences	S	2.1933	us	D	0.3205
quotes_count	C	1.9642	determiners	D	0.3180
difficult_words	D	1.9234	long_sentences	D	0.3127
parenthesis_count	C	1.8511	short_sentences	D	0.3076
automated_readability_index	D	1.7088	modals	D	0.2920
colon_count	C	1.5765	second	D	0.2885
all$_{c}aps$	W	1.2618	many	D	0.2767
rb rb	D	1.2464	coordinating_conjunctions	D	0.2689
comma_count	C	1.2223	part	D	0.2557
capitalized	W	1.1677	either	D	0.2521
spaces_count	C	1.0526	nns in	D	0.2324
exclamation_sentences	S	1.0326	even	D	0.2302
you	D	0.8479	and	D	0.2253
personal_pronouns	D	0.8300	nn prp	D	0.2052
adverbs	D	0.8019	every	D	0.2052
rb jj	D	0.7607	etc	D	0.1972
smog_index	D	0.7501	vbp rb	D	0.1921
however	D	0.7092	prp md	D	0.1788
vb dt nn	D	0.6960	like	D	0.1663
md vb	D	0.6581	vbz rb	D	0.1566
nn vbz	D	0.6331	used	D	0.1435
last	D	0.6141	dt nn nn	D	0.1314
adjectives	D	0.5981	whether	D	0.1178
pronouns	D	0.5885	since	D	0.1144
linsear_write_formula	D	0.5840	on	D	0.1135
jj in	D	0.5646	least	D	0.1122
vb prp	D	0.5341	dt jj	D	0.1050
digits	D	0.4450	be	D	0.1032
vbz dt	D	0.4250	could	D	0.0870
nn in nn	D	0.4202	already	D	0.0852
following	D	0.3785	dale_chall_readability_score	D	0.0759
of	D	0.3562	nn rb	D	0.0757
that	D	0.3454	something	D	0.0706
two	D	0.3402	semicolon_count	C	0.0690
enough	D	0.3385	punctuation_count	C	0.0602

tures are. To understand how the model works in extreme situations we performed two ablation tests.

In the first set of ablation tests we started by training the model using all features and remove each category of features in further experiments. Results are reported in Table 5. Character-level features seem to be important features according to Table 4, but here we see that removing them from the whole set of features does not drop the accuracy that much. We repeat the experiment using character-level features (and N-grams) only. The

Table 5. SVM Accuracy on PAN-2018 dataset

Features	Accuracy
All included	0.6268
− character-level	0.6219
− word-level	0.6174
− sentence-level	0.6102
− document-level (only N-grams)	0.5619

Table 6. SVM Accuracy on PAN-2018 Dataset: Ablation tests

Test	Accuracy
All included	0.6268
No verbs	0.5870
No noun phrases	0.5928
Shuffling	0.6116

model could reach to an accuracy of 0.6187 which is relatively high. This may show that features are not independent and capture overlapping information about each author's writing style: when we leave out character-level features other features capture enough information of the writing style.

In order to do more radical tests and find out about importance of different groups of words we performed second set of ablation tests.

- Remove all noun phrases: to examine significance of noun phrases in determining the writing style of an author we remove all noun phrases in documents.
- Remove all verbs: to explore impact of verbs in identification of an author's writing style we remove all verbs in documents.
- Shuffle all words in each sentence with keeping punctuation marks in place.
  In order to examine if defined features capture sentence structures we shuffled all words within a sentence, but we kept the punctuation marks in their place.

According to the first two tests, it seems that verbs are more important indicators of an author writing style than nouns. This can be explained by topic-specific nouns that are used by all the people talk about the topic and nouns does not have different shapes. But, verbs may be used in many ways and grammatical structures and there are many alternative verbs for each verb. This is consistent with user studies were word clouds with verbs were demonstrated to be more characteristic of authors than those with nouns [9].

In last test, we completely destroy the structure of sentences, but the accuracy of the trained model does not drop significantly. This is surprising as reshuffling all words would make the tasks far more complex for a human. When we shuffle word in each sentence, the only features that affect are n-grams (words and part of speech tags). Other features does not change because they are all based on counting which is not changed. N-grams of part of speech tags are used to give the opportunity to model to capture some grammatical structures of the text. But it seems that it is not enough for understanding grammatical skeleton of text. It is also consistent with Table 4, in which compositions of part of speech tags gain relatively low weights.

Our general conclusion is again two-fold. First, the linguistic features seem redundant and interdependent, suggesting the value of further investigating their relationships for the specific task at hand. Second, noun phrases (what is said) are less important than verbs (how it is said) and the sentence structure plays a relatively minor role. This last point also suggests that grammatical and semantic features are not captured in our model, and would explain the gain in effectiveness of recent complex neural models.

6 Conclusion

In this paper, we used SVM classifier to investigate the importance of mostly used features in style change detection problem especially the features used in the best-performing system in PAN-2018. Results on PAN-2018 and Canadian Parliamentary dataset show that character-level features are important features. Three other useful features in both datasets are vocabulary richness, readability (with 'difficult words' metric), and number of question sentences. Ablation tests show that currently used features do not capture sentence structures in a proper way which might be a strong indicator of an author's writing style. In [5], the authors proposed a model which only rely on grammatical structures and it was scored as the second-best system in PAN-2018. Studying the best way to use both linguistic and grammatical features or combine linguistic features explained in this paper with the neural network based grammatical model in [5] that may result in a better model for Style Change Detection task can be considered as future work. The dataset we constructed based on Parliamentary Debates is more useful for defined task in PAN-2019 which asks for indicating the number of authors for each document as it includes documents with a wide range of number of authors. So it may be used for more complex subtask defined under author identification task.

Acknowledgments This research was supported in part by the Netherlands Organization for Scientific Research (NWO, ACCESS project, grant # CISC.CC.016).

References

1. Argamon, S., Juola, P.: Overview of the international authorship identification competition at pan-2011. In: CLEF (Notebook Papers/Labs/Workshop) (2011)
2. Beelen, K., Thijm, T.A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., Lijbrink, S., Marx, M., Naderi, N., Rheault, L., Polyanovsky, R., Whyte, T.: Digitization of the Canadian parliamentary debates. Canadian Journal of Political Science 50, 849–864 (2017), https://doi.org/10.1017/S0008423916001165
3. Dehghani, M., Azarbonyad, H., Kamps, J., Marx, M.: On horizontal and vertical separation in hierarchical text classification. In: ICTIR'16: Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval. pp. 185–194 (2016), https://doi.org/10.1145/2970398.2970408
4. Giannella, C.: An improved algorithm for unsupervised decomposition of a multi-author document. Journal of the Association for Information Science and Technology 67(2), 400–411 (2016)

5. Hosseinia, M., Mukherjee, A.: A parallel hierarchical attention network for style change detection. Working Notes of CLEF (2018)
6. Inches, G., Crestani, F.: Overview of the international sexual predator identification competition at pan-2012. In: CLEF (Online working notes/labs/workshop). vol. 30 (2012)
7. Juola, P., Stamatatos, E.: Overview of the author identification task at pan 2013. In: CLEF (Working Notes) (2013)
8. Kaptein, R., Hiemstra, D., Kamps, J.: How different are language models and word clouds? In: ECIR'10: Advances in Information Retrieval, 32nd European Conference on IR Research. pp. 556–568 (2010), https://doi.org/10.1007/978-3-642-12275-0_48
9. Kaptein, R., Marx, M.: Focused retrieval and result aggregation with political data. Inf. Retr. 13(5), 412–433 (2010), https://doi.org/10.1007/s10791-010-9130-z
10. Kestemont, M., Tschuggnall, 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: Working Notes of CLEF 2018 (2018), http://ceur-ws.org/Vol-2125/invited_paper_2.pdf
11. Khan, J.A.: Style breach detection: An unsupervised detection model. In: CLEF (Working Notes) (2017)
12. Khan, J.A.: A model for style change detection at a glance. Working Notes of CLEF (2018)
13. Kuznetsov, M.P., Motrenko, A., Kuznetsova, R., Strijov, V.V.: Methods for intrinsic plagiarism detection and author diarization. In: CLEF (Working Notes). pp. 912–919 (2016)
14. Mendenhall, T.C.: The characteristic curves of composition. Science 9(214), 237–249 (1887)
15. Pervaz, I., Ameer, I., Sittar, A., Nawab, R.M.A.: Identification of author personality traits using stylistic features: Notebook for pan at clef 2015. In: CLEF (Working Notes) (2015)
16. Rangel, F., Rosso, P., Montes-y Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at pan 2018: multimodal gender identification in twitter. Working Notes Papers of the CLEF (2018)
17. 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. Working Notes Papers of the CLEF (2017)
18. 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. In: Working Notes Papers of the CLEF 2016 Evaluation Labs (2016)
19. Rangel Pardo, F.M., Celli, F., Rosso, P., Potthast, M., Stein, B., Daelemans, W.: Overview of the 3rd author profiling task at pan 2015. In: CLEF 2015 Evaluation Labs and Workshop Working Notes Papers. pp. 1–8 (2015)
20. Safin, K., Ogaltsov, A.: Detecting a change of style using text statistics. Working Notes of CLEF (2018)
21. Schaetti, N.: Character-based convolutional neural network for style change detection. Working Notes of CLEF (2018)
22. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)
23. Stamatatos, E., Daelemans, W., Verhoeven, B., Potthast, M., Stein, B., Juola, P., Sanchez-Perez, M.A., Barrón-Cedeño, A.: Overview of the author identification task at pan 2014. In: CLEF 2014 Evaluation Labs and Workshop Working Notes Papers, Sheffield, UK, 2014. pp. 1–21 (2014)
24. Zlatkova, D., Kopev, D., Mitov, K., Atanasov, A., Hardalov, M., Koychev, I., Nakov, P.: An ensemble-rich multi-aspect approach towards robust style change detection: notebook for pan at clef 2018. Working Notes of CLEF (2018)
25. Zu Eissen, S.M., Stein, B., Kulig, M.: Plagiarism detection without reference collections. In: Advances in data analysis, pp. 359–366. Springer (2007)