Overview of the Author Identification Task at PAN-2018

Cross-domain Authorship Attribution and Style Change Detection

Mike Kestemont,[1] Michael Tschuggnall,[2] Efstathios Stamatatos,[3] Walter Daelemans,[1]
Günther Specht,[2] Benno Stein,[4] and Martin Potthast[5]

[1]University of Antwerp, Belgium
[2]University of Innsbruck, Austria
[3]University of the Aegean, Greece
[4]Bauhaus-Universität Weimar, Germany
[5]Leipzig University, Germany

pan@webis.de http://pan.webis.de

Abstract Author identification attempts to reveal the authors behind texts. It is an emerging area of research associated with applications in literary research, cyber-security, forensics, and social media analysis. In this edition of PAN, we study two task, the novel task of cross-domain authorship attribution, where the texts of known and unknown authorship belong to different domains, and style change detection, where single-author and multi-author texts are to be distinguished. For the former task, we make use of fanfiction texts, a large part of contemporary fiction written by non-professional authors who are inspired by specific well-known works, to enable us control the domain of texts for the first time. We describe a new corpus of fanfiction texts covering five languages (English, French, Italian, Polish, and Spanish). For the latter, a new data set of Q&As covering multiple topics in English is introduced. We received 11 submissions for the cross-domain authorship attribution task and 5 submissions for the style change detection task. A survey of participant methods and analytical evaluation results are presented in this paper.

1 Introduction

In recent years, the authenticity of (online) information has attracted much attention, especially in the context of the so-called 'fake news' debate in the wake of the US presidential elections. Much emphasis is currently put in various media on the provenance and authenticity of information. In the case of written documents, an important aspect of this sort of provenance criticism relates to authorship: assessing the authenticity of information crucially relates to identifying the original author(s) of these documents. Consequently, one can argue that the development of computational authorship identification systems, that can assist humans in various tasks in this domain (journalism, law enforcement, content moderation, etc.), carries great significance.

Quantitative approaches to tasks like authorship attribution [42], verification [26], profiling [2] or author clustering [48] rely on the basic assumption that the writing style of documents is somehow quantified, learned, and used to build prediction models. It

is commonly stressed that a unifying goal of the field is to develop modeling strategies for texts that focus on *style* rather than *content*. Any successful author identification system, be it in a attribution setup or in a verification setup, must yield robust identifications across texts in different genres, treating different topics or having different target audiences in mind. Because of this requirement, features such as function words or common character-level n-grams are typically considered valuable characteristics, because they are less strongly tied to the specific content or genre of texts. Such features nevertheless require relatively long documents to be successful and they typically result in sparse, less useful representations for short documents. As such, one of the field's most important goals remains the development of systems that do not overfit on the specific content of training texts and scale well across different text varieties.

This year we focus on so-called fanfiction, where non-professional authors produce prose fiction that is inspired by a well-known author or work. Many fans produce fiction across multiple fandoms, raising interesting questions about the stylistic continuity of these authors across these fandoms. Cross-fandom authorship attribution, which is closely related to cross-topic and cross-genre attribution, is therefore the main focus of the cross-domain authorship attribution task.

Traditional models for authorship attribution are not applicable in the case where multiple authors are involved within a single document. Therefore, it is an important prerequisite to at first determine if a document is single- or multi-author. To this end, the style breach detection task at PAN 2017 aimed to find the exact border positions within a document where the authorship changes. Previous results have shown that the problem is quite hard [50], i.e., to identify the exact borders in terms of character position. Therefore, we substantially relaxed the task for PAN 2018 and broke it down to the simple question: Given a document, are there any style changes or not? An alternative formulation would thus be to predict whether a document is written by a single author or by multiple collaborators. In this sense, it is irrelevant to the task to identify the exact border positions between authors.

To be able to evaluate the submitted approaches sufficiently, a data set is needed which contains single as well as multi-author documents. Thereby a key requirement is that multi-author documents contain the same topic, as otherwise the task would be simplified (e.g., an author discussing about the first world war might be easily distinguished from a second one writing about programming languages by applying simple vocabulary analyses). Therefore we created a novel data set by crawling a popular Q&A network, containing millions of publicly available questions and answers regarding several topics. By applying multiple cleaning steps, we ensure that it represents a realistic and high-quality data set for the style change detection problem.

In what follows, after a brief review of previous work on these two task in the following section, Sections 3 and 4 discuss the two outlined tasks, respectively, including a discussion of its rationale, data set construction, performance measures, survey of submitted approaches, evaluation results, and their analysis.

2 Previous Work

Closed-set authorship attribution is a task with rich relevant literature [42, 29]. Two previous editions of PAN included corresponding shared tasks [1, 19]. However, they only examined the case where both training and test documents belong to the same domain, as it is the case for the vast majority of published studies in this area. Cross-domain authorship attribution has been sporadically studied in the last decade [3, 31, 37, 38, 39, 45, 46]. In such cases, training and test texts belong to different domains that may refer to topic, genre, or language. In this section we focus our on the construction of building suitable resources for evaluating cross-domain attribution methods.

The most frequent scenario examined in previous cross-domain attribution studies considers cross-topic conditions. To control topic, usually general thematic categories are defined and all texts are pre-assigned to a topic. For example, Koppel et al. uses three thematic categories (ritual, business, and family) of religious Hebrew-Aramaic texts [24]. Newspaper articles are considered by Mikros and Argiri [27] (classified into two thematic areas: politics and culture) and Stamatatos (classified into four areas: politics, society, world, and UK) [45]. Another approach is to use a controlled corpus where some individuals are asked to write texts on a specific, well-defined topic [47]. The latter provides fine-grained control over topic. On the other hand, the size of such controlled corpora is relatively small.

Another important cross-domain perspective concerns cross-genre conditions. In general, it is hard to collect texts by several authors in different genres. Kestemont et al. make use of literary texts (theater plays and literary prose) [21] while Stamatatos explores differences between opinion articles and book reviews published in the same newspaper [45]. Another idea is to use social media texts based on the fact that many users are active in different social networks (e.g., Facebook and Twitter) [31]. Finally, a controlled corpus can be built, where each subject (author) is asked to write a text in a set of genres (e.g., email, blog, essay) [47]. The most extreme case concerns cross-language conditions where training and test texts are in different languages [3]. A convenient source of such cases is provided by novels that have been translated to other languages hoping that the translator's preferences do not significantly affect the style of the original author. To the best of our knowledge, so far there is no cross-domain authorship attribution study using fanfiction texts.

With respect to intrinsic analyses of texts, PAN included several shared tasks in the last years. Starting from intrinsic plagiarism detection [33], the focus went from clustering authors within documents [48] to the detection of positions where the style, i.e., the authorship, changes [50]. In general, all those tasks imply an intrinsic, stylometric analysis of the texts, as no reference corpora are available. Thus, stylistic fingerprints are created that include lexical features like character n-grams (e.g., [43]), word frequencies (e.g., [16]) or average word/sentence lengths (e.g., [51]), syntactic features like part-of-speech (POS) tag frequencies/structures (e.g., [49]) or structural features such as indentation usages (e.g., [51]). Approaches specifically tackling the similar style breach detection task at PAN 2017 also utilize typical stylometric features such as bags of character n-grams, frequencies of function words, and other lexical metrics, processed by algorithms operating on top to detect borders [7, 22] or outliers [35].

In general, related work targeting multi-author documents is rare. While there exist several approaches for the related text segmentation problem, where a text is divided into distinct portions of different topics, only few approaches target a segmentation by other criteria, especially not by authorship. One of the first approaches in the latter direction that employs stylometry to automatically detect boundaries of authors of collaboratively written texts has been proposed by Glover and Hirst [10]. Nevertheless, the goal of detecting boundaries is not to reveal multiple authors, but to provide hints such that collaboratively written documents can be homogenized in terms of the global style. Further approaches include Graham et al. [12], who utilize neural networks with several stylometric features, and Gianella [9], who proposes a stochastic model on the occurrences of words to split a document by authorship.

With respect to the proposed style change detection task at PAN 2018, i.e., to solely separate single-author documents from multi-authored ones, no prior studies exist to the best of our knowledge.

3 Cross-domain Authorship Attribution

For this edition of PAN, a challenging and novel source of texts has been targeted that seems well suited to advance the field with respect to the style-content dichotomy that is so central to it: fanfiction [15]. In this section, after a brief introduction into this literary genre of writing, we describe how we constructed a data set from a collection of fanfiction works for the task of cross-domain authorship attribution. After that, the participants' submissions are surveyed, followed by an in-depth evaluation of their attribution performance.

3.1 Fanfiction

Fanfiction refers to the large body of contemporary fiction that is nowadays created by non-professional authors ('fans'), who write in the tradition of a well-known source work, such as the *Harry Potter* series by J.K. Rowling, that is called the 'fandom'. These writings or 'fics' engage in a far-reaching form of 'intertextuality': they heavily and explicitly borrow characters, motives, settings, etc. from the source fandom. The authors are typically 'amateurs' who do not seek any commercial gains. (A famous exception to this rule is the successful *Fifty Shades* trilogy by E.L. James, that was originally a *Twilight* fanfic.) In many cases, the legal status of these fanfics has been the subject of controversy, although many fans have stressed the 'transformative' status of their fics. Overall, one could make the distinction between 'transformative' fanfiction and 'affirmative' fanfiction—the latter staying relatively closer to the original texts in the fandom with respect to tone, style or storylines. At the other side of the spectrum, we find intense reworkings of the fandom that, apart from perhaps the names of the protagonists, have little in common anymore with the original fandom. Many texts are for instance pornographic in nature; this is especially true of the so-called 'slash' fics that focus on the (typically same-sex) encounter of two fictional characters, e.g. "Kirk/Spock".

There are various reasons why fanfiction is an interesting benchmark case for computational authorship identification. Most of the fanfiction is nowadays produced on

online platforms (such as fanfiction.net or archiveofourown.org) that are not strongly mediated or moderated, so that the fics in all likelihood accurately reflect the author's individual style. This is typically not the case with professionally published authors for which editorial interventions are a constant cause of worry. Many fans are moreover active across different fandoms. Because of the explicit intertextuality, it can be anticipated that the style of the original fandom texts—sometimes also called the 'canon'—has a strong influence on the fan's writings, because these often aim to imitate the style of the canon's original authors. An interesting example is for instance the *James Potter* series by Georges N. Lippert, an American writer whose children were so disappointed that the original Potter series has come to an end, that he decided to write a multi-volume continuation of the storyline, featuring Harry's son James as a protagonist. For instance, for Lippert, it is clear that he was maximally trying to faithfully reproduce Rowling's writing style.[1]

Fanfiction thus allows for exciting authorship research: do fanfiction authors succeed in imitating the author's stylome or does their individual fingerprint still show in the style of their fics? This question is especially relevant for fans that are active contributors across different fandoms: can we still identify texts by the same author, even when they are basing themselves on different canons? Naturally, such issues challenge the state of the art in computational authorship identification and can provide ideal benchmark data to test the robustness of state-of-art systems across different domains.

3.2 Task Definition

The task can be defined as *closed-set cross-fandom attribution in fanfiction*. Given a sample of reference documents from a restricted and finite set of candidate authors, the task is to determine the most likely author of a previously unseen document of unknown authorship. Documents of known and unknown authorship belong to different domains (fandoms). More specifically, all documents of unknown authorship are fics of the same fandom (target fandom) while the documents of known authorship by the candidate authors are fics of several fandoms (other than the target-fandom). The participants are asked to prepare a method that can handle multiple cross-fandom attribution problems.

In more detail, a cross-domain authorship attribution problem is a tuple (A, K, U), where A is the set of candidate authors, K is the set of reference (known authorship) texts, and U is the set of unknown authorship texts. For each candidate author $a \in A$, we are given $K_a \subset K$, a set of texts unquestionably written by a. Each text in U should be assigned to exactly one $a \in A$. From a text categorization point of view, K is the training corpus and U is the test corpus. Let D_K be the set of fandoms of texts in K. Then, all texts in U belong to a single (target) fandom $d_U \notin D_K$.

3.3 Data Set Construction

For this shared task, we have harvested a collection of fanfics and their associated metadata from the authoritative community platform *Archive of Our Own*, a project of the

[1] http://www.jamespotterseries.com/

Table 1. The cross-domain authorship attribution corpus.

	Language	Problems	Authors	Texts per author		Text length
			(subsets size)	training	test	(avg. words)
Development	English	2	5,20	7	1-22	795
	French	2	5,20	7	1-10	796
	Italian	2	5,20	7	1-17	795
	Polish	2	5,20	7	1-21	800
	Spanish	2	5,20	7	1-21	832
Evaluation	English	4	5,10,15,20	7	1-17	820
	French	4	5,10,15,20	7	1-20	782
	Italian	4	5,10,15,20	7	1-29	802
	Polish	4	5,10,15,20	7	1-42	802
	Spanish	4	5,10,15,20	7	1-24	829

Organization for Transformative Works.[2] We limited the material to fanfics in English (en), French (fr), Italian (it), Polish (pl), and Spanish (sp) that counted at least 500 tokens, according to the platform's own internal word count. Across all data sets, the 'Harry Potter - J. K. Rowling' fandom was typically the most frequent one. We therefore selected fics in this fandom as the test material and fics from all other fandoms as the training material. We included only material for authors that contributed at least one fic to the 'target fandom' and at least one fic to another, 'training fandom'. As such, the task was operationalized as a standard, closed-set attribution task, where all fics in the test material (belonging to the target fandom) had to be attributed to exactly one fan author in the training material.

For each language we constructed two separate data sets: a development set that participants could use to calibrate their system and an evaluation set on the final evaluation of the competing systems was evaluated (see Table 1). Importantly there was no overlap in authors between the development set and the evaluation set (to discourage systems from overfitting on the characteristics of specific authors in the development material). To maximize the comparability of the data sets across languages, we randomly sampled 20 authors for each language and exactly 7 training texts (from non-target fandoms) for each author. No sampling was carried out in the test material of each attribution problem. In other words, in each attribution problem K is equally distributed over the authors while U is imbalanced. No files shorter than 500 tokens were included and to normalize the length of longer fics, we only included the middle 1,000 tokens of the text. Tokenization was done using NLTK's 'WordPunctTokenizer'. All texts were encoded as UTF8 plain text. To enrich the number of attribution problems in each language, random subsets of candidate authors (5, 10, or 15) were selected. For the early-bird evaluation phase, only the attribution problems with a maximal (20) number of candidate authors were used.

[2] https://github.com/radiolarian/AO3Scraper

3.4 Evaluation Framework

There are several evaluation measures that can be used for this closed-set multi-class and single-label classification task. Given that, in each attribution problem, the texts of unknown authorship are not equally distributed over the candidate authors, we decided to use the macro-averaged F1 score. Given an authorship attribution problem, for each candidate author, recall and precision of the provided answers are calculated and a F1 score (i.e., their harmonic mean) is provided. Then, the average F1 score over all candidate authors is used to estimate the performance of submissions for that attribution problem. Finally, submissions are ranked according to their mean macro-averaged F1 score over all available attribution problems. In addition, we also examine macro-averaged precision, macro-averaged recall, and micro-averaged accuracy to provide a more detailed view of submissions' performance.

Following the practice of previous PAN labs, software submissions were required. All submissions are deployed and evaluated in the TIRA experimentation platform [34]. Participants can apply their software to the evaluation data sets themselves. However, only PAN organizers can view the actual evaluation results. Moreover, the submitted software has no access to the internet during its run to avoid data leaks and to ensure a blind evaluation. Beyond evaluation measures, the runtime of submitted software is recorded.

To estimate the difficulty of a cross-domain authorship attribution problem and to provide a challenging baseline for participants, we developed a simple but quite effective approach already used in previous work for similar purposes [38, 37, 46]. This method is based on character n-gram features and a support vector machine (SVM) classifier. First, all character *n*-grams that appear at least f_t times in the training (known authorship) texts of an attribution problem are extracted and used as features to represent both training and test texts. Then, an SVM with linear kernel is trained based on the training texts and can be used to predict the most likely author of the test texts. As shown in previous work, this simple model can be very effective in cross-domain conditions given that the number of features is appropriately defined for each specific attribution problem [45]. However, in this shared task, we use a simple version where the cutoff frequency threshold (i.e., practically, this defines the number of features) is the same for any attribution problem. More specifically, we use $n = 3$ (i.e., character trigrams) and $f_t = 5$. This approach is called PAN18-BASELINE in the rest of this paper. A Python implementation of this approach[3] has been released to enable participants experiment with its possible variations. This implementation makes use of the scikit-learn library [32] and its SVM classifier based on one-vs-rest strategy and $C = 1$.

3.5 Survey of Submissions

We received 11 submissions from research teams from several countries (Austria, Brazil, Germany, Iran (2), Israel (2), Mexico, the Netherlands, Spain, and Switzerland). In addition, 9 out of 11 submitted approaches are described in working notes papers. Table 2 provides an overview of the received submissions and the PAN18-BASELINE.

[3] https://pan.webis.de/clef18/pan18-code/pan18-cdaa-baseline.py

Table 2. A survey of submissions (ranked alphabetically) to the cross-domain authorship attribution task. The following terms are abbreviated: instance-based (i-b), profile-based (p-b), language-specific (l-s), neural network (NN), Echo State Network (ESN), Support Vector Machine (SVM).

Submission		Features	Weighting / Normalization	Paradigm	Classifier	Parameter settings
Team	Reference					
Custódio and Paraboni	[6]	char & word n-grams	TF-IDF	i-b	ensemble	global
Gagala	[8]	various n-grams	none	i-b	NN	global
Halvani and Graner	[14]	compression	none	p-b	similarity	global
López-Anguita et al.	[25]	complexity	L2-norm.	i-b	SVM	l-s
Martín dCR et al.	[4]	various n-grams	log-entropy	i-b	SVM	l-s
Miller et al.	[13]	various n-grams & stylistic	TF-IDF & TF	i-b	SVM	global
Murauer et al.	[28]	char n-grams	TF-IDF	i-b	SVM	local
PAN18-BASELINE		char n-grams	TF	i-b	SVM	global
Schaetti	[41]	tokens	embeddings	i-b	ESN	local
Yigal et al.	[13]	various n-grams & stylistic	TF-IDF & TF	i-b	SVM	global

Submissions without a working notes paper: Saeed Mosavat; Hadi Tabealhojeh

As can be seen, n-grams are the most popular type of features to represent texts in this task. More specifically, character and word n-grams are used by the majority of the participants. Martín dCR et al. also explore typed character n-grams [37] and function word n-grams [44]. Custódio and Paraboni apply text distortion [46] and then extract character n-grams to highlight the use of punctuation marks, numbers, and characters with diacritics (e.g., ó, é, etc). Part-of-speech (POS) n-grams are used by Gagala, Miller et al., and Yigal et al., while López-Anguita et al. report that they experimented with this type of features, but did not manage to include in their final submission. Other types of explored features are complexity measures [25], word and sentence length, and lexical richness functions [13]. It has to be noted that the approach of Halvani and Graner makes use of text compression methods and does not extract concrete text representation features.

Several weighting/normalization schemes are used by the submitted approaches. TF and TF-IDF are the most popular. Martín dCR et al. prefer log-entropy while López-Anguita et al. apply L2-normalization. Only one approach is based on word embeddings [41]. In addition, some approaches also apply principal components analysis to extract a less sparse representation of reduced dimensionality [6, 13].

Only one approach [14] follows the *profile-based paradigm*, where all available samples of known authorship by a candidate author are treated cumulatively [42]. All the other submitted methods follow the *instance-based paradigm*, where any text of known authorship is represented separately. The relatively small size of candidate author set in the attribution problems as well as the balanced distribution of training texts over the candidate authors have positively affected this preference for instance-based methods.

With respect to the classification method, the majority of submissions used support vector machines (SVM), an algorithm that is both efficient for limited number of classes and able to handle sparse representations of large dimensionality. Gagala explored the use of neural networks while Schaetti focused on a more sophisticated approach based on echo-state network-based reservoir computing, a deep learning algorithm that is easier to be trained in comparison to recurrent neural networks [41]. Halvani and Graner exploit the compression-based cosine similarity measure to estimate the most likely author. Finally, Custódio and Paraboni construct an ensemble of three simple models, each one based on logistic regression.

Each method has its parameters to be tuned, relevant to the type and number of used features, the applied weighting or normalization scheme, or the classifier hyperparameters. There are three basic approaches to do that. Some submitted methods that tune their parameters globally, for all available attribution problems in all languages [6, 14, 13]. Another approach tunes the submitted method for each language separately [25, 4]. Finally, a more detailed approach tunes (at least some) parameters for each attribution problem separately [28, 41]. Certainly, global and language-specific approaches are applied to a larger size of texts and attribution problems and they can extract more reliable statistics. On the other hand, local methods focus on the specific properties of a given attribution problem and they are not confused by irrelevant information.

3.6 Evaluation Results

The results of the 11 submitted approaches and the baseline on the evaluation corpus are presented in Table 3. Beyond macro F1 that is used to rank participants, macro precision, macro recall, and micro accuracy results as well as the total runtime cost are given. As can be seen, the winning submission of Custódio and Paraboni achieves the best scores across all evaluation measures. The ranking of the other approaches according to macro F1 roughly remains the same when other evaluation measures are considered. A notable exception is the approach of Yigal et al. which achieves the 3rd-best micro accuracy score while it is ranked 5th according to macro F1. This indicates an increased potential of that method to recognize majority authors in the test set (i.e., the authors with most unknown texts). For all approaches, macro recall is higher than macro precision. This can be explained by the presence of several candidate authors with very few (or just one) unknown text(s) in most of the attribution problems.

The majority of submissions (6) were able to surpass the baseline and another one was very close to it, according to the macro F1 ranking. The remaining 4 submissions were clearly outperformed by the baseline. Remarkably, simple approaches based on character/word n-grams and well-known classification algorithms [6, 28] are much more effective in this task than more sophisticated methods based on deep learning and linguistic analysis of texts [8, 41]. With respect to the total runtime cost of the submitted approaches, in general, the top-performing methods are also relatively fast. On the contrary, most of the methods that perform significantly lower than the baseline are also the least efficient ones.

Table 4 focuses on the macro F1 scores for all participants and the baseline when the subset of problems in each of the five available languages is examined. The overall

Table 3. Performance of submissions in the cross-domain authorship attribution task using several evaluation measures (ranking is based on macro F1).

Submission	Macro F1	Macro Precision	Macro Recall	Micro Accuracy	Runtime
Custódio and Paraboni	**0.685**	**0.672**	**0.784**	**0.779**	00:04:27
Murauer et al.	0.643	0.646	0.741	0.752	00:19:15
Halvani and Graner	0.629	0.649	0.729	0.715	00:42:50
Mosavat	0.613	0.615	0.725	0.721	00:03:34
Yigal et al.	0.598	0.605	0.701	0.732	00:24:09
Martín dCR et al.	0.588	0.580	0.706	0.707	00:11:01
PAN18-BASELINE	0.584	0.588	0.692	0.719	00:01:18
Miller et al.	0.582	0.590	0.690	0.711	00:30:58
Schaetti	0.387	0.426	0.473	0.502	01:17:57
Gagala	0.267	0.306	0.366	0.361	01:37:56
López-Anguita et al.	0.139	0.149	0.241	0.245	00:38:46
Tabealhoje	0.028	0.025	0.100	0.111	02:19:14

Table 4. Authorship attribution evaluation results (macro F1) per language.

Submission	Overall	English	French	Italian	Polish	Spanish
Custódio and Paraboni	**0.685**	0.744	**0.668**	0.676	0.482	**0.856**
Murauer et al.	0.643	**0.762**	0.607	0.663	0.450	0.734
Halvani and Graner	0.629	0.679	0.536	**0.752**	0.426	0.751
Mosavat	0.613	0.685	0.615	0.601	0.435	0.731
Yigal et al.	0.598	0.672	0.609	0.642	0.431	0.636
Martín dCR et al.	0.588	0.601	0.510	0.571	**0.556**	0.705
PAN18-BASELINE	0.584	0.697	0.585	0.605	0.419	0.615
Miller et al.	0.582	0.573	0.611	0.670	0.421	0.637
Schaetti	0.387	0.538	0.332	0.337	0.388	0.343
Gagala	0.267	0.376	0.215	0.248	0.216	0.280
López-Anguita et al.	0.139	0.190	0.065	0.161	0.128	0.153
Tabealhoje	0.028	0.037	0.048	0.014	0.024	0.018

top-performing submission by Custódio and Paraboni was also the most effective one for French and especially Spanish (with a remarkable difference from the second-best approach). Moreover, the method of Halvani and Graner achieved quite remarkable results for Italian in comparison to the rest of submissions. The most difficult cases appear to be the Polish ones while the highest average results are obtained for English and Spanish.

Table 5 shows the performance (macro-averaged F1 score) of the submitted methods for a varying candidate set size (from 20 authors to 5 authors). For instance, when 20 authors are considered, all 5 attribution problems with that candidate set size in all languages are examined. Apparently, the overall top-performing method of Custódio and Paraboni remains the most effective one for each of the examined candidate set sizes. In most cases, the ranking of participants is very similar to their overall ranking. It's also remarkable that the PAN18-BASELINE is especially effective when there are

Table 5. Performance (macro F1) of the cross-domain authorship attribution submissions per candidate set size.

Submission	20 Authors	15 Authors	10 Authors	5 Authors
Custódio and Paraboni	**0.648**	**0.676**	**0.739**	**0.677**
Murauer et al.	0.609	0.642	0.680	0.642
Halvani and Graner	0.609	0.605	0.665	0.636
Mosavat	0.569	0.575	0.653	0.656
Yigal et al.	0.570	0.566	0.649	0.607
Martín dCR et al.	0.556	0.556	0.660	0.582
PAN18-BASELINE	0.546	0.532	0.595	0.663
Miller et al.	0.556	0.550	0.671	0.552
Schaetti	0.282	0.352	0.378	0.538
Gagala	0.204	0.240	0.285	0.339
López-Anguita et al.	0.064	0.065	0.195	0.233
Tabealhoje	0.012	0.015	0.030	0.056

only a few (5) authors. In general, the performance of submissions improves when the candidate set becomes smaller. However, it seems that the best-performing approaches are less accurate in problems with 5 candidate authors in comparison to problems with 10 authors.

As in previous PAN shared tasks, we have applied statistical significance testing to the attributions provided by the submitted approaches to assess to which extent the differences between their outputs are statistically meaningful. In authorship attribution, the distribution of class labels is often heavily skewed, simply unknown, or hard to estimate. This is why we resort to a non-parametric test known as *approximate randomization testing* [30], which does not make any far-reaching assumptions about any underlying distributions. In Table 6 we present pairwise tests for all submitted approaches, where the predictions for all problems have been analyzed in terms of their respective F1-scores. The probabilities returned by the test (for 1,000 bootstrapped iterations) can be interpreted as the conventional p-values of one-sided, statistical tests: they indicate the probability of failing to reject the null hypothesis (H_0) that the classifiers do *not* output significantly different scores. We use a symbolic notation corresponding to the following thresholds: '=' (not significantly different: $p > 0.5$), '*' (significantly different: $p < 0.05$), '**' (very significantly different: $p < 0.01$), '***' (highly significantly different: $p < 0.001$). As can be seen from the table, the difference between submissions of a neighboring rank are typically less statistically meaningful, although it catches the eye that this is not true for the winner and its immediate runner-up in this edition ($p = 0.183$). Note however that the winner does realize a significant statistical difference with respect to all other participants, which is reassuring. Especially, the differences between the submissions which performed above the baseline seem less meaningful, adding to the relativity of the final rankings of these systems. Participants scoring below the baseline generally reach higher significance scores in comparison to those scoring above the baseline threshold, which attests to the competitiveness of the baseline in this edition.

Table 6. Significance of pairwise differences in output between submissions, across all problems.

	Murauer et al.	Halvani and Graner	Mosavat	Yigal et al.	Martín dCR et al.	Miller et al.	PAN18-BASELINE	Schaetti	Gagala	López-Anguita et al.	Tabealhoje
Custódio and Paraboni	=	***	***	***	***	***	***	***	***	***	***
Murauer et al.		**	***	**	***	***	***	***	***	***	***
Halvani and Graner			=	=	=	=	=	***	***	***	***
Mosavat				=	=	=	=	***	***	***	***
Yigal et al.					=	=	=	***	***	***	***
Martín dCR et al.						=	=	***	***	***	***
Miller et al.							=	***	***	***	***
PAN18-BASELINE								***	***	***	***
Schaetti									***	***	***
Gagala										***	***
López-Anguita et al.											***

3.7 Additional Analyses

We have conducted some additional analyses about the cross-fandom nature of the authorship attribution task. All attribution problems involved an attribution task where the unseen texts belonged to a target fandom ('Harry Potter - J.K. Rowling') that was not represented in the training data, which only contained so-called non-target or training fandoms. How did the variety in training fandoms for a particular problem affect the ultimate attribution performance for that author? In Figure 1 (left) we plot the F1-score for each author individually, as a function of the number of distinct training fandoms which were available in the training material for that author. (Note that the total number of training texts per author was always kept stable inside a single problem.) The scores were calculated across all participant submissions, including the baseline. The application of a simple Pearson test to these results allows us to observe a mild ($r = 0.2$), yet statistically significant ($p = 0.001$) positive correlation: if more distinct training fandoms were available for an author, an author's F1-score for the test material benefited from this. This results are firmly in line with those of Sapkota et al., who, in the context of cross-topic authorship attribution, also 'demonstrated that training on diverse topics is better than training on a single topic' [38].

Finally, we explored the effect of the target fandom, i.e. the Harry Potter novel series by J.K. Rowling. Assuming that some of the authors of the target fandom could have been actively imitating Rowling's writing style, we hypothesized that this might have had an effect on the attribution results: i.e., if an author stayed very close to the style of the canon's author, this might have made it more difficult to identify the fan. Interestingly, the results reported in Figure 1 (right) suggest that this was, perhaps somewhat surprisingly, *not* the case at all. For this analysis, we extracted a list of all 7 original Harry Potter chapters in the canon (original UK version published by Bloomsbury).

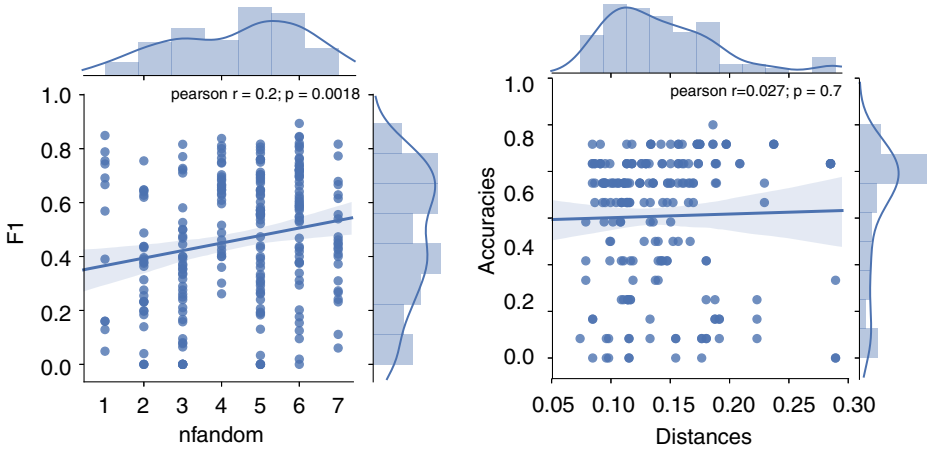


Figure 1. Left: Effect of the number of distinct training fandoms available for an author on the F1-score for that author in the test material. Right: Effect of the stylistic similarity between a test fanfic and the target fandom's original author, J.K. Rowling, and the proportion of correct attributions for this fanfic. Interestingly, no trend can be discerned.

We converted these into an L1-normalized bag-of-words model, capturing the relative frequencies of all character trigrams that appeared in at least 5 chapters. We represent Rowling's style in this analysis as the centroid for the resulting bag-of-words model (column-wise mean). Next, we converted all test fanfics using the same vectorizer and calculated the cosine distance between each test fanfic and Rowling's centroid. Next, we correlated this distance with the number of correct predictions for a test fanfic (across all participants), using Pearson's r. Interestingly, and in spite of the diversity in distances, no trend whatsoever is evident from this analysis ($r = 0.027$; $p = 0.7$). Interestingly, whether or not a fanfic stayed close to Rowling in writing style, thus did not have a clear effect on difficulty of the attribution task.

4 Style Change Detection

The simple, yet challenging question to answer for the style change detection task is as follows: Given a document, is it written by a single author or by multiple authors? To be able to provide an answer, the document has to be intrinsically analyzed, i.e., changes of authorship have to be determined by capturing changes of writing styles. As it is irrelevant at this point to identify the exact change positions, the problem can also be tackled by applying a binary classification over the whole document. Therefore it is also possible to quantify the style of a whole document and to learn feature combinations which separate single-author documents from multi-author texts.

In this section, we present an overview of the style change detection task at PAN 2018. First, we describe the construction of the task's evaluation data set in detail, followed by an overview of its evaluation framework. Then we survey the submitted approaches and report on their evaluation.

4.1 Data Set Construction

Three distinct data sets for training (50%), validation (25%) and testing (25%) have been constructed, where the ground truth for the first two was provided up front to participants. All data sets are based on user posts from 15 heterogeneous sites of the Q&A network StackExchange.[4]

Crawling The basis for the data sets has been crawled from the network[5] as follows:

1. Choosing a site (e.g., *programming*, *politics*, *sports* or *religion*). In the following a site is referred to as *topic*.
2. Retrieving of the 20 most popular tags of the site (e.g., for *politics* the tags *law*, *economy* or *european union* are among the most popular ones). In the following a tag is referred to as *subtopic*.
3. For each subtopic, retrieving of the 30 authors which posted the most questions, as well as the 30 authors who provided the most answers with respect to this specific subtopic.
4. For each author and each subtopic, retrieving of all questions and answers.

Preprocessing Before the final data sets were compiled, all texts have been preprocessed and filtered. The following filtering steps have been applied for both answers and questions:

- removal of very short texts (e.g., only a few words or symbols)
- removal of texts that have been edited by users other than the original author[6]
- removal of external URLs
- removal of embedded images
- removal of code snippets (which is especially needed for the *StackOverflow* site)
- removal of bullet lists
- removal of block quotes
- removal of texts containing Arabic characters (especially needed for the *Islam* site)

After applying these steps, most texts contain sufficiently long and well-formed sentences about a specific subtopic, without any HTML tags or other undesired content. In case the cleaning process shortened a document to less than three sentences, it was removed before creating the actual data set.

Compilation Using the cleaned questions and answers of users belonging to the same topic and subtopic, the final documents have been assembled by varying the parameters listed in Table 7. For single-author documents, one or more texts of the same author have been used to create problems containing 300 to 1000 tokens. The compilation of

[4] https://stackexchange.com

[5] using the StackExchange API, https://api.stackexchange.com, visited June 2018

[6] StackExchange allows questions and answers to be edited easily by any registered member, who may, e.g., correct spelling errors or reformulate texts to be more precise

Table 7. Parameters for constructing the style change detection data set.

Parameter	Value/s
number of style changes	0–3
number of collaborating authors	1–3
document length	300–1000 tokens
change positions	at the end of / within paragraphs, mixed
segment length distribution	equalized / randomly
two-authors distributions	(A1-A2), (A1-A2-A1), (A1-A2-A1-A2)
three-authors distributions	(A1-A2-A3), (A1-A2-A1-A3), (A1-A2-A3-A1) (A2-A1-A3-A1)

multi-author documents has been conducted by combining texts of multiple authors, where the number of authors, changes, segment lengths and author distributions have been varied. This way, 2980 training problems, 1492 validation problems, and 1352 test problems have been created, where for each data set the amount of documents containing style changes is equal to the number of documents containing no changes. A detailed view of the data set's statistics with respect to the parameters is shown in Table 8. Concerning the topics, the number of problems are depicted in Table 9. For each topic and subtopic, single- and multi-author problems are also equally represented. A complete list of subtopics appearing in each topic is shown in Table 12 in the Appendix.

4.2 Evaluation Framework

The participants designed and optimized their approaches with the given, publicly available training and validation data sets described above. Performance could either be measured locally using the provided evaluation script, or by deploying the respective software to TIRA [11, 34] and running it against the respective data set. The test data set was not publicly available, so that the latter option was necessary in this case, i.e., participants submitted their final software and ran it against the test data, without seeing performance results. This way, no information other than that provided by the data set itself was available to participants. Participants were allowed to submit an unlimited number of runs on the test data, but were asked to select one specific run that to be be used for the final ranking and for all results presented in Section 4.4. To evaluate the performances of the approaches, their accuracy was measured, i.e., the portion of correctly predicted style change detection problems compared to the total number of problems. Three baselines were used for comparison:

1. *rnd1-BASELINE*: A guessing baseline that achieves 50% by default due to the balanced distribution of style changing and non-changing documents in the data set.
2. *rnd2-BASELINE*: An enhanced guessing baseline that exploits the data set's statistics document length and number of style changes.
3. *C99-BASELINE*: This baseline employs a commonly used text segmentation algorithm, namely the C99 algorithm proposed by Choi et al. [5], since it is one of the few text segmentation algorithms capable of predicting the number of segments. We utilized this feature by predicting a style change if the algorithm found more than one segment, and no change otherwise.

Table 8. Key figures of the style change detection data set regarding its construction parameters.

Key figures		**Training**	**Validation**	**Test**
Number of documents:		2980 (100%)	1492 (100%)	1352 (100%)
Number of authors	1	1490 (50%)	746 (50%)	676 (50%)
	2	872 (29%)	452 (30%)	384 (28%)
	3	618 (21%)	294 (20%)	292 (22%)
Document length (tokens)	300-500	476 (16%)	233 (16%)	203 (15%)
	500-750	1012 (34%)	528 (35%)	450 (33%)
	750-1000	1126 (38%)	555 (37%)	531 (39%)
	>1000	366 (12%)	176 (12%)	168 (12%)
Multi-authored-documents				
Author distribution	A1-A2	360 (24%)	208 (28%)	149 (22%)
	A1-A2-A1	299 (20%)	155 (21%)	131 (19%)
	A1-A2-A1-A2	213 (14%)	89 (12%)	104 (15%)
	A1-A2-A1-A3	30 (2%)	23 (3%)	18 (3%)
	A1-A2-A3	525 (35%)	244 (33%)	240 (36%)
	A1-A2-A3-A1	31 (2%)	14 (2%)	22 (3%)
	A2-A1-A3-A1	32 (2%)	13 (2%)	12 (2%)
Average segment length (tokens)	< 200	149 (10%)	66 (9%)	71 (11%)
	200-300	806 (54%)	410 (55%)	385 (57%)
	300-400	408 (27%)	199 (27%)	163 (24%)
	> 400	127 (9%)	71 (10%)	57 (8%)
Change positions	end of paragraph	526 (35%)	258 (35%)	255 (38%)
	within paragraphs	494 (33%)	229 (31%)	212 (31%)
	mixed	470 (32%)	259 (35%)	209 (31%)

4.3 Survey of Submissions

This year, 6 teams registered for the style change detection task, five of whom submitted their software to TIRA [11, 34]. In what follows, a short summary of each approach is given:

- *Hosseinia and Mukherjee* [18]: The main idea of this approach is to solely rely on the grammatical structure used by authors in order to detect style changes, i.e., no other lexical features like character or word n-grams are used. To compute corresponding features, first, the parse tree of each sentence of a given document is computed, which is further traversed and linearized. By doing so, the whole document is represented as a consecutive order of parse tree features, which are then fed into a recurrent neural network (RNN) based on the author's previous work on authorship verification [17]. In parallel, a second RNN is constructed of which the input is the parse tree feature representation of the reversed order of sentences of the document. Finally, multiple similarity metrics are computed to estimate the difference between the original and reversed order network representations, where a final softmax layer yields the style change prediction.

Table 9. Overview of the style change detection data set with respect to topics.

Site	Training				Validation				Test			
	Problems	Authors			Problems	Authors			Problems	Authors		
		1	2	3		1	2	3		1	2	3
bicycles	160	80	47	33	82	41	28	13	70	35	27	8
christianity	358	179	107	72	176	88	48	40	172	86	45	41
gaming	178	89	47	42	86	43	23	20	78	39	21	18
history	354	177	104	73	178	89	54	35	170	85	46	39
islam	166	83	49	34	86	43	31	12	72	36	20	16
linguistics	144	72	46	26	72	36	22	14	64	32	12	20
meta	196	98	56	42	94	47	30	17	90	45	30	15
parenting	178	89	54	35	92	46	32	14	78	39	27	12
philosophy	468	234	146	88	232	116	63	53	224	112	65	47
poker	100	50	35	15	48	24	14	10	42	21	13	8
politics	204	102	57	45	102	51	34	17	90	45	22	23
project man.	104	52	24	28	50	25	12	13	44	22	14	8
sports	102	51	34	17	54	27	20	7	40	20	12	8
stackoverflow	112	56	23	33	60	30	16	14	48	24	12	12
writers	156	78	43	35	80	40	25	15	70	35	18	17
$\sum$	**2980**	**1490**	**872**	**618**	**1492**	**746**	**452**	**294**	**1352**	**676**	**384**	**292**

- *Khan* [23]: Here, an algorithmic approach is utilized that operates on the sentence-level. First, the document is split into sentences, for which groups of predefined sizes are formed. By using sliding windows, two consecutive sentence windows are then compared to each other, where exactly one sentence in the middle is shared among both groups. The comparison is based on a similarity function which operates on cooccurrences of word features. More specifically, stop words, most/least frequent words or word pairs, and punctuation frequencies are utilized.
- *Safin and Ogaltsov* [36]: This approach utilizes an ensemble of three individual classifiers, each operating on different kinds of features. First, a random forest classifier is trained using 19 statistical text features including number of sentences, text length, and frequencies of unique words, punctuations, or letters. Second, a classifier is built from a 3,000-dimensional vector containing frequencies of character n-grams. Finally, a logistic regression classifier is trained from the frequencies of all word {1-6}-grams, resulting in a high-dimensional vector with over 3 million dimensions. Using optimized coefficients, a weighted linear combination of the three classifiers is formed, where a predefined threshold determines the final result.
- *Schaetti* [40]: In this approach, a character-based convolutional neural network (CNN) is designed. Each document is represented as a fixed-sized vector of 12,000 consecutive characters which are fed into the network, i.e., into an embedding layer that reduces the dimension to 50 and captures context similarities of occurring characters in a multi-dimensional space. Subsequently, the second layer is composed of three different convolutional layers with 25 filters each to capture the most expressive patterns of 2-4 consecutive character 2-grams. After utilizing a max-pooling layer for each of the convolutional layers, a binary linear layer is finally used to predict the existence of style changes.

Table 10. Evaluation results of the style change detection task.

Submission	Accuracy	Runtime
Zlatkova et al.	**0.893**	01:35:25
Hosseinia and Mukherjee	0.825	10:12:28
Safin and Ogaltsov	0.803	00:05:15
Khan	0.643	00:01:10
Schaetti	0.621	00:03:36
C99-BASELINE	0.589	00:00:16
rnd2-BASELINE	0.560	–
rnd1-BASELINE	0.500	–

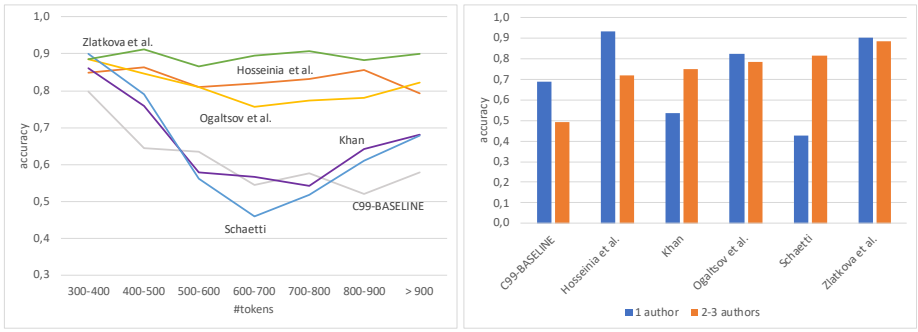
– *Zlatkova et al.* [52]: At a glance, the authors rely on a rather sophisticated, hierarchical ensemble architecture (stacking) to solve the style change detection problem. Prior to building the ensemble, the texts are preprocessed by replacing URL's, file paths and very long words with special tokens, and also by splitting long hyphenated words. Moreover, each document is segmented into three fragments of equal lengths, and also sliding windows are utilized to increase the quantity of the assessed features. Using several distinct feature groups including lexical, syntactical, and other features, four different classifiers (e.g., SVM and random forest) are trained for each group, where a weighted model is subsequently computed for each feature group. These weighted models, in combination with a TF-IDF-based gradient boosting model (using LightGBM [20]) form the input for a logistic regression meta-classifier, which produces the final output.

4.4 Evaluation Results

The overall performance results are depicted in Table 10. With an accuracy of nearly 90%, Zlatkova et al. achieved the best result over all documents across all topics and subtopics. All approaches outperformed all baselines. With respect to runtime, the two best-performing approaches also needed significantly more time (due to the ensemble technique and parse tree generation, respectively), compared to the other participants who produced predictions within minutes for the roughly 1,300 documents in the test data set.

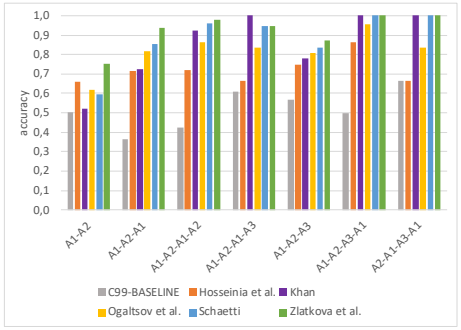
Detailed results with respect to various parameters of the test data set are illustrated in Figure 2. Each sub-figure only show the C99-BASELINE, as it is the best-performing baseline and also the only one which is able to produce differentiated results for the individual text characteristics. As can be seen in Figure 2a, the three best performing approaches are stable to the amount of tokens contained in the problems, and only two approaches are sensitive to it. Concerning single-author and multi-author documents, the results are quite heterogeneous as depicted in Figure 2b. Zlatkova et al. and Ogaltsov et al. achieve similar results for both problem classes, whereas the other approaches as well as the baseline favor either of the two.

Figure 2c shows results for multi-author problems according to the type of author distribution. Interestingly, the simplest compilation A1-A2, i.e., two authors with one style change in between, is the most challenging type. In general, the performance seems to be related to the number of authors involved, i.e., the more authors, the better

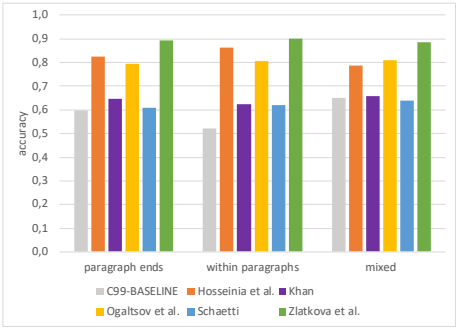


(a) Number of tokens

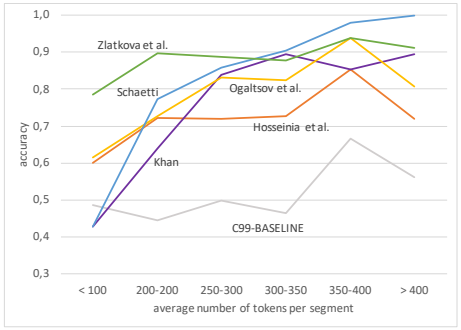
(b) Number of authors



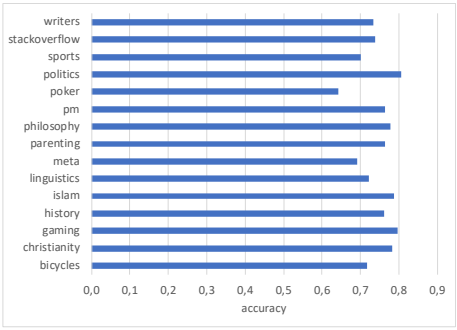
(c) Author distributions



(d) Style change positions



(e) Average number of tokens per segment



(f) Average accuracy per topic

Figure 2. Detailed evaluation results of the style change detection approaches with respect to various parameters.

the results. All submitted approaches are insensitive to the style change position as can be seen from Figure 2d. Thus, it is irrelevant for the respective algorithms if authors switch only at the end of paragraphs or anywhere else. Compared to the results of the previous year's style breach detection task [50] this can be seen as an enhancement, as

Table 11. Evaluation results regarding selected subtopics containing eight or more documents.

	Subtopic (topic)	Docs.	C99-BASELINE	Hosseinia et al.	Khan	Ogaltsov et al.	Schaetti	Zlatkova et al.	Avg.
Best 10 subtopics	starcraft-2 (gaming)	12	0.67	0.83	0.92	**1.00**	**1.00**	0.92	0.93
	political-history (history)	14	0.43	**0.93**	0.86	**0.93**	**0.93**	**0.93**	0.91
	history (christianity)	10	0.80	0.90	0.80	**1.00**	0.80	0.90	0.88
	halal-haram (islam)	10	0.80	**1.00**	0.70	**1.00**	0.70	**1.00**	0.88
	war (history)	8	0.38	0.75	**1.00**	0.88	0.75	**1.00**	0.88
	economy (politics)	8	0.25	**1.00**	0.75	0.88	0.75	**1.00**	0.88
	exegesis (christianity)	22	0.73	0.91	0.82	0.91	0.73	**0.95**	0.86
	prophet-muhammad (islam)	8	0.63	**1.00**	0.88	0.88	0.50	**1.00**	0.85
	syntax (linguistics)	14	0.79	0.86	0.79	0.86	0.71	**1.00**	0.84
	election (politics)	10	0.60	0.80	**0.90**	**0.90**	0.70	**0.90**	0.84
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
Worst 10 subtopics	feature-request (meta)	20	0.45	0.75	0.60	0.75	0.60	**0.80**	0.70
	discipline (parenting)	10	0.30	**0.90**	0.70	0.50	0.60	0.80	0.70
	scrum (pm)	14	**0.93**	**0.93**	0.43	0.79	0.57	0.79	0.70
	ancient-rome (history)	12	0.42	0.83	0.67	0.75	0.25	**0.92**	0.68
	lds (christianity)	14	0.43	0.71	0.57	**0.79**	0.50	0.71	0.66
	fiction (writers)	14	0.36	0.86	0.36	**0.93**	0.29	0.86	0.66
	nature-of-god (christianity)	12	0.75	0.75	0.50	0.67	0.42	**0.83**	0.63
	english (linguistics)	8	0.50	**0.75**	0.63	0.50	0.63	0.50	0.60
	world-war-two (history)	26	0.62	0.73	0.42	0.54	0.46	**0.81**	0.59
	support (meta)	10	0.40	**0.70**	0.50	0.60	0.40	**0.70**	0.58

those approaches showed clearly better results when authors changed only at paragraph ends.

With respect to the average number of tokens per author segment in multi-author documents, Figure 2d shows a clear tendency towards longer segments. That is, the more each author contributes to a document on average, the better the results get. Finally, Figure 2f shows the average accuracy of all submitted approaches (excluding the baseline) for each topic of the test data set. The results are quite homogeneous, yielding the best performance on average for *politics* and the worst for *poker*.

As a final performance analysis, Table 11 shows the accuracies achieved with respect to specific topics and subtopics. Here, the individual results for the on average 10 best-performing subtopics (upper part) as well as the 10 most problematic subtopics (lower part) are shown, where only subtopics containing at least eight documents have been considered. It can be seen that subtopics from various topics are represented, and that approaches achieve perfect accuracy for multiple subtopics. Remarkably, Zlatkova et al. reached the best performance for eight of the top subtopics, predicting 100% of the problems correctly for five of those subtopics. Moreover, for most of the worst-performing subtopics, at least one of the approaches achieved a good accuracy.

5 Summary

Cross-domain authorship attribution studies a challenging, yet realistic scenario where the training and test texts belong to distinct domains. Fanfiction provides excellent material for this task since it enables significant control over the topic of texts. The number of received submissions for this task indicates there is a relatively large research community working on this field. In general, submissions that do not require a deep linguistic analysis of texts were found to be both the most effective and the most efficient ones. Heterogeneous ensembles of simple classifiers and compression models outperformed more sophisticated approaches based on deep learning. Furthermore, the candidate set size is inversely correlated with the attribution accuracy especially when more than 10 authors are considered, while the number of training fandoms positively affects the recognition accuracy of a candidate author.

With the relaxation of the style change detection task, we attracted not only more participants than before, but also rendered the task more tractable for them, as indicated by the the better performance scores achieved. On a novel data set created from a popular Q&A network containing more than 4,000 problems, all participants managed to outperform the three baselines. To predict style changes, a rich set of features has been employed and exploited using various techniques ranging from machine learning ensembles to deep learning. Accuracies of up to nearly 90% over the whole data set and several individual results of 100% for specific topics indicate that the problem can be solved with a high precision. Consequently, the results represent a good starting point to further pursue the style change detection task in future PAN editions.

Bibliography

[1] Argamon, S., Juola, P.: Overview of the international authorship identification competition at PAN-2011. In: CLEF 2011 Labs and Workshop, Notebook Papers, 19-22 September 2011, Amsterdam, The Netherlands (2011)

[2] Argamon, S., Koppel, M., Pennebaker, J.W., Schler, J.: Automatically profiling the author of an anonymous text. Communications of the ACM 52(2), 119–123 (2009)

[3] Bogdanova, D., Lazaridou, A.: Cross-language authorship attribution. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014. pp. 2015–2020 (2014)

[4] Martín del Campo-Rodríguez, C., Gómez-Adorno, H., Sidorov, G., Batyrshin, I.: CIC-GIL Approach to Cross-domain Authorship Attribution. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)

[5] Choi, F.Y.: Advances in Domain Independent Linear Text Segmentation. In: Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. pp. 26–33. Association for Computational Linguistics (2000)

[6] Custódio, J.E., Paraboni, I.: EACH-USP Ensemble cross-domain authorship attribution. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)

[7] Daniel Karaś, M.S., Sobecki, P.: OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)

[8] Gagala, L.: Authorship attribution with neural networks and multiple features. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[9] Giannella, C.: An improved algorithm for unsupervised decomposition of a multi-author document. Technical Papers, The MITRE Corporation (February 2014)
[10] Glover, A., Hirst, G.: Detecting stylistic inconsistencies in collaborative writing. In: The New Writing Environment, pp. 147–168. Springer (1996)
[11] Gollub, T., Stein, B., Burrows, S.: Ousting Ivory Tower Research: Towards a Web Framework for Providing Experiments as a Service. In: Hersh, B., Callan, J., Maarek, Y., Sanderson, M. (eds.) 35th International ACM Conference on Research and Development in Information Retrieval (SIGIR 12). pp. 1125–1126. ACM (Aug 2012)
[12] Graham, N., Hirst, G., Marthi, B.: Segmenting documents by stylistic character. Natural Language Engineering 11(04), 397–415 (2005)
[13] HaCohen-Kerner, Y., Miller, D., Yigal, Y., Shayovitz, E.: Cross-domain Authorship Attribution: Author Identification using char sequences, word unigrams, and POS-tags features). In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[14] Halvani, O., Graner, L.: Cross-Domain Authorship Attribution Based on Compression Models. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[15] Hellekson, K., Busse, K. (eds.): The Fan Fiction Studies Reader. University of Iowa Press (2014)
[16] Holmes, D.I.: The evolution of stylometry in humanities scholarship. Literary and Linguistic Computing 13(3), 111–117 (1998)
[17] Hosseinia, M., Mukherjee, A.: Experiments with neural networks for small and large scale authorship verification. arXiv preprint arXiv:1803.06456 (2018)
[18] Hosseinia, M., Mukherjee, A.: Parallel Attention Recurrent Neural Network for Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[19] 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)
[20] Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems. pp. 3149–3157 (2017)
[21] Kestemont, M., Luyckx, K., Daelemans, W., Crombez, T.: Cross-genre authorship verification using unmasking. English Studies 93(3), 340–356 (2012)
[22] Khan, J.A.: Style breach detection: An unsupervised detection model. In: Working Notes Papers of the CLEF 2017 Evaluation Labs (2017)
[23] Khan, J.A.: A Model for Style Breach Detection at a Glance. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[24] Koppel, M., Schler, J., Bonchek-Dokow, E.: Measuring differentiability: Unmasking pseudonymous authors. Journal of Machine Learning Research 8, 1261–1276 (2007)
[25] López-Anguita, R., Montejo-Ráez, A., Díaz-Galiano, M.C.: Complexity measures and POS n-grams for author identification in several languages: SINAI at PAN@CLEF 2018.

In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[26] Luyckx, K., Daelemans, W.: Authorship attribution and verification with many authors and limited data. In: Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. pp. 513–520. Association for Computational Linguistics (2008)
[27] Mikros, G., Argiri, E.: Investigating Topic Influence in Authorship Attribution. In: Stein, B., Koppel, M., Stamatatos, E. (eds.) SIGIR 07 Workshop Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07). CEUR-WS.org (Jul 2007), http://ceur-ws.org/Vol-276
[28] Murauer, B., Tschuggnall, M., Specht, G.: Dynamic Parameter Search for Cross-Domain Authorship Attribution. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[29] Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., Woodard, D.: Surveying stylometry techniques and applications. ACM Computing Surveys 50(6), 86:1–86:36 (2017)
[30] Noreen, E.W.: Computer-Intensive Methods for Testing Hypotheses: An Introduction. John Wiley and Sons (1989)
[31] Overdorf, R., Greenstadt, R.: Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution. Proceedings on Privacy Enhancing Technologies 2016(3), 155–171 (2016)
[32] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
[33] Potthast, M., Eiselt, A., Barrón-Cedeño, A., Stein, B., Rosso, P.: Overview of the 3rd International Competition on Plagiarism Detection. In: Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN). Amsterdam, The Netherlands (September 2011)
[34] 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)
[35] Safin, K., Kuznetsova, R.: Style Breach Detection with Neural Sentence Embeddings. In: Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2017)
[36] Safin, K., Ogaltsov, A.: Detecting a change of style using text statistics. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[37] Sapkota, U., Bethard, S., Montes, M., Solorio, T.: Not all character n-grams are created equal: A study in authorship attribution. In: Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 93–102 (2015)
[38] Sapkota, U., Solorio, T., Montes, M., Bethard, S., Rosso, P.: Cross-topic authorship attribution: Will out-of-topic data help? In: Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers. pp. 1228–1237 (2014)

[39] Sapkota, U., Solorio, T., Montes-y-Gómez, M., Bethard, S.: Domain adaptation for authorship attribution: Improved structural correspondence learning. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (2016)
[40] Schaetti, N.: UniNE at CLEF 2018: Character-based Convolutional Neural Network for Style Change Detection. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[41] Schaetti, N.: UniNE at CLEF 2018: Echo State Network-based Reservoir Computing for Cross-domain Authorship Attribution. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)
[42] Stamatatos, E.: A Survey of Modern Authorship Attribution Methods. Journal of the American Society for Information Science and Technology 60, 538–556 (2009)
[43] Stamatatos, E.: Intrinsic Plagiarism Detection Using Character n-gram Profiles. In: Notebook Papers of the 5th Evaluation Lab on Uncovering Plagiarism, Authorship and Social Software Misuse (PAN). Amsterdam, The Netherlands (September 2011)
[44] Stamatatos, E.: Plagiarism detection using stopword n-grams. Journal of the Association for Information Science and Technology 62(12), 2512–2527 (2011)
[45] Stamatatos, E.: On the robustness of authorship attribution based on character n-gram features. Journal of Law and Policy 21, 421–439 (2013)
[46] Stamatatos, E.: Authorship attribution using text distortion. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers. pp. 1138–1149. Association for Computational Linguistics (2017)
[47] Stamatatos, E.: Masking topic-related information to enhance authorship attribution. JASIST 69(3), 461–473 (2018)
[48] Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Clustering by Authorship Within and Across Documents. In: Working Notes Papers of the CLEF 2016 Evaluation Labs. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (Sep 2016), http://ceur-ws.org/Vol-1609/
[49] Tschuggnall, M., Specht, G.: Countering Plagiarism by Exposing Irregularities in Authors' Grammar. In: Proceedings of the European Intelligence and Security Informatics Conference (EISIC). pp. 15–22. IEEE, Uppsala, Sweden (August 2013)
[50] Tschuggnall, M., Stamatatos, E., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., Potthast, M.: Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering. In: Cappellato, L., Ferro, N., Goeuriot, L., Mandl, T. (eds.) Working Notes Papers of the CLEF 2017 Evaluation Labs. CEUR Workshop Proceedings, vol. 1866. CLEF and CEUR-WS.org (Sep 2017), http://ceur-ws.org/Vol-1866/
[51] Zheng, R., Li, J., Chen, H., Huang, Z.: A framework for authorship identification of online messages: Writing-style features and classification techniques. Journal of the American Society for Information Science and Technology 57(3), 378–393 (2006)
[52] 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. In: Cappellato, L., Ferro, N., Nie, J.Y., Soulier, L. (eds.) Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CLEF and CEUR-WS.org (2018)

Appendix

Table 12. Topics and subtopics found in the style change detection data set.

Topic	Subtopics
bicycles	brakes, chain, frames, maintenance, mountain-bike, repair, road-bike, shimano, tire, wheels
christianity	bible, biblical-basis, catholicism, church-history, exegesis, history, jesus, lds, nature-of-god, soteriology
gaming	diablo-3, league-of-legends, minecraft, minecraft-commands, pc, pokemon-go, skyrim, starcraft-2, steam, technical-issues
history	20th-century, ancient-history, ancient-rome, europe, middle-ages, military, political-history, united-states, war, world-war-two
islam	fiqh, hadith, halal-haram, nikah, practical-islam, prophet-muhammad, quran, salat, sharia, tafseer
linguistics	computational-linguistics, english, etymology, historical-linguistics, morphology, phonetics, phonology, semantics, syntax, terminology
meta	bug, comments, discussion, feature-request, reputation, review, stackoverflow, status-completed, support, tags
parenting	behavior, development, discipline, infant, newborn, pre-schooler, primary-schooler, sleep, teen, toddler
philosophy	epistemology, ethics, history-of-philosophy, kant, logic, metaphysics, philosophy-of-mathematics, philosophy-of-mind, philosophy-of-science, reference-request
proj. man.	agile, communication, estimating, kanban, ms-project, planning, pm-software, scrum, software-development, team-management
poker	betting-strategy, cash-game, nlhe, odds, online, poker-strategy, poker-theory, rules, texas-hold-em, tournament
politics	congress, constitution, donald-trump, economy, election, european-union, law, president, united-kingdom, united-states
sports	american-football, baseball, cricket, equipment, football, officiating, rules, statistics, tennis, trivia
stackoverflow	android, c++, html, ios, java, javascript, jquery, php, python
writers	character-development, characters, creative-writing, fiction, novel, plot, publishing, style, technical-writing, technique