Overview of the Cross-Domain Authorship Attribution Task at PAN 2019

Mike Kestemont,[1] Efstathios Stamatatos,[2] Enrique Manjavacas,[1]
Walter Daelemans,[1] Martin Potthast,[3] and Benno Stein[4]

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

pan@webis.de https://pan.webis.de

Abstract. Authorship identification remains a highly topical research problem in computational text analysis, with many relevant applications in contemporary society and industry. In this edition of PAN, we focus on authorship attribution, where the task is to attribute an unknown text to a previously seen candidate author. Like in the previous edition we continue to work with fanfiction texts (in four Indo-European languages), written by non-professional authors in a cross-domain setting: the unknown texts belong to a different domain than the training material that is available for the candidate authors. An important novelty of this year's setup is the focus on open-set attribution, meaning that the test texts contain writing samples by previously unseen authors. For these, systems must consequently refrain from an attribution. We received altogether 12 submissions for this task, which we critically assess in this paper. We provide a detailed comparison of these approaches, including three generic baselines.

1 Cross-Domain Authorship Attribution

Authorship attribution [1,2,3] continues to be an important problem in information retrieval and computational linguistics, and also in applied areas such as law and journalism where knowing the author of a document (such as a ransom note) may enable law enforcement to save lives. The most common framework for testing candidate algorithms is the closed-set attribution task: given a sample of reference documents from a finite set of candidate authors, the task is to determine the most likely author of a previously unseen document of unknown authorship. This task is quite challenging under cross-domain conditions where documents of known and unknown authorship come from different domains (such as a different thematic area or genre). In addition, it is often more realistic to assume that the true author of a disputed document is not necessarily included in the list of candidates [4].

 CLEF 2019, 9-12 September 2019, Lugano, Switzerland.

This year, we again focus on the attribution task in the context of transformative literature, more colloquially know as 'fanfiction'. Fanfiction refers to a rapidly expanding body of fictional narratives, typically produced by non-professional authors who self-identify as 'fans' of a particular oeuvre or individual work [5]. Usually, these stories (or 'fics') are openly shared online with a larger fan community on platforms such as `fanfiction.net` or `archiveofourown.org`. Interestingly, fanfiction is the fastest growing form of writing in the world nowadays [6]. When sharing their texts, fanfiction writers explicitly acknowledge taking inspiration from one or more cultural domains that are known as 'fandoms'. The resulting borrowings take place on various levels, such as themes, settings, characters, story world, and also style. Fanfiction usually is of an unofficial and unauthorized nature [7], but because most fanfiction writers do not have any commercial purposes, the genre falls under the principle of 'Fair Use' in many countries [8]. From the perspective of writing style, fanfiction offers valuable benchmark data: the writings are unmediated and unedited before publication, which means that they should accurately reflect an individual author's writing style. Moreover, the rich metadata available for individual fics presents opportunities to quantify the extent to which fanfiction writers have modeled their writing style after the original author's style [9].

In the previous edition of PAN we also dealt with authorship attribution in fanfiction and added extra difficulty with a cross-domain condition (i.e., different fandoms). This year we have further increased the difficulty of the task by focusing on *open-set* attribution conditions, meaning that the true author of a test text is not necessarily included in the list of candidate authors. More formally, an open cross-domain authorship attribution problem can be expressed as a tuple (A, K, U), with A as the set of candidate authors, K as the set of reference (known authorship) texts, and U as 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 author a. Each text in U should either be assigned to exactly one $a \in A$, or the system should refrain from an attribution if the target author is not supposed to be 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 names of) the fandoms of the texts in K. Then, all texts in U belong to a single fandom d_U, with $d_U \notin D_K$.

2 Datasets

This year's shared tasks use datasets in four major Indo-European languages: English, French, Italian, and Spanish. For each language, 10 "problems" were constructed on the basis of a larger dataset obtained from `archiveofourown.org` in 2017. Per language, five problems were released as a development set to the participants in order to calibrate their systems. The final evaluation of the submitted systems was carried out on the five remaining problems that were not publicly released before the final results were communicated. Each problem had to be solved independently from the other problems. It should be noted that the development material could not be used as mere training material for supervised learning approaches since the candidate authors of the development corpus and the evaluation corpus do not overlap. Therefore, approaches

should not be designed to particularly handle the candidate authors of the development corpus but should focus on their generalizability to other author sets.

One problem corresponds to a single open-set attribution task, where we distinguish between the "source" and the "target" material. The source material in each problem contains exactly 7 training texts for exactly 9 candidate authors. In the target material, these 9 authors are represented by at least one test text (possibly more). Additionally, the target material also contains so-called "adversaries", which were not written by one of the candidate authors (indicated by the author label "<UNK>"). The proportion of the number of target texts written by the candidate authors in problems, as opposed to <UNK> documents, was varied across the problems in the development dataset, in order to discourage systems from opportunistic guessing.

Let U_K be the subset of U that includes all test documents actually written by the candidate authors, and let U_U be the subset of U containing the rest of the test documents not written by any candidate author. Then, the *adversary ratio* $r = |U_U|/|U_K|$ determines the likelihood of a test document to belong to one of the candidates. If $r = 0$ or close to 0), the problem is essentially a closed-set attribution scenario since all test documents belong to the candidate authors, or very few are actually written by adversaries. If $r = 1$, then it is equally probable for a test document to be written by a candidate author or by another author. For $r > 1$ it is more likely for a test document to be written by an adversary not included in the list of candidates.

We examine cases where r ranges from 0.2 to 1.0. In greater detail, as can be seen in Table 1, the development dataset comprises 5 problems per language that correspond to $r = [0.2, 0.4, 0.6, 0.8, 1.0]$. This dataset was released for the participants in order to develop and calibrate their submissions. The final evaluation dataset also includes 5 problems per language but with fixed $r = 1$. Thus, the participants are implicitly encouraged to develop generic approaches, because of the varying likelihood that a test document is written by a candidate or an adversary. In addition, it is possible to estimate the effectiveness of submitted methods when $r < 1$ by ignoring their answers for specific subsets of U_U in the evaluation dataset.

Each of the individual texts belongs to a single fandom, i.e., a certain topical domain. Fandoms were made available in the training material so that systems could exploit this information, as done by Seroussi et al. [10] for instance. We only selected works counting at least 500 tokens (according to the original database's internal token count), which is already a challenging text length for authorship analyses. Finally, we normalized document length: for fics longer than 1 000 tokens, we only included the middle 1 000 tokens of the text.

Another novelty this year was the inclusion of a set of 5 000 problem-external documents per language written by "imposter" authors (the authorship of these texts is also encoded as <UNK>). These documents could be freely used by the participants to develop their systems, for instance in the popular framework of imposter verification [4]. The release of this additional corpus should be understood against the backdrop of PAN's benchmarking efforts to improve the reproducibility and comparability of approaches. Many papers using imposter-based approaches do so on ad-hoc collected corpora, making it hard to compare the effect of the composition of the imposter collection. The imposter collection is given for the language as a whole and is thus not

Table 1. Details about the fanfiction datasets built for the cross-domain authorship attribution task. $|A|$ refers to the size of candidates list, $|K_a|$ is the amount of training documents per author a, $a \in A$, $|U|$ is the amount of test documents, r is the adversary ratio, and $\bar{l}$ denotes the average length of the documents in words.

	Language	**# Problems**	$\|A\|$	$\|K_a\|$	$\|U\|$	r	$\bar{l}$
Development	English	5	9	7	137-561	0.2 - 1.0	804
	French	5	9	7	38-430	0.2 - 1.0	790
	Italian	5	9	7	46-196	0.2 - 1.0	814
	Spanish	5	9	7	112-450	0.2 - 1.0	846
Evaluation	English	5	9	7	98-180	1.0	817
	French	5	9	7	48-290	1.0	790
	Italian	5	9	7	34-302	1.0	821
	Spanish	5	9	7	172-588	1.0	838

problem-specific. We provide the information that these texts were not written by authors who appear in the source or target sets for the problems in the language. When selecting these texts from the base dataset, we have given preference to texts from the fandoms covered in the problems, but when this selection was smaller than 5 000 texts, we have completed it with a random selection of other texts.

3 Evaluation Framework

In the final evaluation phase, submitted systems were presented with 5 problems per language: for each problem, given a set of documents (known fanfics) by candidate authors, the systems had to identify the authors of another set of documents (unknown fanfics) in a previously unencountered fandom (target domain). Systems could assume that each candidate author had contributed at least one of the unknown fanfics to the problem, which all belonged to the same target fandom. Some of the fanfics in the target domain, however, were not written by any of the candidate authors. Like in the calibration set, the known fanfics belonged to several fandoms (excluding the target fandom), although not necessarily the same for all candidate authors. An equal number of known fanfics per candidate author was provided: 7 fanfics for 9 authors. By contrast, the unknown fanfics were not equally distributed over the authors.

The submissions were separately evaluated in each attribution problem, based on their open-set macro-averaged F1 score (calculated over the training classes, i.e., when <UNK> is excluded) [11]. Participants were ranked according to their average open-set macro-F1 across all attribution problems of the evaluation corpus. A reference implementation of the evaluation script was made available to the participants.

3.1 Baseline Methods

As usual, we provide the implementation of three baseline methods that provide an estimation of the overall difficulty of the problem given the state of the art in the field.

These implementations are in Python (2.7+) and rely on Scikit-learn and its base packages [12,13] as well as NLTK [14]. Participants were free to base their approach on one of these reference systems, or to develop their own approach from scratch. The provided baseline are as follows:

1. BASELINE-SVM. A language-independent authorship attribution approach, framing attribution as a conventional text classification problem [15]. It is based on a character 3-gram representation and a linear SVM classifier with a reject option. First, it estimates the probabilities of output classes based on Platt's method [16]. Then, it assigns an unknown document to the <UNK> class when the difference of the probabilities of the top two candidates is less than a predefined threshold. Let a_1 and a_2, $a_1, a_2 \in A$, be the two most likely authors of a certain test document while Pr_1 and Pr_2 are the corresponding estimated probabilities (i.e., all other candidates obtained lower probabilities). Then, if $Pr_1 - Pr_2 < 0.1$, the document is left unattributed. Otherwise it is attributed to a_1.
2. BASELINE-COMPRESSOR. A language-independent approach that uses text compression to estimate the distance of an unknown document to each of the candidate authors. This approach was originally proposed by [17] and was later reproduced by [18]. It uses the Prediction by Partial Matching (PPM) compression algorithm to build a model for each candidate author. Then, it calculates the cross-entropy of each test document with respect to the model of each candidate and assigns the document to the author with the lowest score. In order to adapt this method to the open-set classification scenario, we introduced a reject option. In more detail, a test document is left unattributed when the difference between the two most likely candidates is lower than a predefined threshold. Let a_1 and a_2, $a_1, a_2 \in A$, be the two most likely candidate authors for a certain test document while S_1 and S_2 are their cross-entropy scores (i.e., all other candidate authors have higher scores). If $(S_1 - S_2)/S_1 < 0.01$, then the test document is left unattributed. Otherwise, it is assigned to a_1.
3. BASELINE-IMPOSTERS. Implementation of the language-independent imposters approach for authorship verification [4,19], based on character tetragram features. During a bootstrapped procedure, the technique iteratively compares an unknown text to each candidate author's training profile, as well as to a set of imposter documents, on the basis of a randomly selected feature subset. Then, the number of times the unknown document is found more similar to the candidate author's documents rather than to the imposters indicates how likely it is for that candidate to be the true author of the document. Instead of performing this procedure separately for each candidate author, we examine all candidate authors within each iteration (i.e., in each iteration, a maximum of one candidate author's score is increased). If after this repetitive process the highest score (corresponding to the most likely author) does not pass a fixed similarity threshold (here: 10% of repetitions), the document is assigned to the <UNK> class and is left unattributed. This baseline method is the only one that uses additional, problem-external imposter documents. We provided a collection of 5 000 imposter documents (fanfics on several fandoms) per language.

Finally, we also compare the participating systems to a plain "majority" baseline: through a simple voting procedure with random tie breaking, this baseline accepts a

candidate for a given unseen text if the majority of submitted methods agree on it; otherwise, the <UNK> label is predicted. No meta-learning is applied to weigh the importance of the votes of individual systems.

4 Survey of Submissions

In total, 12 methods were submitted to the task and evaluated using the TIRA experimentation framework. All but one (Kipnis) of the submissions are described in the participants' notebook papers. Table 5 presents an overview of the main characteristics of the submitted methods as well as the baselines. We also record whether approaches made use of the language-specific imposter material or language-specific NLP resources, such as pretrained taggers and parsers. As can be seen, there is surprisingly little variance in the approaches. The majority of submissions follow the paradigm of the BASELINE-SVM or the winner approach [20] of PAN-2018 cross-domain authorship attribution task [21], which is an ensemble of classifiers each of which based on a different text modality.

Compared to the baselines, most submitted methods attempt to exploit richer information that corresponds to different text modalities as well as variable-length n-grams [22] in contrast to fixed-length n-grams [23]. The most popular features are n-grams extracted from plain text modalities, such as character, word, token, part-of-speech tag, or syntactic level sequences. Given the cross-domain conditions of the task, several participants attempted to use more abstract forms of textual information such as punctuation sequences [24,22,25] or n-grams extracted from distorted versions [26] of the original documents [27,22,25]. There is limited effort to enrich n-gram features with alternative stylometric measures like word and sentence length distributions [28] or features related to syntactic analysis of documents [24,29]. Only one participant used word embeddings [30]. Other participants report that they were discouraged to use more demanding types of word and sentence embeddings due to hardware limitations of TIRA [31], which points to important infrastructural needs that may be addressed in future editions. Those same teams, however, informally reported that the difference in performance (in the development dataset) when using such additional features is negligible.

With respect to feature weighting, tf-idf is the most popular option while the baseline methods are based on the simpler tf scheme. There is one attempt to use both of these schemes [30]. A quite different approach uses a normalization scheme based on z-scores [29]. In addition, a few methods apply dimension reduction (PCA, SVD) to the features [32,33,22]. Judging from the results, such methods for dimension reduction have the potential to boost performance.

As concerns the classifiers, the most popular choices are SVMs and ensembles of classifiers, usually exploiting SVM base models followed by Logistic Regression (LR) models. In a few cases, the participants informally report that they have experimented with alternative classification algorithms (random forests, k-nn, naive Bayes) and found that SVM and LR are the most effective classifiers for this kind of task [27,30]. None of the participant's methods is based on deep learning algorithms, most probably due

to hardware limitations of TIRA or because of the discouraging reported results in the corresponding task of PAN-2018 [21].

Given the fact that the focus of PAN-2019 edition of the task is on open-set attribution, it can be noted that none of the participants attempted to build a pure open-set classifier [34]. By contrast, they just use closed-set classifiers with a reject option (the classification prediction is dropped when the confidence of prediction is low), similar to the baseline methods [35].

A crucial issue to improve the performance of authorship attribution is the appropriate tuning of hyperparameters. Most of the participants tune the hyperparameters of their approach globally based on the development dataset, that is, they estimate the most suitable parameter values that are applied to any problem of the test dataset. In contrast, a few participants attempt to tune the parameters of their method in a language-specific way, estimating the most suitable values for each language separately. None of the submitted methods attempts to tune parameter values for each individual attribution problem.

The submission of van Halteren focuses on the cross-domain difficulty of the task and attempts to exploit the availability of multiple texts of unknown authorship in the target domain within each attribution problem [29]. This submission performs a sophisticated strategy composed of different phases. Initially, a typical cross-domain classifier is built and each unknown document is assigned to its most likely candidate author but the prediction is kept only for the most confident cases. Then, a new in-domain classifier is built using the target domain documents (for which the predictions were kept in the previous phase) and the remaining target domain documents are classified accordingly. However, this in-domain classifier can only be useful for certain candidate authors, the ones with enough confident predictions in the initial phase. A final phase combines the results of cross-domain and in-domain classifiers and leaves documents with less confident predictions unattributed.

5 Evaluation Results

Table 2 shows an overview of the evaluation results of participants and their ranking according to their macro-F1 (averaged across all attribution problems of the dataset). As can be seen, all but one submission surpass the three baseline methods. In general, the submitted methods and the baselines achieve better macro-recall than macro-precision—which, interestingly, is not the case for the more prudent majority baseline. The two top-performing submissions obtain a very similar macro-F1 score. However, the winning approach of Muttenthaler et al. has better macro-precision while Bacciu et al. achieve better macro-recall. In terms of elapsed runtime, the winning approach of Muttenthaler et al. also proved to be a very efficient one.

Table 3 demonstrates the effectiveness (averaged macro-F1) of the submitted methods for each one of the four languages of the evaluation dataset. The winning approach of Muttenthaler et al. is more effective in English and French while the approach of Bacciu et al. achieves comparable performance in Italian and Spanish. In general, the variation of top-performing approaches across the four languages is low. On average, the highest performance is obtained for attribution problems in Italian; English proved

Table 2. The final evaluation results of the cross-domain authorship attribution task. Participants and baselines are ranked according to macro-F1.

Submission	Macro-Precision	Macro-Recall	Macro-F1	Runtime
Muttenthaler et al.	**0.716**	0.742	**0.690**	00:33:17
MAJORITY	0.748	0.708	0.686	
Bacciu et al.	0.688	**0.768**	0.680	01:06:08
Custodio & Paraboni	0.664	0.717	0.65	01:21:13
Bartelds & de Vries	0.657	0.719	0.644	11:19:32
Rodríguez et al.	0.651	0.713	0.642	01:59:17
Isbister	0.629	0.706	0.622	01:05:32
Johansson	0.593	0.734	0.616	01:05:30
Basile	0.616	0.692	0.613	00:17:08
van Halteren	0.590	0.734	0.598	37:05:47
Rahgouy et al.	0.601	0.633	0.580	02:52:03
Gagala	0.689	0.593	0.576	08:22:33
BASELINE-SVM	0.552	0.635	0.545	
BASELINE-COMPRESSOR	0.561	0.629	0.533	
BASELINE-IMPOSTERS	0.428	0.580	0.395	
Kipnis	0.270	0.409	0.259	20:20:21

to be the most difficult case. It is also remarkable that the baseline-compressor method achieves the best baseline results for English, French, and Italian, but it is not as competitive in Spanish. Furthermore, note that Muttenthaler et al.'s submission is the only one to outperform the otherwise very competitive majority baseline, albeit by a very small margin. The latter reaches a relatively high precision, but must sacrifice quite a bit of recall in return. That the winner outperforms the majority baseline is surprising: in previous editions of this shared task (e.g. [36]), similar meta-level approaches proved very hard to beat. This result is probably an artifact of the lack of diversity among the submissions in the top-scoring cohort, which seem to have produced very similar predictions (see below), thus reducing the beneficial effects of a majority vote among those systems.

In order to examine the effectiveness of the submitted methods for a varying adversary ratio, we performed the following additional evaluation process. As can be seen in Table 1, all attribution problems of the evaluation dataset have a fixed adversary ratio $r = 1$, meaning that an equal number of documents written either by the candidate authors or adversary authors is included in the test set of each problem. Once the submitted methods processed the whole evaluation dataset, we calculated the evaluation measures with decreasing proportions of adversary documents at 100%, 80%, 60%, 40%, and 20%, resulting in an adversary ratio that ranges from 1 to 0.2. Table 4 presents the evaluation results (averaged macro-F1) for such a varying adversary ratio. In general, the performance of all methods increases when the adversary ratio decreases. Recall that $r = 0$ corresponds to a closed-set attribution case. The performance of the two top-performing approaches is very similar in the whole range of examined r-values. However, the method of Muttenthaler et al. is slightly better for high r-values while Bacciu et al. is slightly better for low r-values.

Table 3. Results (macro-F1) per language of the cross-domain authorship attribution task. Participants and baselines are ranked according to their overall macro-F1.

Submission	English	French	Italian	Spanish
Muttenthaler et al.	**0.665**	**0.705**	**0.717**	0.673
Bacciu et al.	0.638	0.689	0.715	**0.679**
Custodio & Paraboni	0.587	0.686	0.682	0.647
Bartelds & de Vries	0.558	0.687	0.700	0.629
Rodríguez et al.	0.597	0.624	0.696	0.651
Isbister	0.529	0.644	0.691	0.623
Johansson	0.613	0.593	0.655	0.602
Basile	0.555	0.628	0.656	0.613
van Halteren	0.532	0.554	0.653	0.652
Rahgouy et al.	0.550	0.583	0.595	0.592
Gagala	0.554	0.564	0.566	0.619
BASELINE-SVM	0.490	0.548	0.566	0.577
BASELINE-COMPRESSOR	0.493	0.595	0.580	0.464
BASELINE-IMPOSTERS	0.359	0.409	0.410	0.400
Kipnis	0.301	0.232	0.285	0.220

Moreover, we have applied statistical significance tests to the systems' output. Especially since many systems have adopted a similar approach, it is worthwhile to discuss the extent to which submissions show statistically meaningful differences. Like in previous editions, we have applied approximate randomization testing, a non-parametric procedure that accounts for the fact that we should not make too many assumptions as to the underlying distributions for the classification labels. Table 6 lists the results for pairwise tests, comparing all submitted approaches to each other, based on their respective F1-scores for all labels in the problems. For 1 000 bootstrapped iterations, the test returns probabilities which we can interpret as the conventional p-values of one-sided, statistical tests—i.e., the probability of failing to reject the null hypothesis (H0) that the classifiers do not output significantly different scores. The symbolic notation takes into account the following straightforward thresholds: '=' (not significantly different: $p > 0.5$), '*' (significantly different: $p < 0.05$), '**' (very significantly different: $p < 0.01$), '***' (highly significantly different: $p < 0.001$). Interestingly, systems with neighboring ranks often do not yield significantly different scores; this is also true for the two top-performing systems. Note that almost all systems have produced an output that is significantly different from the three baselines (which also display a high degree of difference among one another). According to this test, the difference between Muttenthaler et al. and Bacciu et al. is not statistically significant, although the former is significantly different from the majority baseline.

Table 4. Evaluation results (macro-F1) of the cross-domain authorship attribution task for different values of the adversary ratio r. Participants and baselines are ranked according to their overall macro-F1.

	r				
Submission	**100%**	**80%**	**60%**	**40%**	**20%**
Muttenthaler et al.	**0.690**	**0.709**	**0.727**	0.746	0.773
Bacciu et al.	0.680	0.701	0.724	**0.749**	**0.777**
Custodio & Paraboni	0.650	0.666	0.686	0.704	0.728
Bartelds & de Vries	0.644	0.663	0.683	0.708	0.736
Rodríguez et al.	0.642	0.664	0.684	0.704	0.733
Isbister	0.622	0.642	0.664	0.685	0.716
Johansson	0.616	0.641	0.666	0.700	0.735
Basile	0.613	0.633	0.654	0.675	0.706
van Halteren	0.598	0.622	0.645	0.672	0.701
Rahgouy et al.	0.580	0.599	0.619	0.642	0.664
Gagala	0.576	0.586	0.597	0.610	0.624
BASELINE-SVM	0.545	0.563	0.585	0.611	0.642
BASELINE-COMPRESSOR	0.533	0.548	0.569	0.592	0.620
BASELINE-IMPOSTERS	0.395	0.409	0.429	0.453	0.484
Kipnis	0.259	0.270	0.285	0.302	0.324

Table 5. Comparison of the core components of the submitted systems.

Participant	Features	Weighting	Feature transformation/selection	Parameter tuning	Classifier	Open-set criterion	Use imposters data	Language-dependent resources
Bacciu et al.	n-grams (char, word, POS, stem, distortion)	tf-idf	NA	per language	Ensemble (SVM)	Reject	No	stemming, POS
Bartelds and de Vries	n-grams (char, token, POS, punctuation, ...	tf-idf	NA	global	SVM	Reject	No	POS, syntactic parse
Basile	n-grams (char and word)	tf-idf	NA	global	SVM	Reject	No	None
Custodio et al.	n-grams (char, word, distortion), ...	tf-idf	PCA	global	Ensemble (LR)	Reject	No	None
Gagala	n-grams (char, word)	tf-idf	PCA	global	Imposters (LR)	Verification	Yes	None
Isbister	n-grams (char, word), word and sentence ...	tf-idf	NA	global	SVM	Reject	No	None
Johansson	n-grams (char, word, POS, distortion), word ...	tf-idf	NA	global	SVM	Reject	No	POS
Muttenthaler et al.	n-grams (char, word, distortion, punctuation)	tf-idf	SVD	global	Ensemble (SVM)	Reject	No	None
Rahgouy et al.	n-grams (char and word), word embeddings	tf-idf and tf	NA	global	Ensemble (SVM)	Reject	No	stemming
Rodríguez et al.	n-grams (char, typed, punctuation, word)	tf-idf	NA	global	Ensemble (SVM)	Reject	No	None
Van Halteren	n-grams (char, token, syntactic)	z-score	NA	per language	Ensemble (distance-based and SVR)	Reject	No	POS, syntactic parse
baseline-SVM	n-grams (char)	tf	NA	global	SVM	Reject	No	None
baseline-Compressor	char sequences	none	NA	global	PPM	Reject	No	None
baseline-Imposters	n-grams (char)	tf	NA	global	distance-based	Verification	Yes	None

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

	Muttenthaler et al.	majority	Bacciu et al.	Custodio & Paraboni	Bartelds & de Vries	Rodríguez et al.	Isbister	Johansson	Basile	Van Halteren	Rahgouy et al.	Gagala	Baseline-svm	Baseline-compressor	Baseline-impostors	Kipnis
Muttenthaler et al.		**	=	***	***	***	***	***	***	***	***	**	***	***	***	***
Majority			***	***	***	***	***	***	***	***	***	***	***	***	***	***
Bacciu et al.				=	***	***	***	***	***	***	***	*	***	***	***	***
Custodio & Paraboni					=	=	**	***	***	***	***	=	***	***	***	***
Bartelds & de Vries						=	=	***	=	**	*	=	***	***	***	***
Rodríguez et al.							*	***	***	***	**	=	***	***	***	***
Isbister								***	=	=	=	=	***	***	***	***
Johansson									**	=	**	***	=	***	***	***
Basile										=	=	***	***	***	***	***
Van Halteren											=	***	***	***	***	***
Rahgouy et al.												**	***	***	***	***
Gagala													***	***	***	***
Baseline-svm														***	***	***
Baseline-compressor															***	***
baseline-impostors																***

6 Conclusion

The paper discussed the 12 submissions to the 2019 edition of the PAN shared task on authorship identification. Like last year, we focused on cross-domain attribution in fanfiction data. An important innovation this year was the focus on the open-set attribution set-up, where participating systems had to be able to refrain from attributing unseen texts as well. The analyses described above call for a number of considerations that are not without relevance to future development in the field of computational authorship identification. First of all, this year's edition was characterized by a relative low degree of diversity in approaches: especially the higher-scoring cohort almost exclusively adopted a highly similar approach, involving a combination of SVMs as classifier (potentially as part of an ensembles), character n-grams as features, and a rather simple thresholding mechanism to refrain from attributions. It is not immediately clear which directions future research might explore. Deep learning-based methods, which can be pretrained on external corpora, have so far not led to a major breakthrough in the field, despite the impressive improvements which have been reported for these methods in other areas of NLP. Also, a more promising research direction might be to move away from closed-set classifiers (with a naive reject-option), towards purely open-set classifiers [34]

References

1. M. Koppel, J. Schler, and S. Argamon. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26, 2009.
2. P. Juola. Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3):233–334, 2006.
3. Efstathios Stamatatos. A survey of modern authorship attribution methods. *JASIST*, 60(3):538–556, 2009.
4. Moshe Koppel and Yaron Winter. Determining if two documents are written by the same author. *Journal of the Association for Information Science and Technology*, 65(1):178–187, 2014.
5. Karen Hellekson and Kristina Busse, editors. *The Fan Fiction Studies Reader*. University of Iowa Press, 2014.
6. K. Mirmohamadi. *The Digital Afterlives of Jane Austen. Janeites at the Keyboard*. Palgrave MacMillan, 2014.
7. J. Fathallah. *Fanfiction and the Author. How FanFic Changes Popular Cultural Texts*. Amsterdam University Press, 2017.
8. R. Tushnet. Legal fictions: Copyright, fan fiction, and a new common law. *Loyola of Los Angeles Entertainment Law Review*, 17(3), 1997.
9. Efstathios Stamatatos, Francisco M. Rangel Pardo, Michael Tschuggnall, Benno Stein, Mike Kestemont, Paolo Rosso, and Martin Potthast. Overview of PAN 2018 - author identification, author profiling, and author obfuscation. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings*, pages 267–285, 2018.
10. Yanir Seroussi, Ingrid Zukerman, and Fabian Bohnert. Authorship attribution with topic models. *Computational Linguistics*, 40(2):269–310, 2014.

11. Pedro R. Mendes Júnior, Roberto M. de Souza, Rafael de O. Werneck, Bernardo V. Stein, Daniel V. Pazinato, Waldir R. de Almeida, Otávio A. B. Penatti, Ricardo da S. Torres, and Anderson Rocha. Nearest neighbors distance ratio open-set classifier. *Machine Learning*, 106(3):359–386, Mar 2017.
12. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
13. Travis Oliphant. NumPy: A guide to NumPy. USA: Trelgol Publishing, 2006.
14. Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009.
15. F. Sebastiani. Machine Learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.
16. J. Platt. Probabilistic outputs for support vector machines and comparison to regularize likelihood methods. In A.J. Smola, P. Bartlett, B. Schoelkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74, 2000.
17. William J. Teahan and David J. Harper. *Using Compression-Based Language Models for Text Categorization*, pages 141–165. Springer Netherlands, Dordrecht, 2003.
18. Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Proc. of the European Conference on Information Retrieval*, pages 393–407. Springer International Publishing, 2016.
19. Mike Kestemont, Justin Anthony Stover, Moshe Koppel, Folgert Karsdorp, and Walter Daelemans. Authenticating the writings of julius caesar. *Expert Systems with Applications*, 63:86–96, 2016.
20. José Eleandro Custódio and Ivandré Paraboni. EACH-USP Ensemble cross-domain authorship attribution. In Linda Cappellato, Nicola Ferro, Jian-Yun Nie, and Laure Soulier, editors, *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings. CLEF and CEUR-WS.org, 2018.
21. Mike Kestemont, Michael Tschuggnall, Efstathios Stamatatos, Walter Daelemans, Günther Specht, Benno Stein, and Martin Potthast. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum*. CEUR-WS.org, 2018.
22. Lukas Muttenthaler, Gordon Lucas, and Janek Amann. Authorship Attribution in Fan-fictional Texts Given Variable Length Character and Word n-grams. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
23. Angelo Basile. An Open-Vocabulary Approach to Authorship Attribution. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
24. Martijn Bartelds and Wietse de Vries. Improving Cross-Domain Authorship Attribution by Combining Lexical and Syntactic Features. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.

25. Carolina Martín del Campo Rodríguez, Daniel Alejandro Pérez Álvarez, Christian Efraín Maldonado Sifuentes, Grigori Sidorov, Ildar Batyrshin, and Alexander Gelbukh. Authorship Attribution through Punctuation n-grams and Averaged Combination of SVM. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
26. Efstathios Stamatatos. 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*, pages 1138–1149. Association for Computational Linguistics, 2017.
27. Andrea Bacciu, Massimo La Morgia, Alessandro Mei, Eugenio Nerio Nemmi, Valerio Neri, and Julinda Stefa. Cross-domain Authorship Attribution Combining Instance Based and Profile Based Features. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
28. Fredrik Johansson and Tim Isbister. FOI Cross-Domain Authorship Attribution for Criminal Investigations. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
29. Hans van Halteren. Cross-Domain Authorship Attribution with Federales. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
30. Mostafa Rahgouy, Hamed Babaei Giglou, Taher Rahgooy, Mohammad Karami Sheykhlan, and Erfan Mohammadzadeh. Cross-domain Authorship Attribution: Author Identification using a Multi-Aspect Ensemble Approach. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
31. Martin Potthast, Tim Gollub, Matti Wiegmann, and Benno Stein. TIRA Integrated Research Architecture. In Nicola Ferro and Carol Peters, editors, *Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF*. Springer, 2019.
32. José Custódio and Ivandre Paraboni. Multi-channel Open-set Cross-domain Authorship Attribution. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
33. Lukasz Gagala. Authorship Attribution with Logistic Regression and Imposters. In Linda Cappellato, Nicola Ferro, David E. Losada, and Henning Müller, editors, *CLEF 2019 Labs and Workshops, Notebook Papers*. CEUR-WS.org, 2019.
34. Walter Scheirer, Anderson Rocha, Archana Sapkota, and Terrance Boult. Toward open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, 2013.
35. Chuanxing Geng, Sheng-jun Huang, and Songcan Chen. Recent advances in open set recognition: A survey. *arXiv preprint arXiv:1811.08581*, 2018.
36. Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Benno Stein, Martin Potthast, Patrick Juola, Miguel A. Sánchez-Pérez, and Alberto Barrón-Cedeño. Overview of the author identification task at pan 2014. In *CLEF 2014 Evaluation Labs and Workshop – Working Notes Papers*, Sheffield, UK, September 2014.