Overview of the Cross-Domain Authorship Verification Task at PAN 2021

Mike Kestemont[1], Enrique Manjavacas[1], Ilia Markov[1], Janek Bevendorff[2], Matti Wiegmann[2], Efstathios Stamatatos[3], Benno Stein[2] and Martin Potthast[4]

[1]*University of Antwerp*
[2]*Bauhaus-Universität Weimar*
[3]*University of the Aegean*
[4]*Leipzig University*
pan@webis.de https://pan.webis.de

Abstract

Idiosyncrasies in human writing styles make it difficult to develop systems for authorship identification that scale well across individuals. In this year's edition of PAN, the authorship identification track focused on open-set authorship verification, so that systems are applied to unknown documents by previously unseen authors in a new domain. As in the previous year, the sizable materials for this campaign were sampled from English-language fanfiction. The calibration materials handed out to the participants were the same as last year, but a new test set was compiled with authors and fandom domains not present in any of the previous datasets. The general setup of the task did not change, i.e., systems still had to estimate the probability of a pair of documents being authored by the same person. We attracted 13 submissions by 10 international teams, which were compared to three complementary baselines, using five diverse evaluation metrics. Post-hoc analyses show that systems benefitted from the abundant calibration materials and were well-equipped to handle the open-set scenario: Both the top-performing approach and the highly competitive cohort of runner-ups presented surprisingly strong verifiers. We conclude that, at least within this specific text variety, (large-scale) open-set authorship verification is not necessarily or inherently more difficult than a closed-set setup, which offers encouraging perspectives for the future of the field.

1. Introduction

This paper provides a full-length description of the authorship verification shared task at PAN 2021. This edition was the second task installment in a renewed three-year program on the PAN authorship track (2020–2022), in which the scope, the difficulty and, the realism of the tasks are gradually increased each year. After last year's edition focused on providing participants with the largest pool of calibration material by far of any previous authorship shared task at PAN—a technical challenge in its own right—, we sought to improve the difficulty this year by sampling a fully disjunct test set. This is different to last year's edition where the overall task difficulty was kept in check by means of resorting to a *closed-set* evaluation scenario in which the test set was restricted to only authors and fandom domains also included in the calibration set (hence a clever participant could re-cast the task as an *attribution* task). This year's test set,

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

on the other hand, comes with document pairs of exclusively unseen authors writing in unseen fandom domains, which results in an *open-set* or "true" authorship *verification* scenario, which is conventionally considered a much more demanding problem than attribution. For the next year, we planned a consecutive and final "surprise task", on which more details will be released in due time.

In the following, we first contextualize and motivate the design choices outlined above. Next, we shall formalize the task, describe the composition of this year's test set, and detail the employed evaluation metrics as well as the three generic baseline systems that were applied as a point of reference. In the sections following that, we shall briefly discuss the participating systems through a summary of their respective notebooks and the results of the task in tandem with a statistical analysis to assess whether pairs of systems in fact produced significantly different outcomes. In our discussion, we present post hoc analyses—including a comparison with last year's results regarding the distribution of scores—, the effect of non-answers, and the relationship of stylistic and topical similarities. Finally, we assess the contributions of this year's edition regarding the closed-set vs. open-set debate and offer an outlook into the future.

1.1. Motivation and Design Rationale

Much of the research in present-day computational authorship identification is implicitly underpinned by a basic assumption that could be summarized as the "Stylome Hypothesis". This hypothesis, seminally formulated by van Halteren et al. [1], states that all writing individuals would leave a unique stylistic and linguistic "fingerprint" in their work, i.e., a set of stable empirical characteristics that can be extracted from and identified in a large-enough writing sample. In the analogy of the human genome, the assumption is that this fingerprint is a sufficient means to identifying the author of any given writing sample, provided it is long enough. The Stylome Hypothesis is an attractive working hypothesis, but remains hard to demonstrate, let alone prove. Experimental studies in the past decades have enabled scholars to close in on the experimental conditions that must be met for an authorship identification: we know that texts have to be long enough to be analyzed in the first place and verification across different text varieties has proven to be very challenging, not to mention issues of collaborative authorship or copy editors who inject additional stylistic noise. Cases where a reliable set of candidate authors is already available, are easier to solve than those where such a list cannot be established.

One general property of human authorship that has emerged in various studies appears to be its ad-hoc nature: Even within a single genre, textual features that work well to differentiate author A from a set of peers, might fail to separate author B from the same set of peers. Due to the many idiosyncracies that occur in an individual's writing style, this makes it challenging to develop systems that can be robustly scaled across many different individuals. Modeling authorial writing style requires bespoke models that are tailored to the characteristics of a single author or a specific set of authors. These observations tie in with two important scenarios that are commonly distinguished in the field: closed-set and open-set authorship verification. The former term describes the situation in which a system is applied to a set of texts by authors who are already known to the system (as they were seen during the training or calibration phase). The latter term describes the scenario in which a system is applied to texts whose authors are

(potentially) unknown. This open-set scenario is supposedly much more challenging, since one would expect verification systems to overfit on textual properties that are significant for distinguishing this author from their known peers, but which may eventually turn out not to be a general characteristic of their style and hence not distinguish them from other, unknown authors.

This state of affairs has clearly motivated and shaped the shared task in authorship identification at PAN over the years. In particular, three factors have informed the design of the tasks: (1) issues of scale, (2) methodological developments, and (3) the ad-hoc nature of authorship. First of all, to reliably assess the plausibility of the Stylome Hypothesis, much larger corpora are required than were previously available. It is only in recent years, in fact, that larger datasets for authorship attribution have become more widespread. This concern relating to scale is closely related to methodological developments in the field. In the 2018 task overview paper, the organizers voiced serious concerns about a noticeable lack of diversity in the submitted systems. Save a few exceptions, most of the systems then took the form of a simple classifier (typically a linear SVM or decision tree) that was applied to a bag-of-words representation of documents on the basis of character n-grams and other conventional feature sets. This methodological dearth was remarkable, since deep (neural) representation learning had already been shaping the landscape of NLP for several years. Such late adoption of deep neural models for authorship identification was very likely an immediate result of a lack of sufficient training resources as are typically required for representation learning (in particular for the data-hungry pre-training and finetuning of sentence- or document-level embeddings).

2. Authorship Verification

The most central element of authorship analysis is the identification of the document's author(s) [2, 3, 4]. In various fields, scholars have been studying how stylistic and linguistic properties of documents can be harnessed for the achievement of this goal. Because of the variety in authorial styles, including diachronic and synchronic shifts, progress in the field of style-based document authentication is hard to monitor, as it requires extensive, transparent, and repeated benchmarking initiatives [5]. The long-running authorship identification track at PAN hopes to contribute in this area and has organized tasks on authorship identification in various guises. The following section offers an overview of the central concepts as an update on a previously published survey [6]:

- Authorship Attribution: Given a document and a set of candidate authors, determine who wrote the document (PAN 2011–2012, 2016–2020);
- Authorship Verification: Given a pair (or collection) of documents, determine whether they are written by the same author (PAN 2013–2015, 2021);
- Authorship Obfuscation: Given a set of documents by the same author, paraphrase one or all of them so that its author cannot be identified anymore (PAN 2016–2018);
- Obfuscation Evaluation: Devise and implement performance measures that quantify safeness, soundness, and / or sensibleness of an obfuscation software (PAN 2016–2018).

The formal goal of authorship verification is to approximate the target function $\phi : (D_k, d_u) \rightarrow \{T, F\}$, where D_k is a set of documents of known authorship by the same author and d_u is a document of unknown or questioned authorship.[1] If $\phi(D_k, d_u) = T$, then the author of D_k is also the author of d_u and if $\phi(D_k, d_u) = F$, then the author of D_k is not the same as the author of d_u. In the case of cross-domain verification, D_k and d_u stem from a different text variety or encompass considerably different content (e.g. topics or themes, genres, registers, etc.). For the present task, we considered the simplest (and most challenging) formulation of the verification task, i.e., we only considered cases where D_k is a singleton, thus only pairs of two documents are examined. Given a training set of such problems, the verification systems of the participating teams had to be trained and calibrated to analyze the authorship of the unseen text pairs (from the test set). We shall distinguish between same-author text pairs (SA: $\phi(D_k, d_u) = T$) and different-author (DA: $\phi(D_k, d_u) = F$) text pairs. In terms of setup, the novelty this year was that (1) the authors and (2) the stories' fandom domains in the test set were not part of any of the provided calibration materials, which, theoretically speaking, should make this year's task more challenging than last year's.

3. Datasets

Given our aim to benchmark authorship identification systems at a much larger scale, our tasks in recent years [8, 9] focused on transformative literature, or so-called "fanfiction" [10], a text variety that is nowadays abundantly available on the internet [11] with rich metadata and in many languages. Additionally, fanfiction is an excellent source of material for studies of cross-domain scenarios, since users often publish "fics" ranging over multiple topical domains ("fandoms"), such as Harry Potter, Twilight, or Marvel comics. The datasets we provided for our tasks at PAN 2020 and PAN 2021 were crawled from the long-established fanfiction community `fanfiction.net`. Access to the data can be requested on Zenodo.[2] The 2021 edition of the authorship verification task built upon last year's [7] with the same general task layout and training data, but with a conceptually different test set. We retained the overall cross-domain setting, in which the texts in a pair stem from different fandoms, but we replaced the closed-set setting with an open-set setting, where both the authors and the fandoms in the test set are entirely "new" and do not occur in the training set.

The training resources were identical to those from last year and came in a "small" and a "large" variant. The large dataset contains 148,000 same-author and 128,000 different-author pairs across 1,600 fandoms. Each single author has written in at least two, but not more than six fandoms. The small training set is a subset of the large training set with 28,000 same-author and 25,000 different-author pairs from the same 1,600 fandoms. The new test was sampled with the same general strategy (19,999 text pairs in total), but in a way so as to fulfill the previously described open-set constraints to make the task—at least in theory—more difficult.

[1] This paragraph is based on last year's overview paper [7] and included for the sake of completeness.
[2] `https://zenodo.org/record/3716403`

4. Evaluation Framework

For each of the 19,999 problems (or document pairs) in the test set, the systems had to produce a scalar score a_i in the range $[0, 1]$ indicating the (scaled) probability that the pair was written by the same author ($a_i > 0.5$) or different authors ($a_i < 0.5$). Systems could choose to leave problems they deemed too difficult to decide unanswered by submitting a score of precisely $a_i = 0.5$. Such a non-answer is rewarded by some of the metrics over a wrong answer.

4.1. Performance Measures

Similar to year, we adopted a diverse mix of evaluation metrics that focused on different aspects of the verification task at hand. We reused the four evaluation metrics from the 2020 edition, but also included the (complement of the) Brier score [12] as an additional fifth metric (following discussions with participants and audience from the 2020 workshop[3]). The following performance measures were used:

- AUC: the ROC area-under-the-curve score,
- c@1: a variant of the conventional accuracy measure, which rewards systems that leave difficult problems unanswered [13],
- F_1: the well-known F_1 performance measure (*not* taking into account non-answers),
- $\mathrm{F}_{0.5u}$: a newly-proposed $\mathrm{F}_{0.5}$-based measure that emphasizes correctly-answered same-author cases and rewards non-answers [14],
- Brier: the Brier score (more precisely: the complement of the Brier score loss function [12] as implemented in `sklearn` [15]), a straightforward, strictly proper scoring rule that measures the accuracy of probabilistic predictions.

The inclusion of the Brier score was an addition which was meant to measure the probabilistic confidence of the verifiers in a more fine-grained manner. This metric rewards verifiers that produce bolder but correct scores (i.e., a_i close to 0.0 or 1.0). Conversely, the metric would indirectly penalize less committal solutions, such as non-answers ($a_i = 0.5$).

To produce a final ranking for a system, we used the mean score across all individual measures.

4.2. Baselines

In total, we provided three baseline systems (calibrated on the small training set) for comparison, of which the first two were also employed during last year's competition. These were a compression-based approach [16] and a naive distance-based, first-order bag-of-words model [17]. Both were made available to participants at the start. The third baseline was a post-hoc addition for this overview paper and consisted of a short-text variant of Koppel and Schler's unmasking [18, 19], which had yielded good empirical results in the recent past.

[3] Thanks to Fabrizio Sebastiani (Consiglio Nazionale delle Ricerche, Italy) for this suggestion.

5. Survey of Submissions

The authorship verification task received 13 submissions from 10 participating teams. In this section, we provide a short and concise overview of the submitted systems. For further details (including bibliographic references), we refer the interested reader to the full versions of these notebooks. Teams were allowed to hand in exactly one submission per training dataset (large and small). Three teams submitted two systems, the other teams either deliberately chose to submit only a single variant or were unable to produce a valid run in time. The systems listed below are described in the order in which the notebooks were initially submitted.

1. **ikae21** [20] used a hard majority-voting ensemble that incorporated five different machine-learning classifiers (i.e., linear discriminant analysis, gradient boosting, extra trees, support vector machines, and stochastic gradient descent). The features used were top-800 TF-IDF-weighted word unigrams.
2. **menta21** [21] exploited two types of stylometric features, character n-grams and punctuation marks, to train a neural network on each type of feature separately. The outputs were concatenated and fed into another neural network in order to obtain the predictions.
3. **liaozhihao21** [22] used four retrieval models from the Lucene framework. Each retrieval model assigned a probability to a piece of text that it was written by the corresponding author. Later on, a weighted average of the probabilities was calculated to get the final score. The approach assumes that both texts were written by the same author if the highest final score corresponds to the same author.
4. **weerasinghe21** [23] extracted stylometric features from each text pair and used the absolute differences between the feature vectors as input to the logistic regression classifier. The features included character and POS n-grams, special characters, function words, vocabulary richness, POS-tag chunks, and unique spellings.
5. **boenninghoff21** [24] presented a hybrid neural-probabilistic end-to-end framework, which included neural feature extraction and deep metric learning, deep Bayes factor scoring, uncertainty modeling and adaptation, a combined loss function, and additionally an out-of-distribution detector for defining non-responses. In the final step, the model was extended to a majority-voting ensemble.
6. **peng21** [25] proposed an approach that split the texts into fragments and used BERT to extract feature vectors from each fragment, which were then concatenated and fed into a neural network for the final predictions.
7. **futrzynski21** [26] proposed an approach based on the cosine similarities of output representations extracted from BERT. These similarities were compared to several thresholds and were rescaled in order to classify a text pair. The BERT model was trained on the following tasks: masked language modeling, author classification, fandom classification, and author-fandom separation. In addition, the authors proposed a method for decreasing the computational costs by combining embeddings of many short text sequences.
8. **embarcaderoruiz21** [27] proposed a novel approach consisting of a graph representation to represent the texts, which served as input to a Siamese network. The feature extraction network consisted of node embedding layers to obtain vector representations for each

Table 1
System rankings for all PAN 2021 submissions across five evaluation metrics: AUC, c@1, F_1, $F_{0.5u}$, Brier, and an overall mean score (as the final ranking criterion). The dataset column indicates which calibration dataset was used. Bold digits reflect the per-column maximum. Horizontal lines indicate the range of scores yielded by the baselines (in italics).

Team	Dataset	AUC	c@1	F_1	$F_{0.5u}$	Brier	Overall
boenninghoff21	large	**0.9869**	**0.9502**	**0.9524**	**0.9378**	**0.9452**	**0.9545**
embarcaderoruiz21	large	0.9697	0.9306	0.9342	0.9147	0.9305	0.9359
weerasinghe21	large	0.9719	0.9172	0.9159	0.9245	0.9340	0.9327
weerasinghe21	small	0.9666	0.9103	0.9071	0.9270	0.9290	0.9280
menta21	large	0.9635	0.9024	0.8990	0.9186	0.9155	0.9198
peng21	small	0.9172	0.9172	0.9167	0.9200	0.9172	0.9177
embarcaderoruiz21	small	0.9470	0.8982	0.9040	0.8785	0.9072	0.9070
menta21	small	0.9385	0.8662	0.8620	0.8787	0.8762	0.8843
rabinovits21	small	0.8129	0.8129	0.8094	0.8186	0.8129	0.8133
ikae21	small	0.9041	0.7586	0.8145	0.7233	0.8247	0.8050
unmasking21	small	0.8298	0.7707	0.7803	0.7466	0.7904	0.7836
tyo21	large	0.8275	0.7594	0.7911	0.7257	0.8123	0.7832
naive21	small	0.7956	0.7320	0.7856	0.6998	0.7867	0.7600
compressor21	small	0.7896	0.7282	0.7609	0.7027	0.8094	0.7581
futrzynski21	large	0.7982	0.6632	0.8324	0.6682	0.7957	0.7516
liaozhihao21	small	0.4962	0.4962	0.0067	0.0161	0.4962	0.3023

node in the graph as well as a global pooling. The authors also incorporated stylometric features, combining them with the graph components to an ensemble.

9. **tyo21** [28] used BERT within a Siamese network. The embedding space was optimized so that texts written by the same author are adjacent in that space, while texts written by different authors are farther apart. At inference time, the distance between embeddings was compared to a threshold (selected based on a grid search) to make the predictions.
10. **rabinovits21** [29] relied on regression models. The authors incorporated the cosine distance for a set of vector-based features (word-, and character frequencies, POS tags, POS chunk n-grams, punctuation, stopwords) and absolute differences for scalar features (vocabulary richness, average sentence length, Flesch reading ease score) as measures of text-pair similarity. The concatenated similarity scores were used as input to a random forest model (adapted as a regressor).

Overall, we observe a healthy diversity of methods, with several novel approaches, for instance from representation learning with neural networks, appearing among more established methods from text classification or information retrieval. Multiple teams employed a so-called "Siamese" neural network approach [30], which seems to be a natural choice for the analysis of text pairs.

6. Evaluation Results

Table 1 offers a tabular representation of the final results of the submitted systems on the PAN 2021 test set. The overall ranking is based on the mean performance of the five evaluation metrics (last column). The dataset column indicates whether a system was calibrated on the "large" or "small" dataset. In the following, we refer to these as "large" and "small" systems or

Table 2
Significance of pairwise differences in F_1 scores between submissions. Notation: '=' (not significant: $p \geq 0.05$), '*' (significant with $p < 0.05$), '**' (significant with $p < 0.01$), '***' (significant with $p < 0.001$).

	embarcaderoruiz21-large	weerasinghe21-large	weerasinghe21-small	menta21-large	peng21-small	embarcaderoruiz21-small	menta21-small	rabinovits21-small	ikae21-small	unmasking21-small	tyo21-large	naive21-small	compressor21-small	futrzynski21-large	liaozhihao21-small
boenninghoff21-large	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
embarcaderoruiz21-large		*	=	***	**	***	***	***	***	***	***	***	***	***	***
weerasinghe21-large			***	***	=	***	***	***	***	***	***	***	***	***	***
weerasinghe21-small				***	***	***	***	***	***	***	***	***	***	***	***
menta21-large					***	***	***	***	***	***	***	***	***	***	***
peng21-small						***	***	***	***	***	***	***	***	***	***
embarcaderoruiz21-small							***	***	***	***	***	***	***	***	***
menta21-small								***	***	***	***	***	***	***	***
rabinovits21-small									***	***	***	***	***	***	***
ikae21-small										***	*	***	***	***	***
unmasking21-small											***	***	***	***	***
tyo21-large												***	***	***	***
naive21-small													=	***	***
compressor21-small														***	***
futrzynski21-large															***

submissions. In Table 2, we show a pairwise comparison of all combinations of systems to assess whether their solutions are significantly different from each other (based on their F_1 scores). The statistical procedure we applied for this is the approximate randomization test [31], for 10,000 bootstraps per comparison.

The top-performing system this year was contributed by the participant who submitted last year's strongest system. Team boenninghof21 achieved an exceptionally solid and robust performance, including the overall highest score across *all* evaluation metrics. The team in first place is followed by a tight cohort of strong runner-ups (embarcaderoruiz21, weerasinghe21, menta21, and peng21) who all achieved similar scores in the same ballpark. With the exception of three systems (tyo21, futrzynski21, liaozhihao21), most approaches significantly outperformed the three (unoptimized) baselines. The baselines themselves all yielded surprisingly similar performances, with unmasking21 being the best-performing baseline with a slight edge. Somewhat surprisingly, the system by tyo21 turned out to not be significantly different from the unmasking baseline, although it was based on a completely different verification approach.

For most systems, the pairwise F_1 scores significantly differ (Table 2), though in the upper echelons we see a few exceptions. This is to be expected with such exceptional (and hence necessarily similar) performances. The top-performing approach did, in fact, produce a significantly different solution from the runner-up, though the same is not true for all systems in the next cohort, which indicates that their particular ranking order does not necessarily indicate their quality, but incorporates a certain amount of chance. Some participants did well for some

scoring metrics, but showed a more pronounced drop in others. The system by ikae21 for instance, achieved a more than respectable AUC in the lower nineties, but an $F_{0.5u}$ only in the lower seventies (which should primarily be attributed to the different treatment of same-author pairs by this metric). Overall, the non-responses played an important part in the rankings, primarily affecting the c@1 and $F_{0.5u}$ scores. Systems such as liaozhihao21-small, that delivered binary answers without any non-responses were at a clear disadvantage in this regard.

Of particular importance is the observation that if teams submitted separate systems for the large and the small dataset, they invariably yielded significantly different solutions. Most importantly, the "large" variant always outperformed the "small" one. It should be emphasized that last year, the stronger performance of the large systems might could have been attributed to the closed-set scenario, in which a sufficiently complex model could have fully memorized each author's individual characteristics. This effect cannot serve as an explanation in this year's edition, because *none* of the test set authors or fandoms were present in the calibration materials. The performance improvements this year must therefore be attributed to the mere scope or size of the dataset or other characteristics not pertaining directly to the individual authors. This serves as additional evidence that systems were generally able to benefit from the increased training dataset size and could capitalize on accessing more abundant and more diverse material by more authors, *even in an open-set verification scenario*. It also signals clearly that the supposed ad-hoc nature of authorship identification should not be over-estimated. At least within a single textual domain, the results demonstrate the feasibility of modeling authorship quite reliably and at a large scale.

7. Discussion

In this section, we provide a more in-depth analysis of the submitted approaches and their evaluation results, also in comparison with last year's task. First, we take a look at the distribution of the submitted verification scores, including a meta classifier. We go on to inspect the effect of non-responses, and finally try to analyze how topic similarities between texts in a pair might have affected the results.

7.1. Comparison 2020–2021

Due to the intricate similarities and differences between the 2020 and 2021 editions of the task, a more detailed comparison is worthwhile. A clear advantage of the software submission procedure through `tira.io` is that we were able to rerun the systems from one year on the test dataset of another year in most cases. This way we were able to perform a cross-evaluation of quite a few systems with some exceptions due to unresolvable failures when running systems on datasets which they were not designed for. These were mostly a result of hard-coded assumptions that were violated by the new data. For example, several 2020 systems assumed all fandoms in the test set to be known, which was in clear violation with the 2021 dataset design. In Table 3, we present the performance of these system and data combinations in terms of c@1. This comparison is necessarily incomplete but allows us to glean some interesting trends. Across systems, the scores for the 2020 dataset are consistently lower than for 2021 in all instances but one (ikae). We must therefore draw the counter-intuitive conclusion that

Table 3
Cross-comparison of the performances (in terms of c@1) across different combinations of submissions (2020 vs. 2021) and test datasets (also 2020 vs. 2021). Some combinations could not be evaluated due to failures when running the system on a dataset it was not designed for.

Team	2020 System		2021 System	
	2020 Data	2021 Data	2020 Data	2021 Data
niven	0.786	–	–	–
araujo	0.770	0.81	–	–
boenninghoff	0.928	–	0.917	0.950
weerasinghe	0.880	0.913	0.885	0.917
ordonez	0.640	–	–	–
faber	0.331	–	–	–
ikae	0.544	0.503	0.742	0.758
kipnis	0.801	0.815	–	–
gagala	0.786	0.804	–	–
halvani	0.796	0.822	–	–
embarcaderoruiz	–	–	0.914	0.930
menta	–	–	0.878	0.902
peng	–	–	–	0.917
rabinovits	–	–	0.795	0.812
tyo	–	–	–	0.759
futrzynski	–	–	0.662	0.663
liaozhihao	–	–	–	0.496

the open-set formulation with unseen authors and topical domains was, in fact, easier to solve than the closed setting. On the other hand, the new 2021 systems tended to underperform on the 2020 dataset in comparison with the original 2020 submission by the same team—at least in the preciously rare cases in which we were able to make this comparison (i.e., boeninghoff, weerasinghe, ikea).

7.1.1. Distributions

Figure 7.1.1 (left) visualizes the overall distribution of the submitted answers for the systems that outperformed the baselines (best-performing system per team). We see a clear trimodal distribution with peaks around 0, 0.5 and 1, respectively. We noticed that systems submitted "bolder" answers than last year, i.e., only few answers lie in between the three peaks. The middle peak around 0.5 leads to the assumption that some systems deliberately optimized for non-responses. This assumption is further supported by Figure 7.1.1 (right), which shows the same observation, but broken down by individual systems.

In Figure 2, we plot the precision-recall curves for the above-mentioned submissions, including that of a naive meta classifier that predicts the mean score over all systems (dotted line). Whereas in previous years, the meta-classifier often suffered from a lack of methodological diversity in participant systems, this year, the mean verification score outperforms most individual systems. Nevertheless, while the meta classifier can compete with boenninghoff21 in terms of precision, it clearly falls short with regards to recall.[4]

[4]Meta classifier performance: AUC: 0.917, c@1: 0.917, F_1: 0.916, $F_{0.5u}$: 0.919, Brier: 0.917, Overall: 0.917.

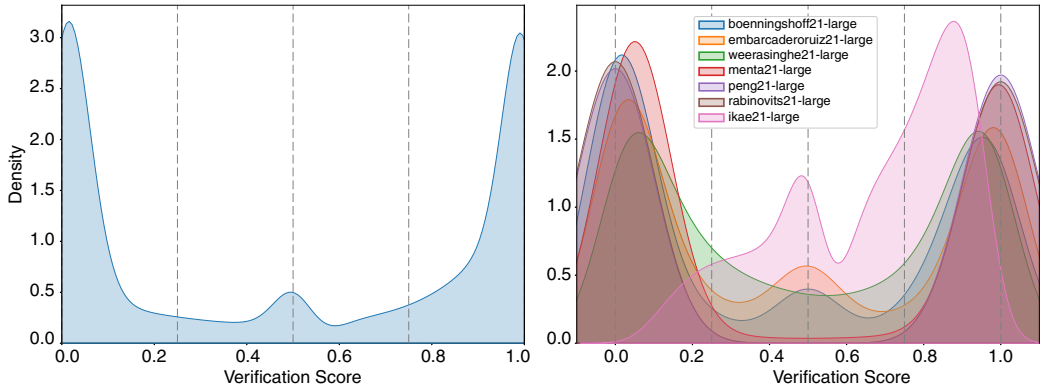

Figure 1: *Left:* Kernel density estimate across all answer scores submitted. Limited to highest ranking system per team which outperformed the baselines. *Right:* Same as left plot, but broken down by system.

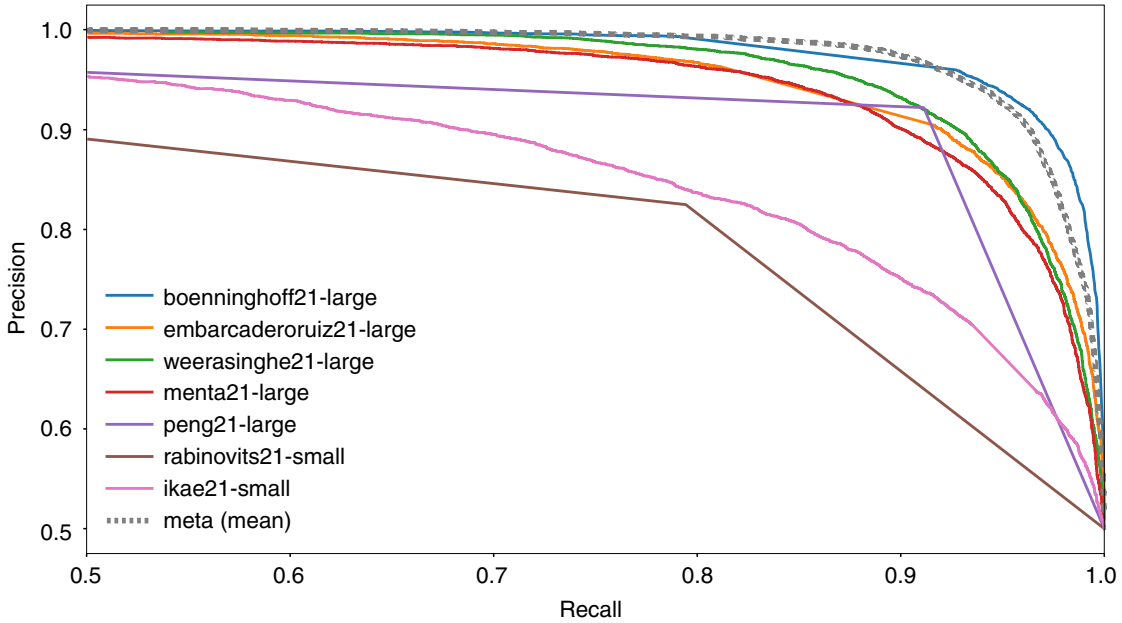

Figure 2: Precision-recall curves for individual systems, as well as a meta classifier that is based on the mean verification score across systems. Limited to highest ranking system per team which outperformed the baselines.

7.1.2. Non-answers

Non-answers were an integral aspect of the evaluation procedure. In the submitted scores, but also in the participants' notebooks, we observed that particularly returning participants,

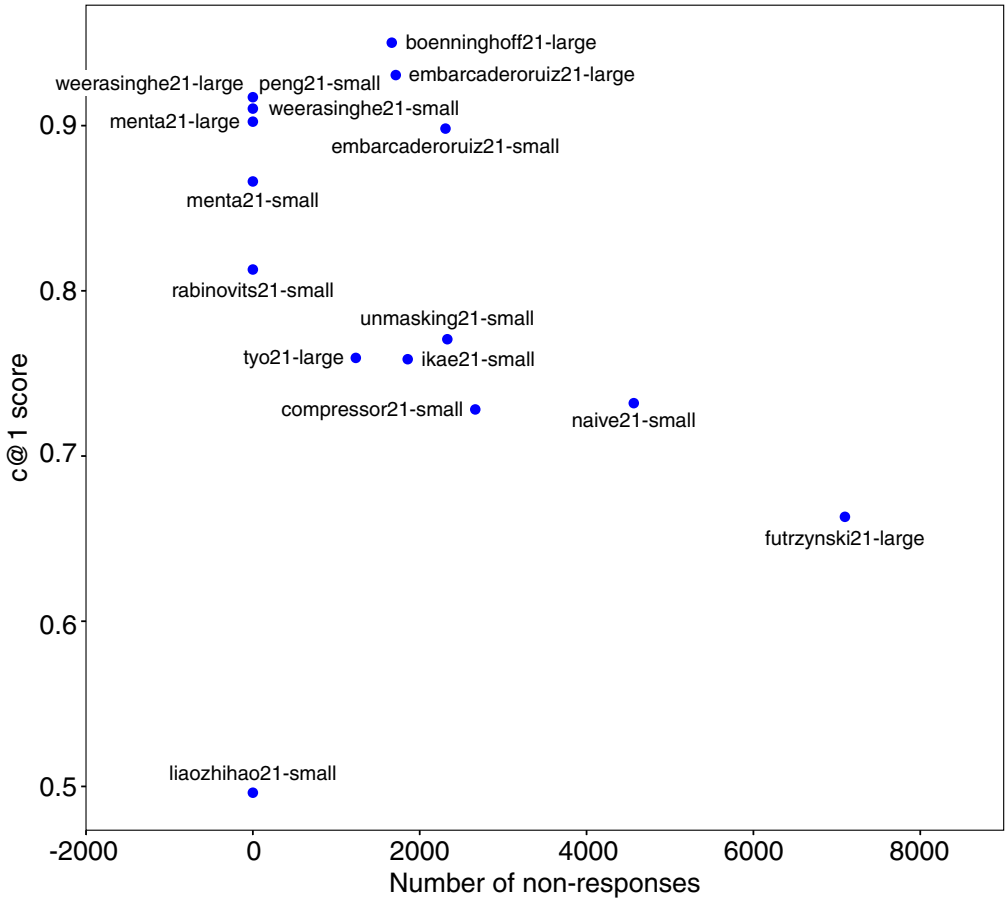

Figure 3: The c@1 scores per system as a function of the absolute number of non-answers.

such as boenninghoff21 and ikae21 took greater care to fine-tune this aspect of their systems (and were indeed successful in doing so). The different systems used non-responses to varying degrees. In Figure 3, we plot the c@1 performance as a function of the absolute number of non-responses per system. We see that futrzynski21 returned overall the most non-responses, though at the cost of a below-baseline performance. The three baselines, too, gave non-answers in comparably many cases, but were convincingly outperformed by most participant systems. The top-performing systems (boenninghoff21, embarcaderoruiz21) refused to answer cases to a more moderate degree, resulting in an overall very good performance. Many of the other high-ranking systems, such as weerasinghe21, menta21, or peng21, appeared as if they did not pay particular attention to optimizing for this aspect of the task and submitted only very few non-answers, if any. We performed a paired, non-parametric Wilcoxon signed-rank test ($n = 16$) to assess whether the number of non-responses of a system (including the baselines) correlated positively with its c@1 score. The result ($W = 28.0; p = 0.019$) offers some ground to accept this positive correlation and thus supports the hypothesis that it generally paid off for systems to submit non-answers for difficult cases.

Like last year, these observations raise the question as to which extent boenninghoff21's

Table 4
Evaluation results for top-performing systems (one per team), excluding any test problems for which boenninghoff21-large submitted a non-response ($a_i = 0.5$).

Team	**Dataset**	AUC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
boenninghoff21	large	0.991	0.957	0.952	0.976	0.963	0.968
embarcaderoruiz21	large	0.977	0.946	0.947	0.927	0.942	0.948
weerasinghe21	large	0.979	0.934	0.930	0.932	0.944	0.944
menta21	large	0.971	0.920	0.913	0.926	0.930	0.932
peng21	small	0.929	0.929	0.925	0.925	0.929	0.928

competitive edge can be attributed to the system's ability to correctly identify such difficult cases in order to leave them unanswered. Table 4 summarizes the performances of the top systems (one per participant) on all cases on which boenninghoff21 submitted a score of $a_i \neq 0.5$. Interestingly, the differences in performance stay the same, as well as the ranking, which indicates that the treatment of difficult cases is not the only magic ingredient (we should emphasize boenninghoff21's exceptional $F_{0.5u}$ score on this subset; indicating that they primarily backed off for different-author document pairs).

7.1.3. The Influence of Topic (*continued*)

In last year's overview paper, we applied a generic topic model to analyze the test problems from a semantic perspective. To avoid repetition, we will not reintroduce this model (non-negative matrix factorization with 150 dimensions applied to a TF-IDF-normalized bag-of-words representation of content words) at length, but it remains an interesting challenge to analyze this year's test data from the same topical perspective. We applied the same pipeline to this year's test data for assessing topic similarities between the document pairs, in which we calculated the cosine similarity between the L1-normalized topic vectors for each document. Overall, the topical distances over all the document pairs in both the 2020 and 2021 test sets show a very similar distribution (2020: $\mu = 0.656, \sigma = 0.147$; 2021: $\mu = 0.641, \sigma = 0.153$). This is reassuring, as it show that while both datasets are cross-fandom, the open-set vs. closed-set reformulation did not introduce any obvious topical artifacts.

Generally speaking, all of the trends reported last year also hold on this year's test set:

1. Same-author pairs displayed a higher topical similarity then different-author pairs, indicating that authors do have an inclination to write about the same topics (see Figure 4 (left)). A non-parametric (one-sided, but unpaired) Mann-Whitney U test ($n_1 = 10,000$, $n_2 = 9,999$) lends support to this view ($U > 68,687K$, $p < 0.001$).
2. There is a mild but real correlation between the topical similarity of a document pair in a test problem and the average verification score submitted by systems.
3. Results for the standard linear regression model reported last year were: $\beta = 0.16$, $R^2 = 0.15$. When limited to the correctly answered cases of the meta classifier, the resulting model this year is comparable ($\beta = 0.16$, $R^2 = 0.15$), but for the incorrect predictions, the coefficients markedly drop ($\beta = 0.09$, $R^2 = 0.01$).

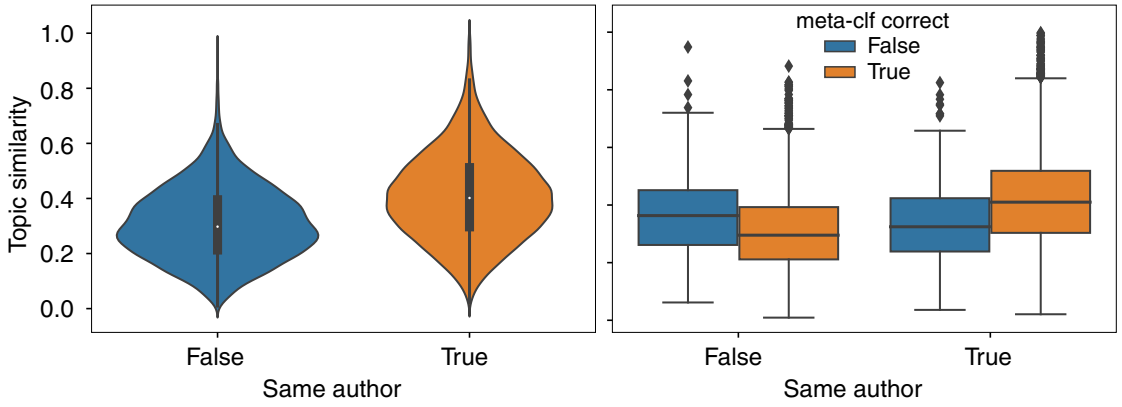

Figure 4: *Left:* Distribution of topical similarity, separate for same-author and different-author pairs. *Right:* The distribution of topical similarity within document pairs in the test set for same-author and different-author pairs broken down by whether the meta-classifier answered the pairs correctly.

All in all, we can hypothesize that this year again, the models were generally susceptible to a misleading influence of topic similarity, as indicated by Figure 4 (right): Correctly solved different-author pairs tended to be of lower topical similarity than those answered incorrectly. Same as last year, this relationship was reversed for the same-author pairs. Thus, topical information *can* be very useful for authorship verification, but it cannot necessarily be taken at face value.

8. Outlook

Last year's edition proved to be a turning point in the history of the authorship identification track at PAN: Through the release of large-scale calibration materials, the performance of authorship models could, for the first time, be benchmarked on a scale sufficient for deep representation learning. This stimulated the adoption of new neural models which produced competitive and, in some cases, outstanding results. Interestingly, this size increase did attract new participants, while at the same time, some of the regular participants from previous years found the sudden increase in data size rather intimidating and struggled in the adaptation of their pre-existing systems to the new data. To counteract this effect and to maximize the inclusivity of our initiative, the separate submission of systems trained on the small and on the large dataset variant was introduced.

Another critical change compared to previous installments was the fact that the new dataset was limited to English-language documents only, a mere result of the availability of the source material. While we assume that most systems would also generalize to other (at least European) languages, we are aware that this might be a potential source of bias and it remains to be seen to which extent exactly the results reported here will be reproducible in other (more heavily inflected) languages. Also, the effect of (potentially very many) non-native speakers of English that appear as authors in the data is hard to quantify at this time. To the best of our knowledge, very few studies have looked at authorship identification across different writing languages. One might hypothesize that authors, when active in their native language, will demonstrate

greater mastery and diversity of style, while in a second language, less refined writing and typical errors might increase their identifiability. Another deserving field for future studies is the comparison of fanfiction material that was exclusively written by authors who self-identify as (non-professional) "fans"—hence received very little (if any) moderation or editing—to writing samples by professional authors.

In spite of these critical remarks, the central take-away message from this year's shared task remains positive: Modern, large-scale authorship verification systems can perform extremely well within the fanfiction domain. Contrary to our expectations, recasting last year's task as an open-set setup did not degrade, but in fact improve their performance. Most systems were more than capable to accurately answer the cases, even though none of the authors and fandoms were seen in the training data. This is highly encouraging, though it remains to be seen whether this holds true for other textual domains outside of transformative fiction. In light of the outstanding results, we should certainly raise the uncomfortable question of whether cross-domain authorship verification in the fanfiction domain is simply too easy. Perhaps the variance between different fandoms is limited (e.g., due to a focus on erotic and pornographic content [32]) and should thus not be taken as a proxy for domain differences in other text varieties. Nevertheless, the findings demonstrate that the issue of the ad-hoc nature of authorship identification can be overcome, at least within a single textual domain, which is certainly a positive and encouraging message.

Acknowledgements

As in previous years, this initiative would not have been possible without the generous contributions of the participating teams, whose patience and enthusiasm we wish to acknowledge in what has been an unusually trying edition. Our thanks also go to the CLEF organizers for the continuation of their hard annual work. Finally, we would like to extend our appreciation to Sebastian Bischoff, Niklas Deckers, Marcel Schliebs, and Ben Thies for assembling the fanfiction.net corpus.

References

[1] H. van Halteren, H. Baayen, F. Tweedie, M. Haverkort, A. Neijt, New machine learning methods demonstrate the existence of a human stylome, Journal of Quantitative Linguistics 12 (2005) 65–77. doi:10.1080/09296170500055350.

[2] E. Stamatatos, A survey of modern authorship attribution methods, JASIST 60 (2009) 538–556. URL: https://doi.org/10.1002/asi.21001. doi:10.1002/asi.21001.

[3] P. Juola, Authorship attribution, Foundations and Trends in Information Retrieval 1 (2006) 233–334.

[4] M. Koppel, J. Schler, S. Argamon, Computational methods in authorship attribution, Journal of the American Society for Information Science and Technology 60 (2009) 9–26.

[5] M. Potthast, S. Braun, T. Buz, F. Duffhauss, F. Friedrich, J. M. Gülzow, J. Köhler, W. Lötzsch, F. Müller, M. E. Müller, R. Paßmann, B. Reinke, L. Rettenmeier, T. Rometsch, T. Sommer, M. Träger, S. Wilhelm, B. Stein, E. Stamatatos, M. Hagen, Who wrote the web? revisiting influential author identification research applicable to information retrieval, in: N. Ferro, F. Crestani, M.-F. Moens, J. Mothe, F. Silvestri, G. M. Di Nunzio, C. Hauff, G. Silvello (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2016, pp. 393–407.

[6] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Shared tasks on authorship analysis at PAN 2020, in: J. M. Jose, E. Yilmaz, J. Magalhães, P. Castells, N. Ferro, M. J. Silva, F. Martins (Eds.), Advances in Information Retrieval, Springer International Publishing, Cham, 2020, pp. 508–516.

[7] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the cross-domain authorship verification task at PAN 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_264.pdf.

[8] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection, in: Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al., 2018, pp. 1–25.

[9] M. Kestemont, E. Stamatatos, E. Manjavacas, W. Daelemans, M. Potthast, B. Stein, Overview of the Cross-domain Authorship Attribution Task at PAN 2019, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/.

[10] K. Hellekson, K. Busse (Eds.), The Fan Fiction Studies Reader, University of Iowa Press, 2014.

[11] J. Fathallah, Fanfiction and the Author. How FanFic Changes Popular Cultural Texts, Amsterdam University Press, 2017.

[12] G. W. Brier, et al., Verification of forecasts expressed in terms of probability, Monthly weather review 78 (1950) 1–3.

[13] A. Peñas, A. Rodrigo, A simple measure to assess non-response, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11, Association for Computational Linguistics, USA, 2011, p. 1415–1424.

[14] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 654–659. URL: https://doi.org/10.18653/v1/n19-1068. doi:10.18653/v1/n19-1068.

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

[16] O. Halvani, L. Graner, Cross-domain authorship attribution based on compression: Notebook for PAN at CLEF 2018, in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/paper_90.pdf.

[17] M. Kestemont, J. A. Stover, M. Koppel, F. Karsdorp, W. Daelemans, Authenticating the writings of julius caesar, Expert Systems with Applications 63 (2016) 86–96. URL: https://doi.org/10.1016/j.eswa.2016.06.029. doi:10.1016/j.eswa.2016.06.029.

[18] M. Koppel, J. Schler, Authorship verification as a one-class classification problem, in: C. E. Brodley (Ed.), Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004, volume 69 of *ACM International Conference*

Proceeding Series, ACM, 2004. doi:10.1145/1015330.1015448.

[19] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing Unmasking for Short Texts, in: J. Burstein, C. Doran, T. Solorio (Eds.), 14th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2019), Association for Computational Linguistics, 2019, pp. 654–659. URL: https://www.aclweb.org/anthology/N19-1068.

[20] C. Ikae, UniNE at PAN-CLEF 2021: Author verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[21] A. Menta, A. Garcia-Serrano, Authorship verification with neural networks via stylometric feature concatenation, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[22] Z. Liao, Z. Hong, Z. Li, G. Liang, Z. Mo, Z. Li, Authorship verification of language models based on Lucene architecture, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[23] J. Weerasinghe, R. Singh, R. Greenstadt, Feature vector difference based authorship verification for open world settings, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[24] B. Boenninghoff, R. M. Nickel, D. Kolossa, O2D2: Out-of-distribution detector to capture undecidable trials in authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[25] Z. Peng, L. Kong, Z. Zhang, Z. Han, X. Sun, Encoding text information by pre-trained model for authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[26] R. Futrzynski, Author classification as pre-training for pairwise authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[27] D. Embarcadero-Ruiz, H. Gómez-Adorno, I. Reyes-Hernández, A. García, A. Embarcadero-Ruiz, Graph-based Siamese network for authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[28] J. Tyo, B. Dhingra, Z. Lipton, Siamese Bert for authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[29] M. Pinzhakova, T. Yagel, J. Rabinovits, Feature similarity-based regression models for authorship verification, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[30] D. Chicco, Siamese Neural Networks: An Overview, Springer US, New York, NY, 2021, pp. 73–94. URL: https://doi.org/10.1007/978-1-0716-0826-5_3. doi:10.1007/978-1-0716-0826-5_3.

[31] E. W. Noreen, Computer-Intensive Methods for Testing Hypotheses: An Introduction, A Wiley-Interscience publication, 1989.

[32] G. Barlas, E. Stamatatos, A transfer learning approach to cross-domain authorship attribution, Evolving Systems (2021). URL: https://link.springer.com/10.1007/s12530-021-09377-2. doi:10.1007/s12530-021-09377-2.