Overview of the Authorship Verification Task at PAN 2022

Efstathios Stamatatos[1], Mike Kestemont[2], Krzysztof Kredens[3], Piotr Pezik[3], Annina Heini[3], Janek Bevendorff[4], Benno Stein[4] and Martin Potthast[5]

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

Abstract

The authorship verification task at PAN 2022 follows the experimental setup of similar shared tasks in the recent past. However, it focuses on a different, and very challenging scenario: given two texts belonging to different discourse types, the task is to determine whether they are written by the same author. Based on a new corpus in English, we provide pairs of texts using four discourse types: essays, emails, text messages, and business memos. The differences in communicative purpose, intended audience, and the level of formality render the cross-discourse-type authorship verification task very hard. We received 7 submissions and evaluated them using the TIRA integrated research architecture, along with two baseline approaches. This paper reviews the submissions and presents a detailed discussion of the evaluation results.

1. Introduction

Author identification (or authorship attribution) aims to reveal information about the individual(s) who wrote a text [1, 2]. There are several relevant tasks that emulate real-world conditions, mainly closed-set authorship attribution (where there is a finite list of candidate authors) and open-set authorship attribution (where there is a set of candidate authors but this does not necessarily include the true author(s)) [3]. The former scenario suits cases where only a short list of persons could eventually be the authors of disputed texts while the latter can be applied to cases where such lists of candidates are not available (or reliable enough). A special case of open-set attribution is *authorship verification* where there is only one candidate author [4]. Among author identification tasks, authorship attribution plays a key role since any given case can be decomposed into a series of authorship verification instances.

In authorship verification, texts of known authorship by one author are presented to a system, which is then tasked to verify whether another text has also been written by that same author [5, 6]. In its simplest form, only one text of known authorship is given [7]. In that case,

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

for a pair of texts (typically one of known authorship and another of unknown authorship), we are asked to determine whether they are written by the same author.

During the last decade, an extensive list of authorship verification methods have been proposed [4, 6, 8, 9]. In addition, several previous PAN editions included a relevant shared task [10, 11, 12, 13, 14]. The effectiveness of authorship verification approaches depends on several factors. Naturally, text length is a crucial factor and usually the effectiveness of systems deteriorates when only short or very short texts are given. Another very challenging form of the task considers cases where texts of known and unknown authorship belong to different domains. In *cross-domain authorship verification*, texts of known and unknown authorship may differ in topic (politics vs. sports), genre (review vs. essay) or even language (English vs. German). In PAN 2015, Both cross-topic and cross-genre authorship verification were considered, and results were with relatively low accuracy were obtained, especially for a cross-genre dataset of essays and reviews in Dutch [12]. In the last two editions of PAN [13, 14] *fanfiction* texts (i.e., non-professional fiction published online by fan authors) belonging to different fandoms (i.e., fanfiction inspired by certain highly popular works) were used. A large training dataset of more than 350,000 verification instances was compiled for this task that enabled the application of powerful deep learning models [15]. Perhaps surprisingly, the best results obtained were rather high, suggesting that most fanfiction authors may retain their stylistic choices over different fandoms, albeit other factors that may have artificially boosted the results could not be ruled out.

The current edition of PAN focuses on *cross-discourse type authorship verification* where texts of known and unknown authorship belong to different discourse types. In particular, these discourse types have significant differences concerning communicative purpose, intended audience, or level of formality. For example, the discourse types of argumentative essays and text messages sent to family members have important stylistic differences imposed by the norms of discourse types. It is therefore very challenging to distinguish authorial characteristics that remain intact across discourse types. In addition, discourse types strongly correlates with text length (e.g., essays are much longer than text messages) and cross-discourse type authorship verification can also be used to study the effect of text length in the effectiveness of authorship verification approaches approaches.

In this paper, we first present the new datasets and the evaluation framework for the cross-discourse type authorship verification shared task at PAN 2022. Next, we shall survey the received submissions and evaluate in detail their effectiveness. Finally, we discuss the main conclusions and possible directions for future work.

2. The PAN Cross-Discourse Type Authorship Verification Corpus 2022

A novel dataset was created from a subset of the recent Aston 100 Idiolects Corpus in English (Kredens, Heini and Pezik 2021),[1] including a rich set of discourse types authored by 112 individuals. We used the following discourse types of written language: emails, essays, text messages, and business memos. All individuals represented in the corpus have a similar age (18–22) and are native speakers of English. The topic of text samples is not restricted, while the level of

[1] https://fold.aston.ac.uk/handle/123456789/17

Table 1
Key statistics of the new dataset for 2022 cross-discourse type authorship verification task.

Subset	Training	Test
Author match	*Text pairs*	
Positive (same author)	6,132 (50.0%)	5,239 (50.0%)
Negative (different author)	6,132 (50.0%)	5,239 (50.0%)
Discourse type pairings	*Text pairs*	
Email–Text message	7,484 (61.0%)	6,092 (58.1%)
Essay–Email	1,618 (13.2%)	1,454 (13.9%)
Essay–Text message	1,182 (9.6%)	1,128 (10.8%)
Business memo–Email	1,014 (8.3%)	900 (8.6%)
Business memo–Text message	780 (6.4%)	718 (6.9%)
Essay–Business memo	186 (1.5%)	186 (1.8%)
Discourse type	*Text length (avg. chars)*	
Essay	11,098	10,117
Email	2,385	2,323
Business memo	1,255	1,042
Text message	611	601

formality can vary within a certain discourse type (e.g., text messages may be addressed to family members or other acquaintances). Table 1 gives an overview of the data and the parts of it used of training and testing different aspects of cross-discourse type authorship verification.

This corpus has been anonymized in that named-entities such as mentions of locations, person names, addresses, etc. were manually replaced with generic placeholder tags. This is very useful for evaluating authorship verification methods in cross-discourse type scenarios since the presence of author-specific and topic-specific information is reduced.

In order to compile the required training and test datasets for the shared task at hand, the corpus needed further preprocessing. First, we split the available individuals into two equal and non-overlapping sets, one to be used for the training dataset and the other for the test dataset. That way, it is ensured that any kind of particularities among the training authors will not affect the effectiveness on the test dataset. In addition, we took advantage of the demographic metadata available and ensured a stable gender distribution of individuals in both the training and test dataset. More specifically, the training and test datasets represent writings by 56 authors each (10 male, 45 female and 1 of unidentified gender).

The dataset comprises a set of text pairs and in each pair the two texts belong to two different discourse types. All six combinations of the four available discourse types are taken into account. However, the distribution of text pairs over the combination of discourse types is not homogeneous since it depends on the available texts belonging to each discourse type. For example, the corpus comprises only one business memo and multiple email messages per individual. Nevertheless, the distribution of verification instances per discourse type combination is similar in both training and test datasets as can be seen in Table 1. Similarly, both training and test datasets have a balanced distribution of positive/negative verification cases. This is also valid for each combination of discourse types (e.g., half of the pairs belonging to the combination essay–email is positive and the other half is negative).

Since the length of texts belonging to certain discourse types can be limited, we concatenated multiple texts of the same discourse type to produce longer text samples. In more detail, email messages were concatenated so that a text sample of at least 2,000 characters was obtained. The date of email messages was taken into account so that consecutive messages are concatenated. In the case of text messages, we concatenated messages sent either to friends or to family, so that text samples of at least 500 characters were obtained. We inserted the special tag <new> in the concatenated messages to indicate the original message boundaries. The text lengths in Table 1 for email and text messages refer to text samples created in this manner.

3. Evaluating Cross-Discourse Type Authorship Verification

In authorship verification, one has to approximate the target function $\phi : (D_k, d_u) \rightarrow \{T, F\}$, where D_k is a set of texts of known authorship and d_u is a text of unknown or disputed authorship. In the current edition of the task, we consider D_k as singleton. Thus, the task is to approximate the target function $\phi : (d_k, d_u) \rightarrow \{T, F\}$ for a pair of texts. If $\phi(d_k, d_u) = T$, then the author of d_k is also the author of d_u (positive instance) and if $\phi(d_k, d_u) = F$, then the author of d_k is not the same as the author of d_u (negative instance). The main novelty of the current edition is that d_k and d_u belong to different discourse types.

The evaluation framework is similar to the one used in recent shared tasks at PAN [14]. For each authorship verification instance (a pair of texts) of the test dataset, participants have to produce a scalar score a_i (in the $[0, 1]$ range) indicating the probability that the pair was written by the same author. It is possible for participants to leave text pairs unanswered by submitting a score of precisely $a_i = 0.5$.

3.1. Evaluation Measures

Similar to recent editions of the authorship verification task [14], we adopt a diverse set of effectiveness measures to highlight different aspects of the capabilities of an authorship verification model. We reused the four measures from the 2020 edition, but also included the Brier score [16] as an additional fifth measure (following discussions with participants and the audience at the 2020 workshop). In total, the following effectiveness measures were used:

- AUROC: the area under the ROC curve,
- c@1: a variant of the conventional accuracy measure, which rewards systems that leave difficult problems unanswered [17],
- F_1: the well-known F_1 effectiveness 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 [18],
- Brier: the complement of the Brier loss function [16] focusing on the accuracy of probabilistic predictions (as implemented in sklearn [19]). This measure rewards verifiers that make "bold" but correct predictions (i.e., a_i close to 0.0 or 1.0) and it indirectly penalizes less confident ones, including non-answers ($a_i = 0.5$). In line with the other measures we take its complement so that higher scores correspond to better effectiveness.
- The average of the above measures is used as final score to rank submitted systems.

We also report runtime on TIRA to give an indication of relative efficiency.

3.2. Baselines

In order to facilitate the comparison of the submitted methods with established approaches from the literature in the field, we provide two baseline methods that are based on character n-grams or character sequences. The source code of the following two methods were made available to the participants at the start of the campaign (together with an official implementation of the evaluation measures):

- Compression-based model. Given a pair of texts t_1 and t_2, the cross-entropy of t_2 based on the Prediction by Partial Matching (PPM) model of t_1 is computed, and vice-versa [20]. Then, a logistic regression classifier is trained using the mean and absolute difference of the two cross-entropies. In addition, using a small radius verification scores around 0.5 are set to exactly 0.5.
- Distance-based character n-gram model [21]. The most frequent character 4-grams are extracted from the training texts and used to represent each text. Then, given a pair of texts, the cosine similarity between them is calculated. During training, two threshold values p_1 and p_2 are optimized to scale the verification scores. All verification scores lower than p_1 correspond to negative answers, all scores greater than p_2 are scaled to positive answers and the remaining scores are set to 0.5, implying that these are hard instances that deliberately are left unanswered.

The baselines are not tailored to particular discourse types, e.g., by tuning hyperparameters.

4. Survey of Submissions

We received seven submissions and evaluated their effectiveness and efficiency using the TIRA integrated research architecture [22]. All participants also submitted a notebook paper describing their approach. The main characteristics of each approach are provided in Table 2.

Most participants followed the recent trend in natural language processing and used pre-trained language models like BERT, T5, or MPNET to obtain text embeddings. Konstantinou et al. [23] report that several such models were compared and the most effective one selected. Approaches not using pre-trained language models exploit graph-based text representations [24], spectral analysis [25], or representations based on traditional feature engineering including features like frequencies of part-of-speech (POS) tags and word unigrams (najafi22).

Regarding the classification model, most participants rely on fully-connected layers that combine the information from the text representation step. It is also reported that several traditional machine learning algorithms, such as support vector machines and random forests were examined but their effectiveness was found to be comparatively low [23]. Other deep learning methods used are convolutional and siamese neural networks. Since the use of deep learning technology usually requires a considerable amount of training and some extra validation data, some participants attempted to augment the provided dataset by generating new authorship verification instances with the help of the available metadata.

Surprisingly, no participant studied discourse type-specific approaches for the given combinations despite their substantial differences.

Table 2
Review of the basic characteristics of the submitted approaches: POS, NEs, SOM, and FC denote part-of-speech tags, named entities, self-organizing maps, and fully-connected layers, respectively.

System	Ref.	Representation	Classification	Augmentation	Type-specific
CRESPOSANCHEZ22	[25]	word unigrams, doc2vec	FC	Yes	No
		(text and POS), SOM	FC	Yes	No
GALICIA22	[24]	graph-based, POS	Siamese network	Yes	No
HUANG22	[26]	BERT	FC	No	No
JINLI22	[23]	MPNET	FC	No	No
LEI22	[27]	BERT	FC	No	No
NAJAFI22	[28]	T5, word unigrams, POS,	CNN, attention	No	No
		NEs, Punctuation	block, FC	No	No
YIHUIYE22	[29]	BERT	TextCNN	Yes	No

Table 3
Evaluation results for the cross-discourse type authorship verification task, ranked by overall effectiveness. Bold font highlights best in column.

Participant	AUROC	C@1	F_1	$F_{0.5u}$	Brier	Overall
BASELINE-CNGDIST22	0.546	0.496	0.669	0.542	0.749	**0.600**
NAJAFI22	**0.598**	**0.571**	0.576	**0.571**	0.618	0.587
GALICIA22	0.512	0.499	0.628	0.544	0.741	0.585
JINLI22	0.577	0.557	0.581	0.563	0.589	0.573
BASELINE-COMPRESSOR22	0.541	0.493	0.570	0.478	**0.750**	0.566
LEI22	0.539	0.539	0.399	0.488	0.539	0.501
YIHUIYE22	0.542	0.526	0.398	0.461	0.565	0.499
HUANG22	0.519	0.519	0.196	0.328	0.519	0.416
CRESPOSANCHEZ22	0.500	0.500	0	0	0.748	0.350

5. Evaluation Results

This section presents an in-depth analysis of the effectiveness and efficiency of the submitted approaches regarding overall, dependent on discourse type, with respect to bias, runtime, and in comparison to the previous year's participants.

5.1. Overall results

Table 3 shows the overall results of all participants. In general, the effectiveness of all submissions is quite low, reflecting the difficulty of the task. The approaches of NAJAFI22, GALICIA22, and JINLI22 clearly outperform the rest of the submissions. It is also surprising that a naive baseline achieved the best overall score, despite the fact that most participant models are quite sophisticated. On the other hand, the most effective method submitted (NAJAFI22) outperforms all other submissions and baselines in three out of five evaluation measures. Its main weakness seems to be the low Brier score which means that its probabilistic predictions are in need of improvement (even if its binary class assignments are relatively strong).

Table 4
Evaluation results for the cross-discourse type authorship verification task, dependent on discourse type pairings, ranked by overall effectiveness on the entire test dataset (see Table 3). Bold font highlights best in column.

(a) Email–Text message

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
baseline-cngdist22	0.585	0.505	**0.672**	0.555	0.751	**0.614**
najafi22	**0.613**	**0.583**	0.579	**0.581**	0.628	0.597
galicia22	0.517	0.502	0.641	0.549	0.740	0.590
jinli22	0.599	0.576	0.603	**0.581**	0.607	0.593
baseline-compressor22	0.570	0.480	0.661	0.499	**0.752**	0.592
lei22	0.537	0.537	0.350	0.462	0.537	0.484
yihuiye22	0.547	0.530	0.409	0.470	0.569	0.505
huang22	0.519	0.519	0.197	0.329	0.519	0.417
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(b) Essay–Email

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
baseline-cngdist22	0.514	0.489	0.657	0.542	0.740	0.589
najafi22	0.570	0.549	0.583	0.557	0.605	0.573
galicia22	0.496	0.498	0.608	0.537	0.744	0.577
jinli22	**0.588**	0.559	0.574	0.563	0.600	0.577
baseline-compressor22	0.531	0.481	**0.659**	0.531	**0.750**	**0.590**
lei22	0.585	**0.585**	0.528	**0.575**	0.585	0.571
yihuiye22	0.548	0.529	0.371	0.449	0.558	0.491
huang22	0.553	0.553	0.283	0.444	0.553	0.477
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(c) Essay–Text message

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
baseline-cngdist22	0.540	0.493	**0.673**	0.539	0.750	**0.599**
najafi22	**0.568**	**0.553**	0.567	**0.556**	0.595	0.568
galicia22	0.513	0.493	0.604	0.534	0.743	0.577
jinli22	0.476	0.483	0.486	0.485	0.520	0.490
baseline-compressor22	0.567	0.513	0.130	0.186	**0.751**	0.429
lei22	0.519	0.519	0.299	0.412	0.519	0.453
yihuiye22	0.509	0.508	0.336	0.410	0.542	0.461
huang22	0.516	0.516	0.173	0.301	0.516	0.404
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(e) Business memo–Email

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
baseline-cngdist22	0.509	0.443	**0.67**	0.500	**0.748**	0.574
najafi22	**0.606**	**0.586**	0.612	**0.589**	0.633	**0.605**
galicia22	0.521	0.513	0.647	0.556	0.742	0.596
jinli22	0.562	0.547	0.565	0.552	0.569	0.559
baseline-compressor22	0.514	0.494	0.214	0.269	0.746	0.447
lei22	0.512	0.512	0.497	0.507	0.512	0.508
yihuiye22	0.539	0.520	0.414	0.461	0.571	0.501
huang22	0.493	0.493	0.099	0.185	0.493	0.353
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(d) Business memo–Text message

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
baseline-cngdist22	0.525	0.376	**0.673**	0.455	**0.748**	**0.555**
najafi22	**0.582**	0.534	0.515	0.526	0.589	0.549
galicia22	0.487	0.464	0.553	0.503	0.741	0.549
jinli22	0.551	0.535	0.567	**0.545**	0.566	0.553
baseline-compressor22	0.524	0.518	0.065	0.127	0.746	0.396
lei22	0.539	**0.539**	0.472	0.517	0.539	0.521
yihuiye22	0.553	0.538	0.463	0.499	0.579	0.526
huang22	0.481	0.481	0.126	0.214	0.481	0.356
cresposanchez22	0.500	0.500	0	0	0.748	0.350

(f) Essay–Business memo

Participant	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**
baseline-cngdist22	0.474	0.466	**0.647**	0.512	0.740	0.568
najafi22	**0.545**	**0.538**	0.533	0.536	0.572	0.545
galicia22	0.496	0.500	0.635	**0.547**	0.739	**0.584**
jinli22	0.528	0.500	0.542	0.516	0.532	0.524
baseline-compressor22	0.474	0.477	0.477	0.446	0.744	0.523
lei22	0.500	0.500	0.367	0.437	0.500	0.461
yihuiye22	0.522	0.492	0.214	0.302	0.545	0.415
huang22	0.527	0.527	0.290	0.415	0.527	0.457
cresposanchez22	0.500	0.500	0	0	**0.748**	0.350

5.2. Results by discourse type

Table 4 shows breaks down the results with respect to the six pairings of discourse types. Recall that each discourse type comes with different average text lengths (see Table 1). For instance, essays are much longer than the rest of the examined discourse types. As Table 4b, c, and f show, when essays are part of a pairing, the submission of galicia22 is the most effective system in terms overall effectiveness. Where essays are excluded (Table 4a, e, and d), their approach is outperformed by that of najafi22. On the shortest discourse types (business memos and text messages; Table 4d) the submission of jinli22 seems to be the most effective. This pairing of discourse type also has the lowest overall effectiveness, indicating that text length (plus cross-discourse verification) remains a crucial factor in authorship verification. The baseline-cngdist22 is relatively stable across combinations of discourse types, while baseline-compressor22 achieves its optimal results when the longest discourse types (essays

Table 5
(a) Number of positive and negative answers provided by each verification model along with the number of unanswered instances of the test dataset. (b) Runtime efficiency of the submitted approaches.

(a)

System	Positive	Negative	Unanswered
CRESPOSANCHEZ22	0	10,478	0
GALICIA22	8,874	1,604	0
HUANG22	1,031	9,447	0
JINLI22	5,820	4,658	0
LEI22	2,805	7,673	0
NAJAFI22	5,355	5,083	40
YIHUIYE22	2,841	7,116	521
BASELINE-CNGDIST22	9,199	17	1,262
BASELINE-COMPRESSOR22	3,927	3,268	3,283

(b)

System	Run time
CRESPOSANCHEZ22	00:05:36
GALICIA22	00:07:22
LEI22	06:04:59
YIHUIYE22	07:16:59
NAJAFI22	18:18:32
JINLI22	23:25.62
HUANG22	31:04:56

and emails) are considered. It practically fails, however, when only very short texts are available.

5.3. Bias

Table 5a shows the number of positive and negative answers provided by each verification model. Note that these are not necessarily correct predictions: they merely correspond to the test instances where the estimated verification score is lower/greater than 0.5. In addition, the number of test instances with a verification score equal to 0.5 is also presented—these correspond to instances left unanswered according to the definition of the task. These three numbers indicate the bias of each verification model. We reiterate that the actual distribution of positive/negative instances in the test dataset is balanced. As can be seen, very few submissions leave instances unanswered. This means that their effectiveness, especially in terms of C@1, can be significantly improved by incorporating a mechanism to exclude borderline instances from positive/negative answers, similar to the ones used by the baselines.

It is also remarkable that the approaches of NAJAFI22 and JINLI22 (along with BASELINE-COMPRESSOR22) are unbiased, providing roughly similar numbers of positive and negative answers. In contrast, the submission of GALICIA22 as well as BASELINE-CNGDIST22 are clearly biased towards positive answers, while HUANG22, LEI22, YIHUIYE22, and CRESPOSANCHEZ22 are clearly biased towards negative answers. Note that these biases do not affect AUROC measures.

5.4. Efficiency

Beyond effectiveness, another criterion for evaluating an authorship verification system is in terms of efficiency or its runtime cost. Depending on the application of specific kinds of technology, this is a significant criterion, especially when large volumes of text have to be analyzed. Table 5b shows the elapsed time of the run of each submitted method on TIRA. As can be seen, the approaches that avoid the use of pre-trained language models [25, 24] achieve the lowest runtime by a large margin. The highest runtime is required by the approach of HUANG22 that splits texts into segments and examines all combinations of segments.

Table 6
(a) Evaluation results of the top-performing models submitted to the PAN 2021 shared task on authorship verification on the cross-discourse type test data. (b) Number of positive and negative answers as well as non-answers provided by these models.

	(a)						(b)		
System	AUROC	c@1	F_1	$F_{0.5u}$	Brier	**Overall**	**Positive**	**Negative**	**Unanswered**
boenninghoff21	0.513	0.501	0.002	0.005	0.531	0.310	10	10,370	98
embarcaderoruiz21	0.538	0.502	0.063	0.116	0.581	0.360	309	9,295	874
weerasinghe21	0.488	0.500	0.011	0.027	0.506	0.306	57	10,421	0

5.5. A Transfer-learning Experiment

We applied the top-performing approaches from the previous 2021 edition of PAN [30] to the current test dataset. Thanks to software submissions at TIRA, this can be accomplished with relative ease. This amounts to a transfer-learning experiment, since the three models are trained and fine-tuned on a cross-fandom authorship verification dataset but now tested on our cross-discourse type dataset. The following methods have been employed:

- boenninghoff21 [31]: A deep learning-based approach including neural feature extraction and deep metric learning, deep Bayes factor scoring, uncertainty modeling and adaptation, a combined loss function, and an additional out-of-distribution detector for non-responses. In its final step, the model was extended to a majority-voting ensemble.
- embarcaderoruiz21 [32]: Its main idea is similar to that of galicia22. A graph-based representation approach is combined with a Siamese network.
- weerasinghe21 [33]: A variety of stylometric features, including character and POS n-grams, function words, and vocabulary richness measures and a logistic regression classifier, fed with the absolute differences of these features for each text pair.

We made no attempt to modify these methods before applying them to the new cross-discourse type test dataset.

The effectiveness of the above-mentioned methods on the PAN 2021 test data was exceptional. All of them obtained an overall score (over the same five evaluation measures used in this paper) of greater than 0.93 [30]. Table 6a shows the effectiveness of the 2021 models on the 2022 test data. Unsurprisingly, the three models perform much worse. Their overall effectiveness on the cross-discourse type dataset is very low, much lower than all but one of the seven submissions and the two baselines shown in Table 3. This means that fine-tuning such models to particular datasets hurts their generalizability. Moreover, cross-fandom verification and cross-discourse type verification have different characteristics in terms of the two available datasets.

Table 6b shows the number of positive and negative answers as well as non-answers for each of the three 2021 models, which exert a clear bias of models towards negative answers. Note that in the 2021 cross-fandom dataset, all texts have similar text length. Likely, this factor along with other substantial differences between fanfiction and the discourse types considered in the cross-discourse type dataset confuse these models (or at least that they need appropriate fine-tuning to improve the scaling of the produced verification scores). Note that the AUROC scores (which do not depend on the scaling of verification scores) are also quite low.

6. Conclusion

Previous shared tasks on authorship attribution at PAN played a crucial role to advance research in the field of authorship analysis and modern methods have been using the PAN datasets for evaluation purposes extensively and have incrementally improved the state of the art [6, 8]. Recent editions of PAN focused on fanfiction. The very good results obtained by the top-performing submissions there may have given the false impression that authorship verification is an almost solved problem [13, 14]. This is in fact not the case, as our experiment shows.

This year, we focused on a very challenging version of the authorship verification task where text pairs of different discourse types are used. When texts differ in communicative purpose, intended audience, or level of formality, it is very challenging to identify stable characteristics associated with authors across these discourse types. The effectiveness of all submissions in the cross-discourse type datasets was comparatively low, some as low as a random-guess baseline.

It is also surprising that all submissions, despite their increased level of sophistication in most of the cases, were outperformed by a naive baseline based on character n-grams and cosine similarity (at least according to the overall effectiveness across all five evaluation measures). This suggests that traditional methods based on well-known stylometric features could still be more effective than deep learning approaches using modern pre-trained language models for this challenging task. Another factor is the volume of data available for training (roughly, 12,000 instances) that can be considered too little for deep learning-based approaches.

Another crucial issue is text length. It seems that when the relatively long essays were used as inputs, the graph-based approach of GALICIA22 was more effective. When shorter texts from discourse types like emails, business memos, and text messages were used, the pre-trained language-model-based approaches of NAJAFI22 and JINLI22 were more effective.

The overall low effectiveness achieved shows that there is a lot of room for improvement in cross-discourse type authorship verification. All submitted approaches adopted a unified model that predicts authorship disregarding combinations of discourse types. Having separate models for each combination of discourse types is an obvious next step. This would mean, however, that the training data should also be split into smaller parts based on the combinations of discourse types. An ensemble method combining traditional stylometric models and pre-trained language models appears like a promising approach in this regard.

References

[1] 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.

[2] 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.

[3] M. Koppel, J. Schler, S. Argamon, Authorship attribution in the wild, Language Resources and Evaluation 45 (2011) 83–94. doi:10.1007/s10579-009-9111-2.

[4] E. Stamatatos, Authorship verification: A review of recent advances, Research in Computing Science 123 (2016) 9–25.

[5] O. Halvani, L. Graner, R. Regev, Taveer: an interpretable topic-agnostic authorship verification method, in: M. Volkamer, C. Wressnegger (Eds.), ARES 2020: The 15th International Conference on Availability, Reliability and Security, ACM, 2020, pp. 41:1–41:10.

[6] N. Potha, E. Stamatatos, Improving author verification based on topic modeling, Journal of the Association for Information Science and Technology 70 (2019) 1074–1088. doi:https://doi.org/10.1002/asi.24183.
[7] M. Koppel, Y. Winter, Determining if two documents are written by the same author, Journal of the Association for Information Science and Technology 65 (2014) 178–187.
[8] S. Ding, B. Fung, F. Iqbal, W. Cheung, Learning stylometric representations for authorship analysis, IEEE Transactions on Cybernetics 49 (2019) 107–121.
[9] B. Boenninghoff, R. M. Nickel, S. Zeiler, D. Kolossa, Similarity learning for authorship verification in social media, in: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2457–2461. doi:10.1109/ICASSP.2019.8683405.
[10] T. Gollub, M. Potthast, A. Beyer, M. Busse, F. M. R. Pardo, P. Rosso, E. Stamatatos, B. Stein, Recent trends in digital text forensics and its evaluation - plagiarism detection, author identification, and author profiling, in: P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Visualization - 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, volume 8138 of *Lecture Notes in Computer Science*, Springer, 2013, pp. 282–302.
[11] M. Potthast, T. Gollub, F. M. R. Pardo, P. Rosso, E. Stamatatos, B. Stein, Improving the reproducibility of pan's shared tasks: - plagiarism detection, author identification, and author profiling, in: E. Kanoulas, M. Lupu, P. D. Clough, M. Sanderson, M. M. Hall, A. Hanbury, E. G. Toms (Eds.), Information Access Evaluation. Multilinguality, Multimodality, and Interaction - 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings, volume 8685 of *Lecture Notes in Computer Science*, Springer, 2014, pp. 268–299.
[12] E. Stamatatos, M. Potthast, F. M. R. Pardo, P. Rosso, B. Stein, Overview of the PAN/CLEF 2015 evaluation lab, in: J. Mothe, J. Savoy, J. Kamps, K. Pinel-Sauvagnat, G. J. F. Jones, E. SanJuan, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 6th International Conference of the CLEF Association, CLEF 2015, Toulouse, France, September 8-11, 2015, Proceedings, volume 9283 of *Lecture Notes in Computer Science*, Springer, 2015, pp. 518–538.
[13] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. M. R. Pardo, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of PAN 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, volume 12260 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 372–383.
[14] J. Bevendorff, B. Chulvi, G. L. D. la Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 419–431.
[15] S. Bischoff, N. Deckers, M. Schliebs, B. Thies, M. Hagen, E. Stamatatos, B. Stein, M. Potthast, The Importance of Suppressing Domain Style in Authorship Analysis, CoRR abs/2005.14714 (2020). URL: https://arxiv.org/abs/2005.14714.
[16] G. W. Brier, et al., Verification of forecasts expressed in terms of probability, Monthly weather review 78 (1950) 1–3.

[17] 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.
[18] 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.
[19] 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.
[20] W. J. Teahan, D. J. Harper, Using Compression-Based Language Models for Text Categorization, Springer Netherlands, Dordrecht, 2003, pp. 141–165. URL: https://doi.org/10.1007/978-94-017-0171-6_7. doi:10.1007/978-94-017-0171-6_7.
[21] M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, W. Daelemans, Authenticating the writings of julius caesar, Expert Systems with Applications 63 (2016) 86–96.
[22] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA integrated research architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, volume 41 of *The Information Retrieval Series*, Springer, 2019, pp. 123–160. URL: https://doi.org/10.1007/978-3-030-22948-1_5. doi:10.1007/978-3-030-22948-1_5.
[23] S. Konstantinou, A. Zinonos, J. Li, Different Encoding Approaches for Authorship Verification, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[24] J. A. Martinez-Galicia, D. Embarcadero-Ruiz, A. R.-O. na, H. Gómez-Adorno, Graph-Based Siamese Network for Authorship Verification, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[25] M. Crespo-Sanchez, H. Gómez-Adorno, I. Lopez-Arevalo, E. Aldana-Bobadilla, K. Salas-Jimenez, J. Cortes-Lopez, A Content Spectral-based Analysis for Authorship Verification, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[26] M. Huang, L. Kong, Z. Peng, Y. Ye, Z. Li, X. Jiang, Z. Han, Authorship verification Based On Fully Interacted Text Segments, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[27] Z. Lei, H. Qi, H. Y, Z. Peng, M. Huang, Application of BERT in Author Verification Task, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[28] M. Najafi, E. Tavan, Text-to-Text Transformer in Authorship Verification Via Stylistic and Semantical Analysis, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[29] Y. Ye, H. Y, Z. Peng, M. Huang, L. Kong, Z. Han, Authorship Verification Using Convolutional Neural Network, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.
[30] M. Kestemont, E. Stamatatos, E. Manjavacas, J. Bevendorff, M. Potthast, B. Stein, Overview of the Authorship Verification Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.
[31] 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.
[32] 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.
[33] 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.