Overview of the Authorship Verification Task at PAN 2023

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

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

Abstract

The authorship verification task for PAN 2023 focuses on a very challenging scenario: given a pair of texts belonging to different discourse types, the task is to determine whether they were authored by the same person. In addition, for the first time, we consider discourse types from both written (i.e., essays and emails) and spoken language (i.e., interviews and speech transcriptions). New datasets in English are provided and we adopt the same evaluation setup and measures as similar tasks in recent editions of PAN. A total of eleven teams submitted 27 runs and were evaluated along with several baselines on the TIRA experimental platform. This paper includes a review of the submitted methods and a detailed discussion of the evaluation results.

1. Introduction

There are many cases where the authorship of a text is disputed. These include literary works published anonymously or under a pseudonym, threats on social media, phishing emails, plagiarism in academic papers, etc. [1, 2, 3]. Authorship analysis is based on text analysis to determine information about the author of a particular text. The basic idea is that the personal writing style of authors can be distinguished using appropriate stylometric methods to represent documents [4]. Provided that a set of candidate authors is available, it is possible to define tasks for closed-set or open-set authorship attribution tasks [5].

A more fundamental task in authorship analysis is *authorship verification*, as it addresses the basic question of whether a particular person authored a text of disputed authorship [6, 7, 8]. It is easy to see that any case of authorship attribution can be split into a number of verification instances (i.e., one for each candidate author). Generally, a set of texts known to be written by the author in question is given, and the task aims at identifying stylistic similarities/differences between these texts and the disputed text [9]. In the simplest form of the authorship verification task, there is only one text of known authorship, so this case may be better described as determining whether a pair of texts was authored by the same person [10].

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece

Several factors can influence the effectiveness of an authorship verification method. Certainly, text length is one of them, because it is very difficult to adequately represent the stylistic features of very short documents. Conversely, the limited input length of modern pre-trained language models can make it difficult to deal with long documents [11]. In addition, topic similarities/differences between the documents involved can be misleading (e.g., two documents may appear similar because of a common topic and not because of their writing style) [12]. In cases where the documents of known and unknown authorship belong to different genres or discourse types (e.g., essay vs. e-mail), it is very difficult to focus on the authors' writing style characteristics that are preserved when writing in different degrees of formality or for different audiences. Such cases are not unrealistic, however, since it is not always possible to find undisputed texts written by a particular person in a particular type of discourse.

PAN has included a task on authorship verification in several previous editions which has increased the attention of the international research community to this task [13, 14, 15, 16, 17, 18]. The focus of several editions has been on cross-domain authorship verification, where documents with known and unknown authorship come from different domains (e.g., topic areas or genres) [15, 16, 17]. Recent PAN editions have focused on fanfiction texts (i.e., non-professional fiction published online by fans of well-known works), with documents of known and unknown authorship belonging to various fandoms (e.g., Harry Potter, Sherlock Holmes). The abundance of online fanfiction documents covering multiple fandoms enabled the creation of large datasets. The results obtained show that authorship verification in fanfiction can be performed with relatively high accuracy [16, 17].

In the previous edition of PAN, we considered a more challenging scenario, namely *cross-discourse type authorship verification*, where documents of known and unknown authorship belong to different discourse types (i.e., essays, emails, text messages, and business memos) [19]. Moreover, the nature of discourse also affects the text length of documents (e.g., essays are much longer than text messages). The obtained results confirm that it is extremely difficult to identify the features of writing style related to the author's personal style in such a variety of discourse types.

In the current edition of PAN, we continue to focus on cross-discourse type authorship verification of document pairs. Unlike previous versions of the task, which used only written language discourse types, the main innovation is that we also consider spoken language. This provides an opportunity to examine the ability of authorship verification methods to deal with the completely different expressions of written and oral language. Most forms of written language are more formal, have a larger vocabulary, and are syntactically more complex than oral language. Oral language, by contrast, is mainly conversational and relatively spontaneous [20]. It is therefore very difficult to identify the stylistic features of one and the same person who expresses themselves both in writing and orally.

The remainder of this paper is organized as follows. The next section describes the new datasets available for training and evaluating authorship verification methods under cross-discourse conditions. We then define the framework for evaluating the shared task for authorship verification on PAN'23. Then, the submissions received are examined and their effectiveness and efficiency are analytically evaluated. Finally, we discuss the main conclusions and possible directions for future work.

Table 1
Two pairs of text samples from the dataset written by the same author in two discourse types.

Interview	**Email**
Okay, erm, <cough>, so on a day when I'm going to the gym, I'll decide to have porridge. So, I'll have, erm, they're called, erm, golden syrup... I don't even know what the, it's like golden and it's just like in a packet and, basically, it comes in like a sachet. And you add milk to it and then you microwave it. And it's just like a really easy way to eat porridge. And I normally have that with, erm, cinnamon powder, erm, peanut butter and some honey. And then if you just mix that all together, sometimes, I have blueberries on there, sometimes I'll have almonds, erm, anything like that. Erm, and then on a normal day, I'll choose to make pancakes, erm, which is basically my favourite breakfast. Like I don't know why but I'm just so obsessed with eating pancakes.	Hi <addr6_FN>,<nl><nl>Thank you for clearing that up! I look forward to seeing you tomorrow.<nl><nl>Many thanks,<nl><nl><part_FN><new>Hi <addr6_FN>, <nl><nl>I hope you're well! Apologies for the late email. Just to double check, we are meeting tomorrow at 12pm? Also am I right in assuming we that we will be meeting at <location>, <house_number><road>, <town>, <city>, <postcode>? The same centre in which tommy interview took place? <nl><nl>Many Thanks, <nl><nl><part_FN><new>Hi <addr5_FN>, <nl><nl><nl>Am I right in assuming that I can appeal within 14 days once receiving my transcript? Would I be able to contact the Welfare Advisor to help me with the appeals process if I require it? <nl><nl><nl>Many thanks, <nl><nl><nl><part_FN>
Essay	**Speech transcription**
Subjective expected utility (SEU) theory is normative theory of a decision making according to which a decision maker chooses an alternative or strategy, the key concepts of SEU are decision making under risk, value and probability. There are five axioms that provide the foundation of SEU theory; also there are three principles which are transitivity, invariance and independence. The invariance principle states that given the same options the decision process always should yield the same decision and the independence principle outcomes which are common to the choices in a decision problem should not affect the decision. The first axiom of SEU Theory is preferences are well ordered, for any two possible outcomes, A and B, either A is preferred to B or B to A or the decision maker is indifferent in the sense of not caring which transpires. The second axiom is preferences are transitive.	Okay. So, image one, er, appears to depict, er, two women who are, <laugh>, one is laughing and one is smiling. The woman that is smiling is standing up and she appears to be grabbing for something, erm, or helping the gir-, the other woman in the picture, er, with something. And the woman sitting down is laughing, er, but not looking at the woman standing up. She appears to be laughing at something or someone in the distance. And both women, <misc> appear to be distracted by different things, I would say, despite the fact that they are standing, er, or quite close together in the image. And they appear to be casually dressed. Er, it looks like a textbook image, er, like one of those models that are in textbooks for students. <misc>Okay.

2. The PAN'23 Authorship Verification Corpus

Similar to the PAN'22 edition of the task [19], the new dataset is based on the current Aston 100 Idiolects Corpus in English.[1] This corpus contains a variety of discourse types written by about 100 people. All subjects are of similar age (18-22) and native English speakers. The topic of the text samples is not restricted. Specifically, we consider four discourse types: two from written language (i.e., emails and essays) and two from oral language (i.e., interviews and speech transcripts). All six possible pairings of discourse types are examined. Note that the essay–email pairing was also included in the PAN'22 dataset.

Since the length of e-mails can be very short, we concatenate consecutive messages (ordered by date) so that we get text samples of at least 2,000 characters. Since the corpus also contains individual interview utterances, we also concatenate consecutive utterances to obtain text samples of at least 2,000 characters. All text samples in the corpus have been pre-processed to replace named entities with general tags. This helps to reduce the potentially confusing

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

Table 2
Statistics of PAN'23 datasets used for the cross-discourse authorship verification task.

Dataset statistic	Training	Test
Text pairs		
Positive	4,418 (50.0%)	4,828 (50.0%)
Negative	4,418 (50.0%)	4,828 (50.0%)
Email - Speech transcription	1,036 (11.7%)	1,074 (11.1%)
Essay - Email	1,454 (16.5%)	1,618 (16.8%)
Essay - Interview	884 (10.0%)	938 (9.7%)
Essay - Speech transcription	256 (2.9%)	206 (2.1%)
Interview - Email	4,564 (51.7%)	5,214 (54.0%)
Speech transcription - Interview	642 (7.3%)	606 (6.3%)
Text length (avg. chars)		
Email	2,308	2,346
Essay	9,894	10,770
Interview	2,503	2,501
Speech transcription	2,395	2,537

influence of topics on classification. Examples of such tags can be seen in the text samples of Table 1. Finally, nonverbal vocalizations (e.g., coughing, laughing) that occur regularly in spoken discourse are also replaced by special tags.

To create training and test datasets, we first split the available individuals into two equally sized, non-overlapping sets. More specifically, the text samples of 56 individuals are used for the training dataset and the test dataset is obtained from another set of 56 individuals. Both groups of authors have a similar gender distribution. Each dataset consists of a set of document pairs, and in each pair the documents belong to different discourse types. Because the distribution of available text samples across discourse types is not balanced, the distribution of document pairs across the six possible pairs of discourse types is also not homogeneous, as can be seen in Table 2. However, it is strikingly similar between training and test datasets. Moreover, both datasets are balanced with respect to pairs with the same author and pairs with different authors. This is also true when each specific pairing of discourse types is considered separately.

3. Evaluation Setup

The evaluation setup is similar to the one used for the previous shared tasks at PAN [16, 17, 18]. Formally, the objective function $\phi : (d_k, d_u) \rightarrow \{T, F\}$, where d_k is a text of known authorship, d_u is a text of unknown or disputed authorship, and $\{T, F\}$ indicate truth values has to be approximated. 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 current edition of the task, d_k and d_u belong to different discourse types of written or oral language.

For each instance of authorship verification (i.e., a text pair) in the test dataset, participants must provide a scalar score a_i (in the range $[0, 1]$) indicating the probability that the pair was authored by the same person. It is possible for participants to leave text pairs unanswered by giving a score of exactly $a_i = 0.5$.

3.1. Evaluation measures

Following the practice of recent editions of the authorship verification task [17, 18], we adopt a number of effectiveness measures to highlight different aspects of the capabilities of an authorship verification model. Specifically, the following measures have been used for evaluation:

- AUROC: the area under the ROC curve.
- c@1: a variant of the conventional accuracy measure that rewards systems that leave difficult verification cases unanswered [21].
- F_1: the well-known F_1-effectiveness measure (*not* considering unanswered cases).
- $\mathrm{F}_{0.5u}$: a newly proposed $\mathrm{F}_{0.5}$-based measure that highlights correctly answered instances of the same author and rewards unanswered instances [22].
- Brier: the complement of the Brier loss function [23] focusing on the accuracy of probabilistic predictions (as implemented in sklearn) [24]). 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 ($a_i = 0.5$). Consistent with the other measures, we take its complement so that higher scores correspond to better effectiveness.
- The average of the above measures has been used as the final overall score for ranking the submitted systems.

All submitted systems are deployed and executed on TIRA [25]. In addition to effectiveness, we also report the runtime to determine the efficiency of the submitted approaches.

3.2. Baselines

Several baseline methods are used to obtain an estimate of the difficulty of the task and the specific evaluation data set. These baselines are established approaches from the relevant literature, representing both n-gram-based and neural network-based methods. The latter involve methods that performed relatively well in the previous edition of the task [19], which is very similar to the current one. Specifically, the following baselines are used:

- *compressor*: given a pair of texts t_1 and t_2, the cross-entropy of t_2 is calculated based on the Prediction by Partial Matching (PPM) model of t_1 and vice versa [26]. A logistic regression classifier is then trained using the mean and absolute difference of the two cross-entropies. In addition, using a small radius sets verification values around 0.5 to exactly 0.5.
- *cngdist*: the most common 4-character frames are extracted from the training texts and used to represent each text. Then, for a pair of texts, the cosine similarity between the two texts is calculated [1]. During training, two thresholds p_1 and p_2 are optimized to scale the verification results. All review results lower than p_1 correspond to negative responses, all results higher than p_2 are scaled to positive responses, and the remaining results are set to 0.5, indicating difficult cases that are intentionally left unanswered.
- *najafi22*: a pre-trained language model (T5) in combination with a convolutional neural network and an attention mechanism is used [27]. In addition to text content, this approach also uses information about parts of speech, punctuation, emoji, and tags for named entities. In terms of overall effectiveness, this method achieved the best results in the PAN'22 edition of the task.

Table 3
Review of the basic features of the submitted approaches (sorted alphabetically). Augm., DTS, and Chunk. denote augmentation, discourse type-specific, and chunking, respectively.

System	Representation	Classification	Augm.	DTS	Chunk.
Guo et al. [32]	BERT	Contrastive learning	Yes	No	No
Huang et al. [33]	BERT	Contrastive learning	Yes	No	No
Ibrahim et al. [34]	S-BERT	Contrastive learning	Yes	No	Yes
Li et al. [35]	BERT	Fully connected	No	No	Yes
Liu et al. [36]	BERT	Fully connected	No	No	Yes
Lv et al. [37]	BERT	Fully connected	No	No	No
Petropoulos [38]	BERT	Contrastive learning, BiLSTM	Yes	Yes	Yes
Qiu et al. [39]	BERT	Contrastive learning	Yes	No	No
Sanjesh and Mangai [31]	n-grams, function words, etc.	cosine similarity	No	No	No
Sun et al. [30]	Adhominem, n-grams, function words, etc.	Bayes factor scoring	No	No	No
Valdez Valenzuela et al. [29]	Graph convolutional network	Fully connected	Yes	No	No

- *galicia22*: a neural graph network to represent text based on part-of-speech labels. Then, a Siamese network with a global attention layer and a final fully linked layer determines the output verification score. This method was very close to the best overall effectiveness on PAN'22.

The first two baselines (*compressor* and *cngdist*) are trained using the PAN'23 training dataset described in the previous section. The last two baselines (*najafi22* and *galicia22*) are used exactly as they were submitted to PAN'22, i.e., they were trained on the PAN'22 training dataset, which only considers discourse types of written language [19].

4. Survey of Submissions

We received eleven submissions from as many research teams, who deployed their software on TIRA. Each team also submitted a notebook describing the details of their method. In this section, a review of the main characteristics of the submitted methods is performed, as shown in Table 3. The majority of participants use contextual embeddings provided by pre-trained language models to represent texts. Despite the availability of a number of pre-trained models, all of these approaches favor BERT models [28]. Other neural network methods for representing text are based on graph convolutional networks [29] and a BiLSTM combined with an attention mechanism [6, 30]. Then again, some participants use more traditional features such as n-grams, function words, measures of vocabulary richness, etc. [30, 31]

In terms of classification, a popular option is to place fully linked layers on top of neural-based text representation layers. Another approach popular with PAN'23 participants is contrastive learning. Finally, cosine similarity or Bayes factor scoring are used by distance-based approaches

that utilize a number of features. Given the relatively small size of the training dataset, several participants are using deep learning approaches in an attempt to expand it. Most of these attempts use available metadata about the people who wrote the texts to identify all possible pairs of texts. In addition, one participant also used the PAN'22 training dataset to supplement the available data [34]. Given the input length limitations of pre-trained language models, multiple PAN'23 participants segment input texts into sections and produce multiple pairs of text sections from the same original text pair [34, 38]. The scores of each text piece pair are then combined to obtain the final score of the original text pair.

Despite the fact that the PAN'23 dataset includes a set of six discourse type pairs (e.g., essay–email, essay–interview, etc.), the vast majority of submitted methods treat all of these pairs in a homogeneous manner. Only one participant used a discourse type-specific approach [38]. More specifically, one model is trained on written discourse texts and another model is trained on spoken discourse texts. These are finally combined with a general model trained from all available texts.

5. Evaluation Results

We received submissions from eleven research teams. Unlike previous editions of the task, we allowed a maximum of three runs per participant to allow for a more thorough evaluation of the submitted methods if certain hyperparameter settings vary or if certain components of the proposed method are changed. In this paper, a total of 27 runs are evaluated. In addition, the four baselines described previously are also evaluated to provide useful insights.

Table 4 contains the final evaluation results for the PAN'23 test dataset of all submitted software and baselines. As can be seen, the overall effectiveness of all evaluated methods is quite low. This reflects the difficulty of the task and shows that it is really difficult to identify personal stylistic features of authors across different discourse types, especially when both written and oral utterances are studied. The most successful approaches seem to follow the same basic architecture, namely a pre-trained language model combined with contrastive learning [34, 32, 38]. However, a naive distance-based baseline trained on character n-grams is very competitive. Moreover, a baseline trained on the PAN'22 dataset (*galicia22*) achieves a slightly better overall effectiveness than a very similar approach trained on the PAN'23 dataset [29]. The winning approach of Ibrahim et al. [34] achieves the best or near-best effectiveness on all evaluation measures. Several participants achieve relatively low Brier scores that significantly affect their overall effectiveness, despite achieving relatively high AUROC and c@1) scores [33, 36, 37, 39]. This can be explained by the fact that they all provide binary answers and not probability values. In most cases where multiple runs are submitted by the same team, there are no significant differences in effectiveness. An exception is Guo et al. [32], especially when F_1 is considered.

5.1. Results by Discourse Type Pairing

From Table 2 we see that the distribution of verification instances across discourse type pairings is far from balanced. In addition, essays in the PAN'23 dataset are on average four times longer than emails, interviews, and voice transcripts. Of the six discourse type pairs, one includes only written language texts (e.g., essay–email), one refers only to spoken language texts (e.g., interview–speech-transcription), while the remaining pairs examine a combination of written and spoken language.

Table 4
Final results for the cross-discourse type authorship verification task at PAN'23. The systems are ranked according to their average effectiveness in five evaluation criteria. The best result per column is in bold.

Systems	Run	AUROC	c@1	F_1	$F_{0.5u}$	Brier	Overall
Ibrahim et al. [34]	reduced-graph	**0.616**	**0.572**	0.617	0.562	0.746	**0.623**
Ibrahim et al. [34]	resolving-globe	**0.616**	**0.572**	0.617	0.562	0.746	**0.623**
Guo et al. [32]	irregular-strategist	0.581	0.557	0.621	**0.571**	0.742	0.614
Ibrahim et al. [34]	golden-ottoman	0.598	0.546	0.622	0.550	0.744	0.612
BASELINE	cngdist	0.516	0.499	**0.666**	0.555	0.741	0.595
Petropoulos [38]	graceful-chianti	0.526	0.514	0.624	0.549	0.743	0.591
Petropoulos [38]	clever-daemon	0.525	0.516	0.622	0.550	0.743	0.591
BASELINE	galicia22	0.504	0.502	0.650	0.552	0.740	0.589
Valdez Valenzuela et al. [29]	GNN-SHORT	0.511	0.508	0.655	0.555	0.705	0.587
Sun et al. [30]	SDML epoch 8	0.504	0.502	0.632	0.546	0.747	0.586
Sun et al. [30]	SDML epoch 24	0.505	0.501	0.601	0.536	0.749	0.578
Guo et al. [32]	uniform-reward	0.595	0.555	0.460	0.527	0.723	0.572
Valdez Valenzuela et al. [29]	GNN-FULL	0.517	0.512	0.628	0.549	0.644	0.570
Sun et al. [30]	SDML epoch 35	0.511	0.508	0.558	0.526	0.749	0.570
Valdez Valenzuela et al. [29]	GNN-MED	0.503	0.502	0.602	0.534	0.709	0.570
BASELINE	najafi22	0.601	0.569	0.466	0.543	0.595	0.555
Huang et al. [33]	isochoric-paint	0.563	0.563	0.511	0.550	0.563	0.550
Liu et al. [36]	coincident-sound	0.548	0.548	0.544	0.547	0.548	0.547
Lv et al. [37]	radioactive-copyright	0.553	0.553	0.504	0.540	0.553	0.541
Huang et al. [33]	steel-coriander	0.500	0.500	0.651	0.551	0.500	0.540
Li et al. [35]	wan-ocean	0.500	0.500	0.646	0.550	0.500	0.539
Lv et al. [37]	tender-bugle	0.551	0.551	0.501	0.537	0.551	0.538
Lv et al. [37]	cold-rotor	0.550	0.550	0.465	0.524	0.550	0.528
Qiu et al. [39]	corn-mall	0.540	0.540	0.421	0.499	0.540	0.508
Qiu et al. [39]	poky-deck	0.540	0.540	0.421	0.499	0.540	0.508
Liu et al. [36]	perpendicular-field	0.534	0.534	0.421	0.493	0.534	0.503
Liu et al. [36]	foggy-raster	0.533	0.533	0.424	0.493	0.533	0.503
BASELINE	compressor	0.506	0.051	0.626	0.076	**0.750**	0.402
Sanjesh and Mangai [31]	calm-lyrics	0.525	0.500	0.030	0.068	0.729	0.370
Sanjesh and Mangai [31]	null-midpoint	0.523	0.499	0.031	0.066	0.730	0.370
Sanjesh and Mangai [31]	Multi-Feature Classifier	0.501	0.010	0.000	0.000	0.750	0.252

Table 5 provides a more detailed look at the effectiveness of the submitted systems with respect to the six pairings of discourse types. Only the total score (i.e., the average of AUROC, c@1, F_1, $F_{0.5u}$, and Brier) for each pairing is reported. First, note that all runs of Ibrahim et al. [34] as well as the baseline of *najafi22* show very high performance for the essay–email pairing. This is easily explained because the evaluation dataset of PAN'23 for this pairing is actually included in the training dataset of PAN'22. As mentioned earlier, Ibrahim et al. extended the training dataset by using the entire training dataset from PAN'22. In addition, the baselines *najafi22* and *galicia22* were trained with the PAN'22 training dataset. Therefore, evaluating these approaches using the essay–email pairing is biased. It is worth noting, however, that *galicia22* does not perform significantly higher on the essay–email pairing than on the other discourse type pairs.

Table 5
Evaluation results overall score for each discourse type pairing. ES, EM, IN, and ST denote essay, email, interview, and speech transcription, respectively. The best result per column is shown in bold.

System	Run	ES-EM	ES-IN	ES-ST	EM-ST	IN-EM	ST-IN	All
Ibrahim et al. [34]	reduced-graph	0.888	0.594	**0.619**	0.584	0.535	0.609	**0.623**
Ibrahim et al. [34]	resolving-globe	0.888	0.594	**0.619**	0.584	0.535	0.609	**0.623**
Guo et al. [32]	irregular-strategist	0.640	**0.605**	0.550	**0.607**	**0.613**	0.611	0.614
Ibrahim et al. [34]	golden-ottoman	0.835	0.600	0.605	0.567	0.539	**0.617**	0.612
BASELINE	cngdist	0.595	0.592	0.570	0.594	0.598	0.606	0.595
Petropoulos [38]	graceful-chianti	0.512	0.592	0.588	0.604	0.596	0.612	0.591
Petropoulos [38]	clever-daemon	0.519	0.592	0.583	0.605	0.596	0.609	0.591
BASELINE	galicia22	0.590	0.568	0.595	0.578	0.595	0.588	0.589
Valdez Valenzuela et al. [29]	GNN SHORT	0.582	0.586	0.580	0.600	0.585	0.579	0.587
Sun et al. [30]	SDML epoch8	0.584	0.583	0.549	0.562	0.592	0.597	0.586
Sun et al. [30]	SDML epoch24	0.565	0.566	0.575	0.569	0.585	0.585	0.578
Guo et al. [32]	uniform-reward	0.593	0.543	0.461	0.568	0.574	0.584	0.572
Valdez Valenzuela et al. [29]	GNN FULL	0.581	0.539	0.555	0.554	0.575	0.575	0.570
Sun et al. [30]	SDML epoch35	0.551	0.547	0.551	0.580	0.575	0.594	0.570
Valdez Valenzuela et al. [29]	GNN MED	0.569	0.572	0.593	0.554	0.576	0.499	0.570
BASELINE	najafi22	**0.918**	0.482	0.497	0.473	0.423	0.575	0.555
Huang et al. [33]	isochoric-paint	0.570	0.569	0.496	0.527	0.541	0.598	0.550
Liu et al. [36]	coincident-sound	0.589	0.498	0.468	0.520	0.544	0.596	0.547
Lv et al. [37]	radioactive-copyright	0.544	0.503	0.428	0.533	0.542	0.604	0.541
Huang et al. [33]	steel-coriander	0.537	0.541	0.535	0.538	0.541	0.546	0.540
Li et al. [35]	wan-ocean	0.537	0.548	0.539	0.531	0.539	0.546	0.539
Lv et al. [37]	tender-bugle	0.593	0.510	0.520	0.527	0.526	0.565	0.538
Lv et al. [37]	cold-rotor	0.538	0.467	0.547	0.523	0.533	0.545	0.528
Qiu et al. [39]	corn-mall	0.499	0.500	0.474	0.468	0.511	0.598	0.508
Qiu et al. [39]	poky-deck	0.499	0.500	0.474	0.468	0.511	0.598	0.508
Liu et al. [36]	perpendicular-field	0.538	0.477	0.350	0.459	0.508	0.513	0.503
Liu et al. [36]	foggy-raster	0.479	0.416	0.372	0.492	0.515	0.562	0.503
BASELINE	compressor	0.457	0.413	0.439	0.252	0.256	0.541	0.402
Sanjesh and Mangai [31]	calm-lyrics	0.366	0.360	0.427	0.383	0.364	0.403	0.370
Sanjesh and Mangai [31]	null-midpoint	0.363	0.359	0.426	0.382	0.363	0.401	0.370
Sanjesh and Mangai [31]	Multi-Feature Classifier	0.253	0.252	0.256	0.250	0.252	0.254	0.252

Apart from these biased results, the *irregular-stategist* run of Guo et al. [32] achieves the best results for the essay–email pairing, as well as for three other pairings combining written and spoken language. For the remaining pairings of discourse types, the runs of Ibrahim et al. obtain the best results. The latter's effectiveness, however, is relatively low for the interview–email pairing, which corresponds to the majority of the review cases (see Table 2). Moreover, both the [29] approach (the *GNN SHORT* run) and two baselines (i.e., *galicia22* and *cngdist*) show relatively high robustness of their effectiveness across all six pairings.

It is also worth noting that the average effectiveness of all teams' methods is higher when the discourse types involved are both from eithger written language (essay–email) or spoken language (speech transcription–interview) than in the cases when one discourse type is from written language and the other is from spoken language. This suggests that the inherent differences between written and spoken language further complicate the task.

Table 6
Biases in the response of submitted systems, sorted by their overall effectiveness in Table 4.

System	Run	Positive (%)	Negative (%)	Non-response (%)
Ibrahim et al. [34]	reduced-graph	49.9	36.2	13.9
Ibrahim et al. [34]	resolving-globe	49.9	36.2	13.9
Guo et al. [32]	irregular-strategist	67.0	33.0	0.0
Ibrahim et al. [34]	golden-ottoman	56.7	29.3	13.9
BASELINE	cngdist	99.8	0.2	0.0
Petropoulos [38]	graceful-chianti	79.3	20.7	0.0
Petropoulos [38]	clever-daemon	77.9	22.1	0.0
BASELINE	galicia22	92.4	7.6	0.0
Valdez Valenzuela et al. [29]	GNN SHORT	92.1	7.4	0.4
Sun et al. [30]	SDML epoch8	85.3	14.7	0.0
Sun et al. [30]	SDML epoch24	75.1	24.9	0.0
Guo et al. [32]	uniform-reward	32.5	67.5	0.0
Valdez Valenzuela et al. [29]	GNN FULL	80.4	18.9	0.7
Sun et al. [30]	SDML epoch35	61.5	38.5	0.0
Valdez Valenzuela et al. [29]	GNN MED	73.4	24.4	2.2
BASELINE	najafi22	30.7	69.0	0.3
Huang et al. [33]	isochoric-paint	39.4	60.6	0.0
Liu et al. [36]	coincident-sound	49.2	50.8	0.0
Lv et al. [37]	radioactive-copyright	40.0	60.0	0.0
Huang et al. [33]	steel-coriander	93.3	6.7	0.0
Li et al. [35]	wan-ocean	91.0	9.0	0.0
Lv et al. [37]	tender-bugle	40.0	60.0	0.0
Lv et al. [37]	cold-rotor	34.2	65.8	0.0
Qiu et al. [39]	corn-mall	29.5	70.5	0.0
Qiu et al. [39]	poky-deck	29.5	70.5	0.0
Liu et al. [36]	perpendicular-field	30.4	69.6	0.0
Liu et al. [36]	foggy-raster	31.1	68.9	0.0
BASELINE	compressor	2.6	2.0	95.4
Sanjesh and Mangai [31]	calm-lyrics	1.5	98.5	0.0
Sanjesh and Mangai [31]	null-midpoint	1.5	95.5	3.0
Sanjesh and Mangai [31]	Multi-Feature Classifier	0.0	1.0	0.0

5.2. Response Bias

As previously reported, the PAN'23 dataset for authorship verification is perfectly balanced in terms of positive (same author) and negative (different author) instances. This is true for both the training and test datasets and for the different discourse type pairs. It is therefore interesting to examine whether the approaches presented provide a roughly balanced set of responses or whether they are biased toward a particular class (i.e., positive or negative responses). Moreover, according to the task's experiment setup, it is possible for participants to leave a verification instance unanswered by returning a probability of exactly 0.5. Some evaluation measures (e.g., c@1) are specifically designed to account for such response non-answers. Therefore, it is interesting to see which submitted approaches actually use this possibility and to what extent.

Table 6 shows the percentage of positive and negative responses and non-answers relative to the total number of possible instances of the test dataset (i.e., 9,656 instances). It should be emphasized that the positive/negative answers reported are not necessarily correct. Also note that the percentages from the last run by Sanjesh and Mangai [31] do not add up to 100 because they only provided answers for a small subset of the evaluation data set.

As can be seen, most of the best performing approaches favor positive responses over negative ones. The two best runs of Ibrahim et al. [34] are relatively balanced in this respect. The runs of Petropoulos [38], Valdez Valenzuela et al. [29], and Sun et al. [30] are clearly focused on positive responses. An extreme case is the baseline *cngdist*, where almost all responses are positive. It is also noticeable that one run of Guo et al. [32] (*irregular-strategist*) tends to have positive responses, while the other run (*uniform-reward*) has the opposite tendency.

As for the percentage of non-answer review cases, it is easy to see that few participants attempt to use this option. More specifically, Ibrahim et al. leaves a moderate number of instances (e.g., 13.9%) unanswered across all three runs. Moreover, Valdez Valenzuela et al. and Sanjesh and Mangai [31] exhibit at most 3% non-answers. An extreme case is the baseline *compressor*, which returns an answer less than 5% of the time. This probably means that a suitable tuning of the hyperparameters for this baseline could significantly improve its effectiveness.

5.3. Efficiency

So far, we have focused on the effectiveness of the methods presented. Another dimension of the evaluation is their efficiency, which indicates how well they apply to large datasets for authorship verification. For this purpose, Table 7 shows the runtime of each submitted run as recorded by TIRA. Comparing the runtime of the two best-performing runs of Ibrahim et al. shows that *resolving-globe* is much faster than *reduced-graph*, although they achieve exactly the same effectiveness (see Table 4). In fact, the former is the most efficient among all the submitted runs, with the exception of *Multi-Feature Classifier* by Sanjesh and Mangai, which, as mentioned before, processed only a small part of the test dataset. Other relatively fast approaches include the runs of Lv et al. [37] as well as Liu et al. [36]. The very competitive run of Guo et al. (*irregular-strategist*) has the highest runtime, much higher than the other run submitted (*uniform-reward*) by this team. Other approaches with relatively high runtime are the runs of Huang et al. [33], Li et al. [35], and Sanjesh and Mangai (*null-midpoint*).

6. Conclusion

Several earlier editions of PAN included an authorship verification task. In all of them, a variety of datasets have been developed and used to evaluate dozens of methods. Several scenarios were considered, e.g., when a small set of documents of known authorship is provided in each verification instance [13, 14, 15], when only text pairs are examined [16, 17, 18], when multiple languages are included in the dataset [13, 14, 15], and when only English fanfiction texts are used [16, 17]. The research community in this area has evolved significantly over the past decade, motivated in part by relevant PAN shared tasks.

Following the practice of the previous PAN edition, we focus on particularly difficult cases in which the input texts belong to different discourse types, i.e., there are crucial differences in

Table 7
Efficiency of the submitted approaches, sorted according to their overall effectiveness in Table 4.

System	Run	Runtime
Ibrahim et al. [34]	reduced-graph	290 minutes 15 seconds
Ibrahim et al. [34]	resolving-globe	15 minutes 34 seconds
Guo et al. [32]	irregular-strategist	2,115 minutes 41 seconds
Ibrahim et al. [34]	golden-ottoman	346 minutes 21 seconds
Petropoulos [38]	graceful-chianti	147 minutes 2 seconds
Petropoulos [38]	clever-daemon	145 minutes 47 seconds
Valdez Valenzuela et al. [29]	GNN SHORT	30 minutes 16 seconds
Sun et al. [30]	SDML epoch8	74 minutes 39 seconds
Sun et al. [30]	SDML epoch24	95 minutes 22 seconds
Guo et al. [32]	uniform-reward	30 minutes 35 seconds
Valdez Valenzuela et al. [29]	GNN FULL	83 minutes 10 seconds
Sun et al. [30]	SDML epoch35	75 minutes 19 seconds
Valdez Valenzuela et al. [29]	GNN MED	26 minutes 17 seconds
Huang et al. [33]	isochoric-paint	717 minutes 45 seconds
Liu et al. [36]	coincident-sound	23 minutes 21 seconds
Lv et al. [37]	radioactive-copyright	18 minutes 0 seconds
Huang et al. [33]	steel-coriander	708 minutes 51 seconds
Li et al. [35]	wan-ocean	768 minutes 57 seconds
Lv et al. [37]	tender-bugle	17 minutes 1 second
Lv et al. [37]	cold-rotor	17 minutes 7 seconds
Qiu et al. [39]	corn-mall	17 minutes 35 seconds
Qiu et al. [39]	poky-deck	603 minutes 14 seconds
Liu et al. [36]	perpendicular-field	23 minutes 28 seconds
Liu et al. [36]	foggy-raster	25 minutes 48 seconds
Sanjesh and Mangai [31]	calm-lyrics	388 minutes 53 seconds
Sanjesh and Mangai [31]	null-midpoint	1,017 minutes 15 seconds
Sanjesh and Mangai [31]	Multi-Feature Classifier	4 minutes 16 seconds

terms of target audience and communicative purpose. To make matters more challenging, in this edition we introduce the use of discourse types from both written and spoken language, with their inherent differences in formality and complexity. Not surprisingly, the effectiveness achieved is overall weak resembling in many cases a baseline with random estimates. When the average effectiveness of all participants is taken into account, the instances of authorship verification involving discourse types from both written and spoken language (e.g., interview–email) are more difficult than those involving only written discourse types or only spoken discourse types. This shows that the PAN'23 dataset is even more difficult than the corresponding PAN'22 dataset.

The majority of the submitted approaches are based on neural methods. Several submissions, including the most effective ones, use a pre-trained language model in combination with contrastive learning [40]. This seems appropriate to discard similarities in topics and hopefully dissimilarities in discourse types [12]. Conversely, there is lack of diversity in the methods used by participants, including the selection of the specific pre-trained language model. A broader

variety of approaches could lead to a more comprehensive analysis and deeper insights. This certainly suggests that there is much room for improvement in cross-discourse authorship verification. A possible direction to achieve better results could be the development of methods able to adapt to each specific discourse type pairing, taking into account the general characteristics of the discourse types involved. In addition, an appropriate way to combine traditional n-gram-based approaches with neural-based models could provide improved performance.

Another promising direction for future work is to leverage the availability of powerful generative language models (e.g., GPT) to improve both authorship verification methods and the quality of their evaluation. Finally, exploring the link between authorship verification and machine-generated text recognition is an area that can be explored in future work.

References

[1] M. Kestemont, J. Stover, M. Koppel, F. Karsdorp, W. Daelemans, Authenticating the writings of julius caesar, Expert Systems with Applications 63 (2016) 86–96.

[2] M. R. Schmid, F. Iqbal, B. C. Fung, E-mail authorship attribution using customized associative classification, Digital Investigation 14 (2015) S116–S126.

[3] V. K, D. Gupta, Detection of idea plagiarism using syntax–semantic concept extractions with genetic algorithm, Expert Systems with Applications 73 (2017) 11–26.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

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

[9] M. Kocher, J. Savoy, A simple and efficient algorithm for authorship verification, Journal of the Association for Information Science and Technology 68 (2017) 259–269.

[10] 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.

[11] T. Nguyen, C. Dagli, K. Alperin, C. Vandam, E. Singer, Improving long-text authorship verification via model selection and data tuning, in: Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 28–37. URL: https://aclanthology.org/2023.latechclfl-1.4.

[12] A. Wegmann, M. Schraagen, D. Nguyen, Same author or just same topic? towards content-independent style representations, in: Proceedings of the 7th Workshop on Representation Learning for NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 249–268. URL: https://aclanthology.org/2022.repl4nlp-1.26. doi:10.18653/v1/2022.repl4nlp-1.26.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pęzik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection, in: A. Barrón-Cedeño, G. Da San Martino, M. Degli Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer International Publishing, Cham, 2022, pp. 382–394.

[19] E. Stamatatos, M. Kestemont, K. Kredens, P. Pezik, A. Heini, J. Bevendorff, M. Potthast, B. Stein, Overview of the Authorship Verification Task at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022.

[20] R. L. Cayer, R. K. Sacks, Oral and written discourse of basic writers: Similarities and differences, Research in the Teaching of English 13 (1979) 121–128. URL: http://www.jstor.org/stable/40170748.

[21] 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.

[22] 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.
[23] G. W. Brier, et al., Verification of forecasts expressed in terms of probability, Monthly weather review 78 (1950) 1–3.
[24] 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.
[25] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.
[26] 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.
[27] 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.
[28] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.
[29] A. Valdez Valenzuela, H. G. Adorno, J. Martinez Galicia, Heterogeneous-Graph Convolutional Network for Authorship Verification, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[30] Y. Sun, S. Afanaseva, K. Patil, Stylometric and Neural Features Combined Deep Bayesian Classifier for Authorship Verification, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[31] R. Sanjesh, A. Mangai, A Multi-Feature Custom Classification Approach to Authorship Verification, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[32] M. Guo, Z. Han, H. Chen, H. Qi, A Contrastive Learning of Sample Pairs for Authorship Verification, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[33] Z. Huang, L. Kong, M. Huang, Authorship Verification based on CoSENT, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[34] M. Ibrahim, A. Akram, M. Radwan, R. Ayman, M. Abd-El-Hameed, N. El-Makky, M. Torki, Enhancing Authorship Verification using Sentence-Transformers, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[35] J. Li, Q. Zhang, M. Huang, Author Verification of text fragments based on the Bert model, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[36] X. Liu, L. Kong, M. Huang, Text-Segment Interaction for Authorship Verification using BERT-based Classification, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[37] J. Lv, Y. Han, Q. Dong, Application of R-Drop in Author Authorship Verification, in:

M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[38] P. Petropoulos, Contrastive Learning for Authorship Verification using BERT and Bi-LSTM in a Siamese Architecture, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[39] Y. Qiu, H. Qi, Y. Han, K. Huang, Authorship Verification Based on SimCSE, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023.
[40] R. A. Rivera-Soto, O. E. Miano, J. Ordonez, B. Y. Chen, A. Khan, M. Bishop, N. Andrews, Learning universal authorship representations, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 913–919. URL: https://aclanthology.org/2021.emnlp-main.70. doi:10.18653/v1/2021.emnlp-main.70.