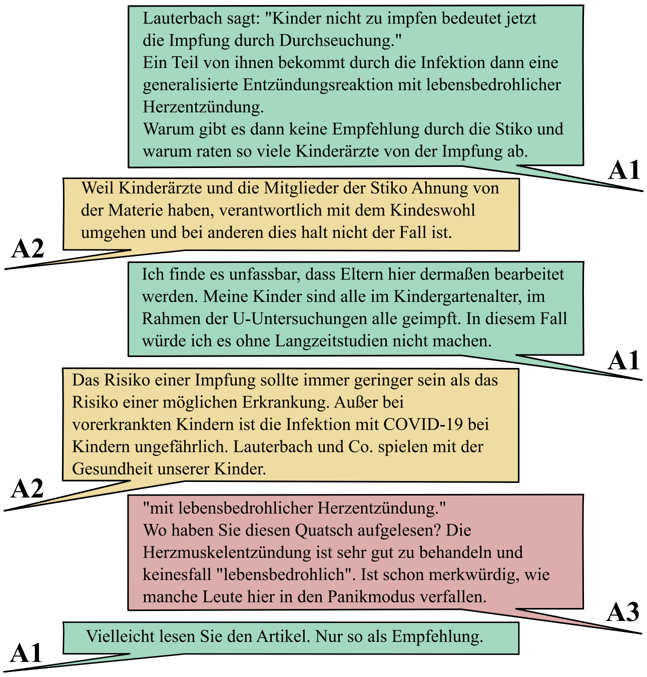
Who Wrote When? Author Diarization in Social Media Discussions

Benedikt Boenninghoff [1], Henry Hosseini [2], Robert M. Nickel [3], Dorothea Kolossa [4],
[1]Ruhr-Universität Bochum, benedikt.boenninghoff@rub.de,
[2]Universität Münster, henry.hosseini@wi.uni-muenster.de,
[3]Bucknell University, rmn009@bucknell.edu,
[4]Technische Universität Berlin, dorothea.kolossa@tu-berlin.de

Abstract

We are proposing a novel framework for author diarization, i.e. attributing comments in online discussions to individual authors. We consider an innovative approach that merges pre-trained neural representations of writing style with author-conditional encoder-decoder diarization, enhanced by a Conditional Random Field with Viterbi decoding for alignment refinement. Additionally, we introduce two new large-scale German language datasets, one for authorship verification and the other for author diarization. We evaluate the performance of our diarization framework on these datasets, offering insights into the strengths and limitations of this approach.



Lauterbach sagt: "Kinder nicht zu impfen bedeutet jetzt die Impfung durch Durchseuchung."
Ein Teil von ihnen bekommt durch die Infektion dann eine generalisierte Entzündungsreaktion mit lebensbedrohlicher Herzentzündung.
Warum gibt es dann keine Empfehlung durch die Stiko und warum raten so viele Kinderärzte von der Impfung ab.

Weil Kinderärzte und die Mitglieder der Stiko Ahnung von der Materie haben, verantwortlich mit dem Kindeswohl umgehen und bei anderen dies halt nicht der Fall ist.

Ich finde es unfassbar, dass Eltern hier dermaßen bearbeitet werden. Meine Kinder sind alle im Kindergartenalter, im Rahmen der U-Untersuchungen alle geimpft. In diesem Fall würde ich es ohne Langzeitstudien nicht machen.

Das Risiko einer Impfung sollte immer geringer sein als das Risiko einer möglichen Erkrankung. Außer bei vorerkrankten Kindern ist die Infektion mit COVID-19 bei Kindern ungefährlich. Lauterbach und Co. spielen mit der Gesundheit unserer Kinder.

"mit lebensbedrohlicher Herzentzündung."
Wo haben Sie diesen Quatsch aufgelesen? Die Herzmuskelentzündung ist sehr gut zu behandeln und keinesfall "lebensbedrohlich". Ist schon merkwürdig, wie manche Leute hier in den Panikmodus verfallen.

Vielleicht lesen Sie den Artikel. Nur so als Empfehlung.

Figure 1: Example discussion taken from the extracted dataset.

1 Introduction

Author Diarization (AD) tries to answer whether a single author or multiple authors have written a chronologically ordered series of texts. The task of author diarization can generally be described as the process of dividing an input text stream into text segments belonging to an individual author or to collaboratively working authors.

This work focuses on online discussions, such as those on a newspaper's forum or chat rooms. Two key assumptions guide our approach. Firstly, style changes are assumed to occur only between comments. Secondly, multiple authors are assumed to not collaborate on the same comment. Consequently, we redefine the task as follows: Given a sequence of comments, determine the number of authors and uniquely assign all discussion comments to their respective authors. Figure 1 illustrates the described scenario and the expected output for the author diarization task[1].

Author diarization is intricately linked to the *Authorship Verification* (AV) and *Style Change Detection* (SCD) domains. While existing AV approaches are proven to be effective for single-instance verification (Ishihara et al., 2024; Tyo et al., 2023; Manolache et al., 2024), AD provides a comprehensive view of the entire conversation, making it more practical for real-world applications where posts are seldom isolated and often part of an ongoing dialogue. Significant advancements have been achieved through the PAN series of SCD competitions spanning from 2017 to 2024[2]. It is commonly used to detect instances of plagiarism or ghostwriting. However, in contrast to author diarization, SCD aims to identify shifts in linguistic style within a text (Zangerle et al., 2022, 2023). This task seeks to pinpoint transitions between sentences or paragraphs where such changes occur, indicating a shift from one style to another, irre-

[1] Example is taken from the Zeit-Online website: https://www.zeit.de/hamburg/2021-07/corona-impfung-kinder-jugendliche-hamburg-bayern\#comments

[2] https://pan.webis.de/shared-tasks.html

Findings of the Association for Computational Linguistics: EACL 2024, pages 15721–15734
November 12-16, 2024 ©2024 Association for Computational Linguistics

Datasets	**#Train**	**#Dev**	**#Test**
Verification	762,690	87,610	88,510
Diarization (small)	313,042	67,744	66,845
Diarization (large)	440,822	95,334	94,398

Table 1: Number of **pairs** for the authorship verification task and number of **discussions** for the author diarization task.

spective of individual authors. Hence, algorithms for detecting stylistic changes are designed for different objectives, outputting binary decisions (e.g., a zero or one).

Our approach mainly consists of three stages: We use an AV task as a proxy to learn neural representations of writing style. AV methods are generally unable to directly perform the diarization task since they cannot resolve ambiguities that inevitably occur when three or more authors receive inconsistent pairwise labels as same-author pairs and different-author pairs. Hence, in a second stage, we train an author-conditional encoder-decoder diarizer in a permutation-invariant fashion. Lastly, we implement a conditional random field (CRF) and Viterbi decoder in a third stage to revise/correct the estimated author-comment alignments.

Due to the absence of publicly available large-scale diarization datasets, we compiled two substantial datasets of German newspaper comments for our analysis: The first AV dataset contains 938,810 comment pairs. The second one contains 630,554 discussions to train and evaluate the author diarizer. More precisely, we conduct an ablation study to demonstrate the impact of revising the estimated alignment between authors and comments.

In summary, our main contributions are the following: (1) We created two novel large-scale German language datasets serving as new benchmarks, one for authorship verification and one for author diarization. (2) We propose and evaluate a novel model for author diarization which can be easily adapted to languages other than German - the framework only requires the substitution of the language-specific BERT model.

2 Dataset Construction

We gathered two large-scale datasets comprising newspaper comments. These comments stem from forum discussions of articles that span from 2005 to 2021 on the German Zeit-Online forum discussion website[3]. All discussions are in German.

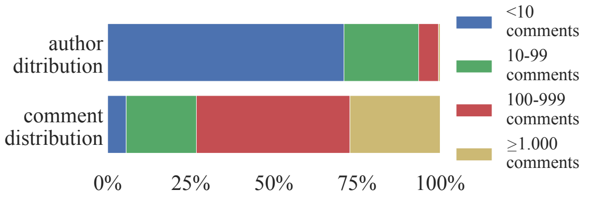


Figure 2: Relationship between the percentage distribution of authors and their contribution to the total number of comments.

Table 1 summarizes the sizes of the datasets: For *verification*, we have 762,690 randomly paired training pairs, 87,610 development (dev) pairs, and 88,510 test pairs.

For *diarization*, both a small and large dataset are provided. The small dataset is a subset of the large dataset and involves discussions with one to four participants, whereas the large dataset includes discussions with up to six participants. The small dataset helps in identifying potential issues early and allows for quicker iterations during model development. The large dataset, on the other hand, is essential for testing the model's performance in more realistic and extensive scenarios. In total, the large diarization dataset is composed of 440,822 discussions for training, 95,334 for development, and 94,398 for testing.

The discussions are constrained to contain between one and ten comments each. Longer discussions are randomly partitioned into shorter ones to adhere to this constraint. While the sets of authors in the verification and diarization datasets overlap, the datasets themselves are distinct, as they do not share the same comments.

The comments are anonymized in that URLs, author names, emails, and direct quotes in each dataset were automatically substituted with placeholders, which were then incorporated as new trainable tokens into the model. Anonymization, or more accurately, text normalization, removes direct identifiers (e.g., names, emails) that could otherwise bias the model. This ensures that the model focuses on stylometric features that are indicative of an author's writing style, rather than any content-specific identifiers.

Figure 2 illustrates the relationship between the percentage distribution of authors and their contribution to the total number of comments. The first bar chart shows the distribution of authors by the

[3] https://www.zeit.de

Dataset	#1	#2	#3	#4	#5	#6
train	4.8%	19.8%	25.3%	21.1%	15.7%	13.3%
dev	4.8%	19.6%	25.3%	21.3%	15.6%	13.3%
test	4.8%	19.9%	25.1%	21.1%	15.8%	13.4%

Table 2: Percentage of unique author counts (from one to six) per discussion across the train, dev, and test splits for the large dataset.

number of comments they produce, while the second bar chart shows the distribution of comments attributed to these authors. For instance, the blue segment represents the percentage of authors who have produced fewer than 10 comments. The first chart indicates that a large majority (nearly 75%) of authors fall into this category. Despite being the largest author group, their contribution to total comments is minimal, as shown in the second chart. There is a clear disparity between the number of authors and their contribution to the total comments. The majority of authors are infrequent commenters, while a smaller fraction of highly active authors generate the majority of the content. This pattern can often be seen on social media and other content platforms, where a smaller number of users contribute most of the content.

Table 2 presents the percentage distribution of unique author counts per discussion across the training, dev, and test sets in the large dataset. The percentage distributions are very consistent across the train, dev, and test sets, indicating a balanced dataset splitting. Less than 25% of the large dataset of the discussions contain one or two commenters. More than 75% of the discussions contain three to six commenters. Our findings indicate that the majority of discussions involve a more diverse set of participants, highlighting a varied and rich interaction landscape within the dataset. This contradicts the expectation that most discussions are dominated by individuals.

Figure 3 depicts the distribution of the top ten topics discussed between 2016 and 2021. The topics were provided in the metadata during data collection. Initially, topics were evenly distributed until 2020, when COVID-19 articles surged and their discussions started dominating the discourse.

Further details and dataset assessment are provided in the Appendix.

3 Method

This section describes the different modules of our author diarization framework in detail. The incoming text data is processed in four stages. A block diagram of the framework is shown in Fig. 4.

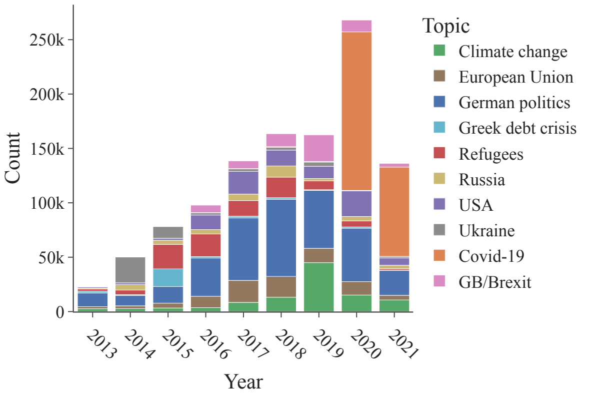


Figure 3: Top ten topic distribution of the discussions. The topic shifts, e.g., COVID-19 or Brexit, are evident.

3.1 Neural Feature Extraction

In stage ①, we train a feature extraction model on an AV dataset. The AV training is used as a proxy to produce features that represent the writing style. Given two input comments, the AV model determines if the same author wrote them.

In this paper, we compare two published models. The first model, AdHominem (Boenninghoff et al., 2019; Boenninghoff et al., 2021b), outperformed all other systems at the PAN 2020 and 2021 AV shared tasks (Kestemont et al., 2021). The second model, SRoBERTa (Zhu and Jurgens, 2021), showed impressive results on social media data such as Amazon reviews and Reddit comments.

We made use of the publicly available source codes[4,5] of both implementations provided by the authors, but inserted the German version of FastText[6] word embeddings in AdHominem, and replaced the pre-trained English-language RoBERTa model by the German equivalent GBERT[7] (Chan et al., 2020). To keep the BERT-based models distinguishable, we refer to the fine-tuned model as the SGBERT model hereinafter.

The AV task can be described as follows (Boenninghoff et al., 2021a): Suppose that we have a pair of comments $\mathcal{C}_1$ and $\mathcal{C}_2$ with an associated ground-truth label $a \in \{0, 1\}$. The value of a indicates whether the two comments were written by the same author ($a = 1$) or by different authors ($a = 0$). Both methods (SGBERT and AdHominem) follow the same paradigm: They per-

[4] https://github.com/boenninghoff/pan_2020_2021_authorship_verification

[5] https://github.com/lingjzhu/idiolect

[6] https://fasttext.cc/docs/en/crawl-vectors.html

[7] https://huggingface.co/deepset/gbert-base

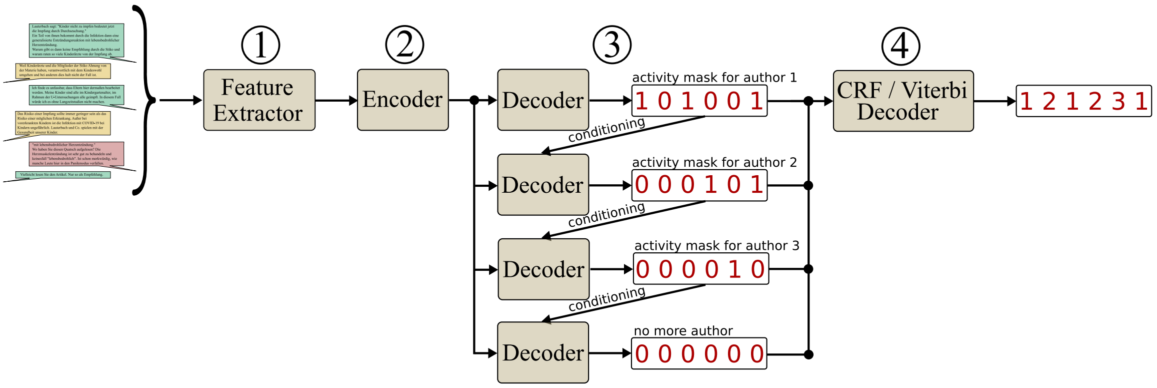


Figure 4: A block diagram of our author diarization framework. The input to the framework is a sequence of text segments, each of which may have been written by a different author. The output is a sequence of author indices. There is one index for each text segment. Text segments with the same index are attributed to the same author.

form neural feature extraction, $m_1\colon\{\mathcal{C}_1, \mathcal{C}_2\} \longrightarrow \{\boldsymbol{f}_1, \boldsymbol{f}_2\}$, followed by a (probabilistic) comparison of the extracted stylometric representations, $m_2\colon\{\boldsymbol{f}_1, \boldsymbol{f}_2\} \longrightarrow p \in [0, 1]$. The estimated label $\widehat{a}$ is obtained from a threshold test applied to the output score p. For both methods, we choose $\widehat{a} = 1$ if $p > 0.5$ and $\widehat{a} = 0$ if $p \leq 0.5$. Since we are only interested in the writing style representations (mapping m_1), we can express the feature extraction stage as

$$\begin{aligned} \boldsymbol{F}_{C\times D} &= \text{AdHominem}\big(\mathcal{C}_1, \ldots, \mathcal{C}_C\big), \\ \boldsymbol{F}_{C\times D} &= \text{SGBERT}\big(\mathcal{C}_1, \ldots, \mathcal{C}_C\big), \end{aligned} \tag{1}$$

where $\boldsymbol{F}_{C\times D} = \big[\,\boldsymbol{f}_1, \ldots, \boldsymbol{f}_C\,\big]^T$ represents the feature matrix, C defines the maximum number of comments in the discussion, and D denotes the dimension of the feature embeddings $\boldsymbol{f}_i$.

3.2 Author-Conditional Chain Model

The stages ② and ③ in Fig. 4 mainly follow the neural probabilistic model proposed in (Fujita et al., 2020). Given the stylometric representations of stage ①, the (binary) author diarizer consists of an encoder-decoder topology that assigns all comments uniquely to some author.

For the encoding part in stage ② we use a three-layer BiLSTM encoder as follows:

$$\boldsymbol{X}_{C\times D} = \text{BiLSTM}\big(\boldsymbol{F}_{C\times D}\big). \tag{2}$$

The motivation behind stage ② is that the feature vectors in $\boldsymbol{F}_{C\times D}$ represent the writing style of a single author while the features in $\boldsymbol{X}_{C\times D}$ are capable of capturing the temporal correlations and dependencies of the authors of comments within a discussion. Matrix $\boldsymbol{X}_{C\times D}$ in Eq. (2) denotes the output of the encoding part and serves as an input for the decoder.

The decoder in stage ③ of Fig. 4 runs through an iterative procedure that computes a binary activity mask for a single author at each stage. The activity mask assigns a one to a comment deemed to belong to the given author and assigns a zero otherwise. The total number of comments in the respective discussion defines the length of the activity mask. In each iteration, the encoder features and the previously estimated activity mask (Fujita et al., 2020) are used as inputs for the next update. Thus, the diarizer can process a variable (unknown) number of authors. For the a-th author, we define the corresponding binary mask $\boldsymbol{y}^{(a)}$ as a C-dimensional vector of zeros and ones, i.e. $\boldsymbol{y}^{(a)} = \big[y_c^{(a)} \in \{0, 1\} | c = 1, \ldots, C\big]$. Again, the ones in the masks indicate that the respective comments were written by the a-th author. The task is to find the most probable author-comment alignments, i.e.:

$$\widehat{\boldsymbol{y}}^{(1)}, .., \widehat{\boldsymbol{y}}^{(A)} = \underset{\boldsymbol{y}^{(1)} \ldots \boldsymbol{y}^{(A)}}{\arg\max} \Pr(\boldsymbol{y}^{(1)}, .., \boldsymbol{y}^{(A)} | \boldsymbol{F}), \tag{3}$$

where the variable A defines the number of authors participating in the discussion. Applying the chain rule results in

$$\begin{aligned} &\Pr(\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(A)} | \boldsymbol{F}) \\ &\quad = \prod_{a=1}^{A} \Pr(\boldsymbol{y}^{(a)} | \boldsymbol{y}^{(1)}, \ldots \boldsymbol{y}^{(a-1)}, \boldsymbol{F}). \end{aligned} \tag{4}$$

The product in Eq. (4) is iteratively realized by the LSTM-based decoder in stage ③. The iterations are initialized with a $C \times 1$-dimensional activity mask estimate $\widehat{\boldsymbol{y}}^{(0)}_{C\times 1}$ of all zeros. In the a-th iteration, we compute:

$$\boldsymbol{H}^{(a)}_{C\times D} = \tag{5}$$
$$\text{LSTM}\left(\begin{bmatrix} \boldsymbol{X}^{T}_{C\times D} \\ \text{Linear}\big(\widehat{\boldsymbol{y}}^{(a-1)}_{C\times 1}\big)^{T}_{C\times D} \end{bmatrix}^{T}, \boldsymbol{H}^{(a-1)}_{C\times D} \right),$$
$$\boldsymbol{z}^{(a)}_{C\times 1} = \text{Sigmoid}\big(\text{Linear}\big(\boldsymbol{H}^{(a)}_{C\times D}\big)_{C\times 1}\big),$$

where the LSTM cell in Eq. (5) receives two arguments, the input of size $C \times 2D$ and the LSTM hidden state from the previous step of size $C \times D$, and outputs the next hidden state $\boldsymbol{H}^{(a)}_{C\times D}$. The parameter T in $\boldsymbol{X}^{T}_{C\times D}$ refers to the transpose of the matrix, and Linear() is a linear projection that maps a $C\times 1$-dimensional vector into a $C\times D$-dimensional matrix (and vice versa) to achieve dimensional consistency with the encoder output $\boldsymbol{X}_{C\times D}$ in Eq. (2). We map the concatenated $2D$-dimensional input to D dimensions while broadcasting both dimensions, the batch size and C. Consequently, the LSTM cell is recursive along the author-axis only. Then, a linear projection with a Sigmoid function transforms the hidden state to a C-dimensional soft mask $\boldsymbol{z}^{(a)}_{C\times 1}$. Finally, thresholding yields the binary activity mask in the a-th iteration,

$$\widehat{\boldsymbol{y}}^{(a)}_{C\times 1} = \big[\mathbf{1}(z^{(a)}_{c} > 0.5) | c = 1, \dots, C\big]. \tag{6}$$

A binary cross-entropy (BCE) loss between the estimated activity masks and the (binary) ground-truth labels is used as the first part of the overall loss function to train the encode-decoder system. Note that the order of the selected authors in Fig. 4 is not essential. All permutations of the identified activity masks are valid. Therefore, we have to ensure that the training of the decoder is *permutation-invariant* (Yu et al., 2017). Stage ③ in Figure 4 shows one possible solution in the order of the selected authors. To mitigate the label ambiguity or permutation problem, the optimal order of the author activity masks is now determined by selecting the order that minimizes the BCE loss:

$$\phi^{\star} = \underset{\phi\in\text{perm(A)}}{\arg\min} \left\{ \sum_{c=1}^{C}\sum_{a=1}^{A} \text{BCE}\big(z^{(a)}_{c}, y^{(\phi_a)}_{c}\big) \right\}, \tag{7}$$

where perm(A) represents all possible author permutations of the given discussion and the terms $y^{(\phi_a)}_{c}$ denote the permutated ground-truth labels. In this way, the computation time is tractable because we run through the decoder in Eq. (4) one time, re-sort the ground-truth labels as shown in Eq. (7) and perform backpropagation. The BCE loss is then computed by two terms,

$$\mathcal{L}_{\text{BCE}} = \sum_{c=1}^{C}\sum_{a=1}^{A} \text{BCE}\big(z^{(a)}_{c}, y^{(\phi^{*}_{a})}_{c}\big) + \sum_{c=1}^{C}\sum_{a=A+1}^{A_{max}} \text{BCE}\big(z^{(a)}_{c}, 0\big), \tag{8}$$

where A_{max} is the pre-defined maximum number of authors, i.e., the maximum number of iterations for the decoder step. Hence, the first term in Eq. (8) measures the decoder output w.r.t. the ground-truth labels, while the second term punishes false alarms.

3.3 CRF and Viterbi Decoding

There is one drawback of producing binary activity masks, as shown in Fig. 4: It is possible that some comments may not be assigned to any author and that some comments may be assigned to multiple authors. For instance, the second comment in Fig. 4 is not assigned to any author and the last comment is assigned to two authors. As pointed out in the Introduction, we do not expect multi-authored comments. We address this problem in stage ④ by transforming the estimated binary activity masks of the optimal permutation into multiclass labels and by performing sequence labeling based on a CRF model. As described in (Ma and Hovy, 2016) and confirmed by our results, it is beneficial to jointly decode and refine the best chain of authorship labels for a given temporal discussion.

We firstly transform the estimated binary activity masks in Eq. (6) of the optimal permutation in Eq. (7) into multiclass labels,

$$\boldsymbol{y}^{\text{(CRF)}}_{C\times 1} = \text{BinaryToMulticlass}\left(\boldsymbol{y}^{(\phi^{*}_{1})}_{C\times 1}, \dots, \boldsymbol{y}^{(\phi^{*}_{A_{\text{max}}})}_{C\times 1}\right). \tag{9}$$

As a result, $\boldsymbol{y}^{\text{(CRF)}}_{C\times 1}$ in Eq. (9) contains the author numbers (e.g. $[1, 2, 1, 2, 3, 1]$ as in Fig. 4) in order of appearance w.r.t. the re-ordered activity masks for the given discussion. By rewriting the labels $\boldsymbol{y}^{\text{(CRF)}}_{C\times 1} = [y^{\text{(CRF)}}_{1}, \dots, y^{\text{(CRF)}}_{C}]^{T}$ and the concatenated hidden cell states $\boldsymbol{H}_{C\times A_{\text{max}}\cdot D} = [\boldsymbol{H}^{(1)}_{C\times D}, \dots, \boldsymbol{H}^{(A_{\text{max}})}_{C\times D}] = [\boldsymbol{h}_1, \dots, \boldsymbol{h}_C]^{T}$ from Eq. (5), we can use the $A_{\text{max}} \cdot D$-dimensional vectors $\boldsymbol{h}_c$ to define a probabilistic layer of the author-comment alignment,

$$p(y_1^{(\text{CRF})}, \ldots, y_C^{(\text{CRF})} | \boldsymbol{H}) \\ = \frac{1}{Z(\boldsymbol{H})} \prod_{c=1}^{C} \psi_c\big(y_{c-1}^{(\text{CRF})}, y_c^{(\text{CRF})}, \boldsymbol{h}_c\big), \tag{10}$$

where the normalization factor in Eq. (10) is defined as $Z(\boldsymbol{H}) = \sum_{\boldsymbol{y}^{(\text{CRF})}} \prod_{c=1}^{C} \psi_c\big(y_{c-1}^{(\text{CRF})}, y_c^{(\text{CRF})}, \boldsymbol{h}_c\big)$ and $\psi_c(y', y, \boldsymbol{h}_c) = \exp\big(\boldsymbol{w}_{y'y}^T\, \boldsymbol{h}_c + b_{y'y}\big)$ represents the potential function. The terms $\boldsymbol{w}_{y'y}^T$ and $b_{y'y}$ are trainable weights corresponding to the label pair (y', y), respectively. The loss function is then given by the negative log-likelihood of the multiclass ground-truth hypothesis,

$$\mathcal{L}_{\text{CRF}} = -\log p\big(y_1^{(\text{CRF})}, \ldots, y_C^{(\text{CRF})} | \boldsymbol{H}\big), \tag{11}$$

and provides the second loss term in our overall loss function. The training is efficiently implemented with the forward-algorithm. During inference, we adopt the Viterbi decoder to search for the most likely label sequence,

$$\widehat{\boldsymbol{y}}_{C\times 1}^{(\text{CRF})} = \underset{y_1^{(\text{CRF})} \ldots y_A^{(\text{CRF})}}{\arg\max}\ p\big(y_1^{(\text{CRF})}, \ldots, y_C^{(\text{CRF})} | \boldsymbol{H}\big). \tag{12}$$

The pseudocodes for training and inference are summarized in Algorithms 1 and 2.

4 Experimental setup

The experimental setup discussion is divided into two parts, i.e., for *authorship verification* and for *author diarization*.

4.1 Authorship Verification

As mentioned in Section 3.1, we trained a feature extraction model that is used to deliver the feature representations for the subsequent author diarization model. In our experiments, we compared two different models to perform the authorship verification task: ADHOMINEM (Boenninghoff et al., 2019; Boenninghoff et al., 2021b), and SROBERTA (Zhu and Jurgens, 2021).

In the first step, we trained the models on the same dataset splits for **Amazon reviews** and **Reddit** posts provided in (Zhu and Jurgens, 2021) to validate the reproducibility of the results. We also used the same discriminative evaluation metrics as in (Zhu and Jurgens, 2021). In the second step, we replaced the English-specific BERT and FastText models with German equivalents in the source code and trained these models on the new **Zeit-Online** dataset. To capture the calibration capacity, we additionally calculated the expected calibration error (ECE) and maximum calibration error (MCE) as described in (Guo et al., 2017). As a result, each model was trained and evaluated on datasets that were exclusively in one language.

Due to the lack of a comparable large English-language author diarization dataset, we included available English-language AV datasets to achieve the following: Showing that models' scores on social media data are comparable across languages and illustrating that the training adjustments we made have a consistent impact.

4.2 Author Diarization

After fine-tuning the feature extraction models on the **Zeit-Online** dataset (stage ①), we proceeded to train the author diarizer. We evaluated the performance of the author diarization model using metrics such as the diarization error rate (DER), the Jaccard error rate (JER), and the F1-score. While DER and JER are well-established and widely recognized in the area of speaker diarization, the F1-score has been extensively employed in various style change detection tasks. We aim to cater to a diverse audience by presenting all three metrics.

Prior to applying evaluation metrics, an optimal mapping between **reference authors** (ground truth) and **system authors** (estimated authors) needs to be established. The Hungarian algorithm (Kuhn, 1955) is employed to determine this optimal mapping, pairing each reference author with at most one system author. In all evaluation metrics described below, we assume that such a commensurate mapping has been performed first.

The first metric utilized is the Jaccard error rate (JER), derived from the Jaccard index, a measure that assesses the similarity between sets. Following the methodology in (Ryant et al., 2020), the author-specific JER is computed for each reference author concerning the corresponding system author. Formally, for each author a, we have

$$\text{JER}(a) = \frac{1}{A} \sum_{a=1}^{A} 1 - \frac{\big|\text{ref}(a) \cap \text{sys}(a)\big|}{\big|\text{ref}(a) \cup \text{sys}(a)\big|}, \tag{13}$$

where A is the total number of reference authors. The set ref(a) denotes the collection of all comment indices associated with author a, i.e., if the

```
Input: Features $\boldsymbol{F}_{C\times D}$, binary author labels $\boldsymbol{y}^{(a)}_{C\times 1}$ $\forall a \in \{1,\dots,A\}$.
Output: Loss $\mathcal{L}_{\text{CRF-BCE-PIT}}$
$\boldsymbol{X}_{C\times D} = \text{BiLSTM}\big(\boldsymbol{F}_{C\times D}\big)$                                   // Eq. (2)
for $a = 1,\dots,A_{max}$ do
    $\boldsymbol{H}^{(a)}_{C\times D} = \text{LSTM}\left(\begin{bmatrix}\boldsymbol{X}^T_{C\times D}\\ \text{Linear}\big(\widehat{\boldsymbol{y}}^{(a-1)}_{C\times 1}\big)^T_{C\times D}\end{bmatrix}^T, \boldsymbol{H}^{(a-1)}_{C\times D}\right)$      // Eq. (5)
    $\boldsymbol{z}^{(a)}_{C\times 1} = \text{Sigmoid}\big(\text{Linear}\big(\boldsymbol{H}^{(a)}_{C\times D}\big)_{C\times 1}\big)$            // Eq. (5)
    $\widehat{\boldsymbol{y}}^{(a)}_{C\times 1} = \big[\mathbf{1}(z^{(a)}_c > 0.5) | c = 1,\dots,C\big]$              // Eq. (6)
end
$\phi^\star = \underset{\phi\in\text{perm(A)}}{\arg\min}\left\{\sum_{c=1}^{C}\sum_{a=1}^{A}\text{BCE}\big(z^{(a)}_c, y^{(\phi_a)}_c\big)\right\}$          // Eq. (7)
$\boldsymbol{y}^{(\text{CRF})}_{C\times 1} = \text{BinaryToMulticlass}\left(\boldsymbol{y}^{(\phi^*_1)}_{C\times 1},\dots,\boldsymbol{y}^{(\phi^*_{A_{\max}})}_{C\times 1}\right)$       // Eq. (9)
$p\big(y^{(\text{CRF})}_1,\dots,y^{(\text{CRF})}_C|\boldsymbol{H}\big) = \text{CRF}\big(\boldsymbol{H}_{C\times A_{\max}\cdot D}, \boldsymbol{y}^{(\text{CRF})}_{C\times 1}\big)$      // Eq. (10)
$\mathcal{L}_{\text{CRF-BCE-PIT}} = -\log p\big(y^{(\text{CRF})}_1,\dots,y^{(\text{CRF})}_C|\boldsymbol{H}\big)$
        $+\sum_{c=1}^{C}\sum_{a=1}^{A}\text{BCE}\big(z^{(a)}_c, y^{(\phi^*_a)}_c\big) + \sum_{c=1}^{C}\sum_{a=A+1}^{A_{\max}}\text{BCE}\big(z^{(a)}_c, 0\big)$    // Eq. (8) and Eq. (11)
```

Algorithm 1: Loss computation for the author diarization framework.

```
Input: Features $\boldsymbol{F}_{C\times D}$.
Output: Estimated multiclass labels $\widehat{\boldsymbol{y}}_{C\times 1}$.
$\boldsymbol{X}_{C\times D} = \text{BiLSTM}\big(\boldsymbol{F}_{C\times D}\big)$                                   // Eq. (2)
for $a = 1,\dots,A_{max}$ do
    $\boldsymbol{H}^{(a)}_{C\times D} = \text{LSTM}\left(\begin{bmatrix}\boldsymbol{X}^T_{C\times D}\\ \text{Linear}\big(\widehat{\boldsymbol{y}}^{(a-1)}_{C\times 1}\big)^T_{C\times D}\end{bmatrix}^T, \boldsymbol{H}^{(a-1)}_{C\times D}\right)$      // Eq. (5)
    $\boldsymbol{z}^{(a)}_{C\times 1} = \text{Sigmoid}\big(\text{Linear}\big(\boldsymbol{H}^{(a)}_{C\times D}\big)_{C\times 1}\big)$            // Eq. (5)
    $\widehat{\boldsymbol{y}}^{(a)}_{C\times 1} = \big[\mathbf{1}(z^{(a)}_c > 0.5) | c = 1,\dots,C\big]$              // Eq. (6)
end
$\widehat{\boldsymbol{y}}^{(\text{CRF})}_{C\times 1} = \text{Viterbi}\big(\boldsymbol{H}_{C\times A_{\max}\cdot D}\big)$                            // Eq. (12)
```

Algorithm 2: Inference of the author diarization framework.

first, third, and sixth comment are from author 1, for example, then $\text{ref}(1) = \{1, 3, 6\}$. Similarly, the set $\text{sys}(a)$ denotes the collection of all comment indices that were estimated to belong to author a by our system. If there is no pairing for a reference author a, then $\text{JER}(a)$ is set to 1. The JER for a discussion is then the average of all author-specific JER values:

$$\text{JER} = \frac{1}{A}\sum_{a=1}^{A}\text{JER}(a). \tag{14}$$

The diarization error rate (DER) is the second metric, measuring the proportion of comments incorrectly attributed to an author. It is expressed as:

$$\text{DER} = 1 - \frac{1}{C}\sum_{a=1}^{A}\big|\text{ref}(a) \cap \text{sys}(a)\big|. \tag{15}$$

where C in Eq. (15) represents the number of comments in the discussion, and A denotes the number of reference authors.

As noted in (Ryant et al., 2020), JER and DER exhibit a strong correlation, as evident in Eq. (13) and (15). The range is given as $0 \leq \text{JER}, \text{DER} \leq 1$ by design. Unlike the DER, the JER assigns equal weight to each author's contribution, irrespective of their comment count. In discussions with numerous authors, the JER tends to be higher.

Our final metric, the macro-averaged F1-score, is also used in the PAN 2021 style change detection task, as documented by Zangerle et al. (2021).

5 Results

We conducted ten repetitions of model training. Results are reported in Table 3 for *authorship verification* and Table 4 for *author diarization*. The shown metrics are the *average* values from the ten training cycles, including respective standard deviations.

5.1 Authorship Verification

Table 3 provides a detailed look at the performance of the employed authorship verification methods.

Configuration	#Tokens	ACC (%) ↑	F1 ↑	AUC ↑	ECE ↓	MCE ↓
Amazon Reviews						
Zhu and Jurgens (2021)	$n = 100$	82.9	83.1	90.9	-	-
Our results (SRoBERTA)	$n = 100$	82.86 ± 0.05	82.98 ± 0.12	90.84 ± 0.05	-	-
SRoBERTA + re-sampling	$n = 100$	83.50 ± 0.06	83.74 ± 0.05	91.36 ± 0.08	-	-
SRoBERTA + re-sampling	$n = 500$	**86.44 ± 0.05**	**86.48 ± 0.10**	**93.88 ± 0.04**	-	-
Reddit Posts						
Zhu and Jurgens (2021)	$n = 100$	73.0	73.7	81.2	-	-
Our results (SRoBERTA)	$n = 100$	73.08 ± 0.12	73.72 ± 0.17	81.00 ± 0.11	-	-
SRoBERTA + re-sampling	$n = 100$	74.28 ± 0.10	74.12 ± 0.21	82.50 ± 0.09	-	-
SRoBERTA + re-sampling	$n = 500$	**76.40 ± 0.07**	**76.38 ± 0.08**	**84.80 ± 0.07**	-	-
Zeit-Online Comments						
AdHominem + re-sampling	$n = 500$	76.13 ± 0.13	77.14 ± 0.23	84.55 ± 0.07	**0.46 ± 0.16**	**0.99 ± 0.34**
SGBERT + re-sampling	$n = 500$	**80.01 ± 0.29**	**79.87 ± 0.33**	**88.62 ± 0.31**	14.59 ± 0.34	20.47 ± 0.39

Table 3: Authorship verification results on the datasets used in (Zhu and Jurgens, 2021) and on the Zeit-Online verification test set described in Table 1.

Method	Features	DER ↓	JER ↓	F1 ↑
Zeit-Online Comments (small)				
Diarizer	GBERT (Chan et al., 2020)	21.32 ± 0.25	23.83 ± 0.42	83.14 ± 0.35
Diarizer	AdHominem (Boenninghoff et al., 2021a)	20.79 ± 0.26	22.83 ± 0.28	83.66 ± 0.31
Diarizer	SGBERT	**16.85 ± 0.35**	**19.79 ± 0.36**	**85.80 ± 0.24**
Diarizer + CRF	GBERT (Chan et al., 2020)	20.73 ± 0.58	22.81 ± 0.62	83.70 ± 0.49
Diarizer + CRF	AdHominem (Boenninghoff et al., 2021a)	18.94 ± 0.54	21.49 ± 0.32	84.79 ± 0.27
Diarizer + CRF	SGBERT	**15.57 ± 0.30**	**20.05 ± 0.55**	**85.60 ± 0.41**
Zeit-Online Comments (large)				
Diarizer	GBERT (Chan et al., 2020)	21.44 ± 0.25	23.94 ± 0.45	83.11 ± 0.21
Diarizer	AdHominem (Boenninghoff et al., 2021a)	20.97 ± 0.33	23.27 ± 0.49	83.54 ± 0.32
Diarizer	SGBERT	**18.83 ± 0.40**	**21.59 ± 0.36**	**84.91 ± 0.31**
Diarizer + CRF	GBERT (Chan et al., 2020)	20.44 ± 0.36	22.82 ± 0.35	83.67 ± 0.23
Diarizer + CRF	AdHominem (Boenninghoff et al., 2021a)	19.14 ± 0.23	21.67 ± 0.64	84.63 ± 0.45
Diarizer + CRF	SGBERT	**17.52 ± 0.20**	**20.93 ± 0.38**	**85.30 ± 0.29**

Table 4: Diarization results for various configurations of the proposed system.

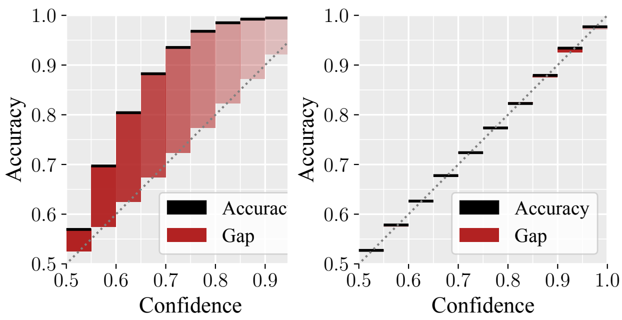
The initial two rows in Table 3 for each of the **Amazon** and the **Reddit** datasets present performance scores as reported in (Zhu and Jurgens, 2021), using identical hyper-parameters for our evaluation runs. In particular, our results are closely aligned with the discriminative scores reported regarding accuracy, F1-score, and AUC.

Additional experiments involve the re-sampling strategy proposed in (Boenninghoff et al., 2019). All datasets contain fixed pairs but provide author identifiers. The idea is to increase the size (only of the training set) by dissembling all predefined pairs and re-sampling new same-author and different-author pairs in each epoch. This ensures a diverse and representative sample of possible author combinations and that the model learns to generalize well across various types of author pairs.

As we can see, the re-sampling strategy leads to significant improvements, particularly when the input token number is increased to $n = 500$, resulting in the highest accuracy, F1-score, and AUC compared to all other configurations.

In the final rows of Table 3, we extended our experiment to the **Zeit-Online** dataset, juxtaposing our results for the two approaches AdHominem and SGBERT. Both methods exhibit competitive results. Remarkably, AdHominem displays well-calibrated predictions, as evidenced by an ECE of 0.46 and MCE of 0.99. Conversely, SGBERT surpasses AdHominem with superior scores in all discriminative metrics.

Fig. 5 shows the reliability diagrams for both models. In Fig. 5a, the red gaps between accuracy and confidence signify miscalibration in the SG-



(a) SGBERT. (b) AdHominem.

Figure 5: Reliability diagrams depict the average confidence and accuracy on the Zeit-Online authorship verification dataset. Darker red bars correspond to bins with a higher number of trials (Guo et al., 2017).

BERT model. This observation suggests that the model consistently provides under-confident predictions, as the accuracy consistently exceeds the confidence. In contrast, Fig. 5b illustrates that for AdHominem, accuracy and confidence are closely aligned, if not equal.

5.2 Author Diarization

Table 4 displays diarization results for two subsets of the **Zeit-Online** dataset, denoted as *small* and *large* (see Table 1). Recall that the *small* dataset includes discussions with up to four contributors, whereas the *large* dataset comprises discussions with up to six contributors. For both **Zeit-Online** datasets, we compare three feature extractors: diarizer with raw GBERT features, diarizer with AdHominem features, and diarizer with SGBERT features. We observe that SGBERT exhibits the best performance, showcasing substantial reductions in DER and JER, along with an enhanced F1-score compared to other methods. The incorporation of the CRF-layer (stage ④ in Section 3.3) further boosts the results, with SGBERT + CRF surpassing other configurations, achieving the lowest DER, JER, and the highest F1-score.

Table 5 presents diarization results given the SGBERT + Diarizer + CRF configuration only. The results are sorted w.r.t. the number of authors per discussion. Unsurprisingly, we observe the following trend: As the number of authors increases, the DER and JER also increase. At the same time, the F1-score decreases, indicating greater difficulty in accurately distinguishing between authors in discussions with multiple contributors.

#Authors	**Ratio**	**DER** ↓	**JER** ↓	**F1** ↑
Zeit-Online Comments (small)				
$A = 1$	6.8%	13.53 ± 1.66	13.59 ± 1.66	90.94 ± 1.11
$A = 2$	28.1%	12.41 ± 0.35	12.70 ± 0.44	90.89 ± 0.40
$A = 3$	35.4%	15.35 ± 0.36	18.49 ± 0.67	86.51 ± 0.54
$A = 4$	29.7%	19.24 ± 0.67	30.43 ± 0.99	78.15 ± 0.78
Zeit-Online Comments (large)				
$A = 1$	4.8%	12.81 ± 1.56	12.87 ± 1.56	91.40 ± 1.03
$A = 2$	19.9%	12.60 ± 0.68	13.00 ± 0.55	90.59 ± 0.45
$A = 3$	25.1%	17.17 ± 0.46	18.95 ± 0.90	86.17 ± 0.75
$A = 4$	21.1%	21.36 ± 0.36	25.00 ± 0.68	82.29 ± 0.55
$A = 5$	15.8%	21.70 ± 0.22	27.29 ± 0.39	81.17 ± 0.37
$A = 6$	13.4%	16.25 ± 0.53	26.44 ± 2.44	82.41 ± 0.90

Table 5: Diarization results vs. number of authors per discussion for the SGBERT + Diarizer + CRF configuration. The **Ratio** value indicates the respective percentage of discussions with the commensurate author number.

5.3 Overall Analysis

The results indicate that SGBERT, in combination with the CRF-layer, provides a powerful tool for the diarization of our **Zeit-Online** dataset. The approach consistently outperforms other configurations, showcasing its effectiveness in segmenting and attributing authors. The incorporation of the CRF-layer as a post-processing step enhances the temporal coherence of the diarization output, contributing to improved overall performance.

6 Conclusion

Digital text forensics is crucial for verifying the authenticity and integrity of electronic texts, aiding in investigations related to cybercrime, fraud, and misinformation. In conclusion, the diarization experiments underscore the significance of feature selection and show promising results for the proposed author-conditional encoder-decoder framework, enhanced by a Conditional Random Field with Viterbi decoding. These findings contribute to the ongoing efforts to enhance author diarization techniques. Future directions of our work include end-to-end training of the proposed system.

Acknowledgements

Our work was partially funded by the German Federal Ministry of Research and Education (BMBF) in conjunction with the *noFake* project, Award Identifier F16KIS1519 and the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – 462287308 (HU 3005/2-1).

7 Limitations

Our author-conditional encoder-decoder diarization framework, as demonstrated by our findings, shows promising results and addresses various challenges in attributing text segments to different authors. Nevertheless, it is essential to acknowledge the limitations of the framework:

Dependency on feature extraction models

Our framework is dependent on the performance of the authorship verification models (ADHOMINEM and GBERT) for feature extraction. Inaccuracies or biases in these models have the potential to propagate to the diarization process, resulting in erroneous author attributions.

Assumption of single authorship

In this work, we assume that each comment is written by a single individual. However, in other scenarios, text segments may involve multiple authors collaborating, leading to inaccuracies in attribution. Stages ① to ③ of our framework, as illustrated in Fig. 4, generally output activity masks that facilitate multi-author assignments.

Permutation-invariant training

The permutation-invariant training process poses a challenge, as finding the optimal author permutation can be computationally expensive, particularly in discussions involving a large number of participants. One workaround is segmenting longer discussions into shorter ones to facilitate computationally tractable permutation-invariant training.

Sensitivity to different text genres

The effectiveness of the extracted feature representation may vary depending on the text genre. It is possible that the presented framework may not generalize well to diverse datasets with different linguistic characteristics.

Dependency on language-specific models

We acknowledge that the author diarization dataset used in this work is in German and that the author diarization framework was not tested in other languages, including English. We leave this aspect for future investigation.

Missing state-of-the-art comparison

Our baseline for the diarization task employs GBERT as a feature extractor, without further finetuning on the AV dataset. This method was chosen because of the absence of publicly available, state-of-the-art models specifically tailored for this task. Current research, including the PAN competition, primarily addresses style change detection, which is not directly applicable to author diarization.

8 Ethical Statement

In conducting this research, constructing the dataset, and developing the author diarization framework described in this paper, we have been mindful of the ethical implications and responsibilities associated with working with textual data, particularly in the realm of authorship analysis. We recognize the potential for misuse of such technology and are committed to maintaining ethical standards throughout our work.

We emphasize the following potential implications of identifying individuals:

- **Unintended identification:** If the AD framework is used on text data containing personal or sensitive information, it may inadvertently reveal the identity of users if the system can link specific writing styles or content to known individuals.
- **Cross-referencing with other datasets:** If diarization results are combined with other datasets, there is a risk of re-identifying users even if their names are not explicitly mentioned in the text.

To mitigate these risks, we have implemented the following strategies:

- Our dataset is semi-automatically anonymized by replacing personally identifiable information like author names or emails. Additionally, we will not provide any person-based meta-data, as well as the URLs linking to the original data source.
- To further mitigate risks, we publish the dataset on request only. This controlled access allows us to approve requests based on ethical considerations, ensuring that the data is used appropriately and responsibly.

We respect the privacy and confidentiality of the individuals whose comments are used in our research. Therefore, we have taken manual and automated measures to anonymize and protect sensitive information to the best of our ability.

We are open to feedback and constructive criticism to improve the ethical integrity of our work.

References

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019. Explainable Authorship Verification in Social Media via Attention-based Similarity Learning. In *IEEE International Conference on Big Data*, pages 36–45.

Benedikt Boenninghoff, Dorothea Kolossa, and Robert M. Nickel. 2021a. Self-calibrating neural-probabilistic model for authorship verification under covariate shift. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 145–158, Cham. Springer International Publishing.

Benedikt Boenninghoff, Robert M. Nickel, and Dorothea Kolossa. 2021b. O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification—Notebook for PAN at CLEF 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Branden Chan, Stefan Schweter, and Timo Möller. 2020. German's next language model. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yusuke Fujita, Shinji Watanabe, Shota Horiguchi, Yawen Xue, Jing Shi, and Kenji Nagamatsu. 2020. Neural speaker diarization with speaker-wise chain rule. *ArXiv*, abs/2006.01796.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *34th International Conference on Machine Learning*, volume 70, pages 1321–1330. PMLR.

Shunichi Ishihara, Sonia Kulkarni, Michael Carne, Sabine Ehrhardt, and Andrea Nini. 2024. Validation in forensic text comparison: Issues and opportunities. *Languages*, 9(2).

Mike Kestemont, Enrique Manjavacas, Ilia Markov, Janek Bevendorff, Matti Wiegmann, Efstathios Stamatatos, Martin Potthast, and Benno Stein. 2020. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

Mike Kestemont, Efstathios Stamatatos, Enrique Manjavacas, Janek Bevendorff, Martin Potthast, and Benno Stein. 2021. Overview of the Authorship Verification Task at PAN 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Harold W. Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1064–1074, Berlin, Germany. Association for Computational Linguistics.

Andrei Manolache, Florin Brad, Antonio Barbalau, Radu Tudor Ionescu, and Marius Popescu. 2024. Veridark: a large-scale benchmark for authorship verification on the dark web. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Neville Ryant, Kenneth Church, Christopher Cieri, Jun Du, Sriram Ganapathy, and Mark Liberman. 2020. Third dihard challenge evaluation plan.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2023. Valla: Standardizing and benchmarking authorship attribution and verification through empirical evaluation and comparative analysis. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 649–660, Nusa Dua, Bali. Association for Computational Linguistics.

Dong Yu, Morten Kolbæk, Z. Tan, and Jesper Højvang Jensen. 2017. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 241–245.

Eva Zangerle, Maximilian Mayerl, , Martin Potthast, and Benno Stein. 2021. Overview of the Style Change Detection Task at PAN 2021. In *CLEF 2021 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2022. Overview of the Style Change Detection Task at PAN 2022. In *CLEF 2022 Labs and Workshops, Notebook Papers*. CEUR-WS.org.

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2023. Overview of the Multi-Author Writing Style Analysis Task at PAN 2023. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023)*, volume 3497 of *CEUR Workshop Proceedings*, pages 2513–2522.

Jian Zhu and David Jurgens. 2021. Idiosyncratic but not arbitrary: Learning idiolects in online registers reveals distinctive yet consistent individual styles. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 279–297, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Appendix

A.1 Dataset construction

As described in Section 2, we collected two large data sets comprised of comments from the discussion forums of the articles published from 2005 to 2021 in the German Zeit-Online newspaper.[8]

We created both, small and large data splits to easily investigate the scalability and robustness of our model. The construction of the dataset was conducted as follows:

1. We extracted the plain text of the collected HTML files that contained the discussions and performed text normalization, including Unicode normalization, whitespace normalization, and normalization of quotation marks.
2. We discarded discussions with comments that consisted of less than 10 tokens, determined using the German BERT model (Chan et al., 2020).
3. We removed discussions in which single comments were written in languages other than German or were modified or deleted by the Zeit-Online moderator team.
4. Longer discussions were randomly partitioned into shorter ones. The distribution of the number of unique authors per number of comments in the discussions is provided in Fig. 7.
5. We collected all comments from removed discussions (in steps 2 and 3) for the authorship verification dataset, verified that this comment is not part of any discussion in the author diarization dataset, and applied the first three steps.
6. We replaced URLs, author names, emails, and direct quotes with placeholders ([URL], [NAME], [EMAIL], [QUOTE]).
7. Finally, we split each dataset into 70%, 15%, and 15% portions to obtain the training, development, and test sets.
8. For ADHOMINEM, we utilized the Tokenizer SoMaJo[9], specialized for Web text, to segment each comment into sentences and each sentence into tokens. We used the German version of FastText[10] to initialize the token-to-embedding matrix.

	Tokens			Characters		
	Train	Dev	Test	Train	Dev	Test
Mean	89	90	88	422	425	414
Std	79	80	80	377	380	381
Median	63	64	62	298	300	290
Min	10	10	10	10	10	10
Max	8,062	7,130	19,239	38,847	35,225	91,535

Table 6: Statistics of the corpus for the **authorship verification** task

	Tokens			Characters		
	Train	Dev	Test	Train	Dev	Test
Mean	92	93	93	440	440	440
Std	79	79	80	379	379	380
Median	67	67	67	316	315	315
Min	10	10	10	12	12	12
Max	2,966	2,168	2,202	14,037	10,177	9,473

Table 7: Statistics of the corpus for the **authorship diarization** task.

B Dataset assessment

In this Section, we evaluate the dataset to determine its quality and suitability for author diarization.

Tables 6 and 7 show descriptive statistics of the number of tokens (using the GBERT tokenizer) and characters in the authorship diarization and authorship verification datasets.

Both datasets comprise a small range of text lengths, with mean token counts of 88 and 93 and constant standard deviations (std), demonstrating that training, test, and development sets were carefully crafted to contain texts with similar statistical properties. The median values are consistently lower than the means, showing a right-skewed distribution, where most comments are relatively short, but a few very long comments raise the average. Both tasks show a minimum of 10 tokens and character counts around 10-12, pointing to the inclusion of very short texts.

Figure 6 presents the number of discussions in the (large) authorship diarization dataset from different perspectives: The top histogram compares the number of discussions to the number of comments, indicating that the dataset is evenly distributed in terms of comment counts. The middle image shows the distribution of discussions with respect to the number of unique authors. Our analysis reveals that the majority of the discussions involve three to six commenters, indicating a rich interaction landscape. The bottom histogram displays the counts of style changes (author turns) occurring

[8] https://www.zeit.de

[9] https://github.com/tsproisl/SoMaJo

[10] https://fasttext.cc/docs/en/crawl-vectors.html

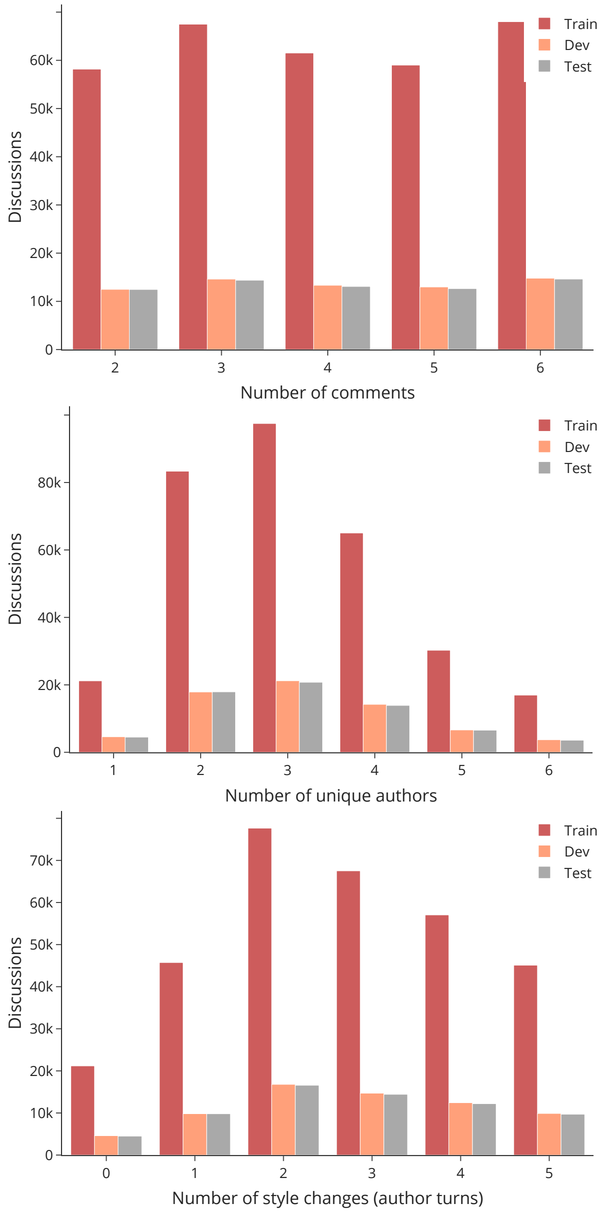


Figure 6: Number of comments (top), unique authors (middle), and style changes (bottom) in the discussions of the (large) author diarization dataset for the train, dev, and test sets.

between comments during a discussion. Again, our dataset shows that the majority of discussions exhibit a high level of dynamic author participation.

Figure 7 provides a more detailed view compared to the previous aggregated histogram in Fig. 6 (top) and presents histograms of the discussion counts regarding the number of unique authors. The figure includes several subplots, each corresponding to a different number of comments C in the discussions. The subplots display data for $C = 2$ through $C = 10$. In each plot, the x-axis represents the number of unique authors participating in a discussion while the y-axis shows the number of discussions. Each bar in the plots corresponds to a count of discussions with a specific number of unique authors. The histograms show that although a significant number of discussions with fewer comments involve only one or two commenters, most discussions feature a more diverse group of participants.

The dataset statistics emphasize that our author diarization dataset is challenging. It involves managing discussions where comments are short, less stylized, and multiple authors contribute frequently and interchangeably on the same topic. As discussed in (Kestemont et al., 2020), traditional authorship verification methods struggle in such scenarios and they can be misled by topic similarity. In our dataset, participants discuss the same topic, which requires diarization techniques to disregard topical content to ensure accurate attribution in these challenging environments.

C Training procedure and inference

Our training and inference follows the pseudo-code scheme in Algorithm 1 and 2. Training our model with our data typically requires a single GPU with 24GB of memory (NVIDIA RTX A5000). On average, and depending on early-stopping, this training process can be completed within two weeks.

We made minimal changes to the original authorship verification source codes, focusing mainly on hyper-parameter tuning and adapting specific parts to train the models on our dataset, i.e., using language-specific BERT and FastText models.

The authorship verification results in Table 1 are based on the fixed train, dev, and test splits for all datasets. We randomly initialized the models (i.e., all weights for ADHOMINEM and the AttentionPool layer for SGBERT (Zhu and Jurgens, 2021)) ten times, trained each model, and averaged the metrics returned from the test set. The results demonstrate the stability of the model training, indicated by the consistent performance across different runs.

The baseline for the author diarization framework used in our experiments is the pretrained GBERT model used as a feature extractor, without fine-tuning on the AV dataset. Next, we trained the ADHOMINEM and SGBERT models using the AV dataset, enabling them to serve as input features for the author diarizer. We randomly initialized the weights of the author diariza-

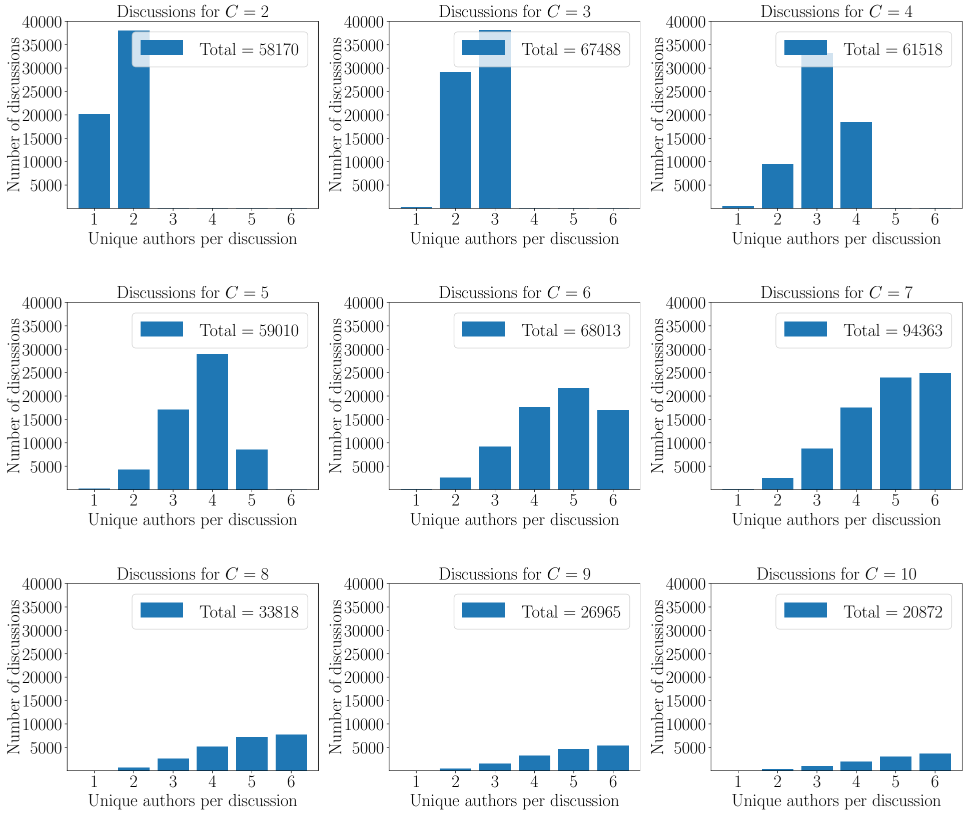


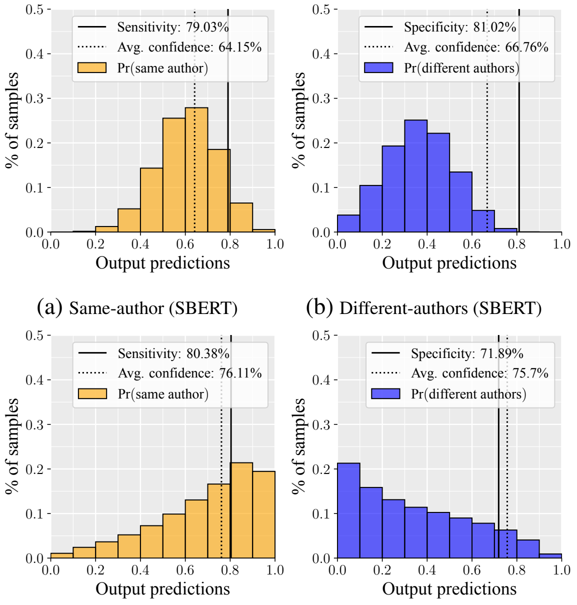
Figure 7: Distribution of the number of unique authors per number of discussion comments.

tion model and commenced training. This process yielded ten trained author diarization models for each: the baseline model (GBERT), ADHOMINEM, and SGBERT.

D Calibration of AV models

Fig. 8 shows the posterior histograms including the averaged confidence as well as the sensitivity and specificity for both methods, SGBERT and ADHOMINEM. All confidence values lie within the interval $[0.5, 1]$, since authorship verification represents a binary classification task in this paper. Hence, to obtain confidence scores, the output scores are transformed w.r.t. to the estimated authorship label, showing Pr(same author) if m_2: $\{\boldsymbol{f}_1, \boldsymbol{f}_2\} \longrightarrow p = 1$ and Pr(different authors) if m_2:$\{\boldsymbol{f}_1, \boldsymbol{f}_2\} \longrightarrow p = 0$.

By comparing the histograms for ADHOMINEM and SGBERT, we see that these models return different output score distributions. While SGBERT follows a bell shaped distribution, with peaks around 0.6 and 0.4, which are the pre-defined thresholds mentioned in (Zhu and Jurgens, 2021), ADHOMINEM tries to align the scores with accuracy on average.



(a) Same-author (SBERT)

(b) Different-authors (SBERT)

(c) Same-author (AdHominem)

(d) Different-authors (AdHominem)

Figure 8: Posterior histograms for both methods (True positive rates or sensitivity, true negative rates or specificity).