Machine Learning with Applications 25 (2026) 100943

Contents lists available at ScienceDirect

Machine Learning with Applications

journal homepage: www.elsevier.com/locate/mlwa

Cross-domain authorship verification with feature interaction networks: Evaluating no-holdout and holdout protocols

Britt van Leeuwen [a,b,*], Sandjai Bhulai [b], Rob van der Mei [a,b]

[a] Centrum Wiskunde en Informatica (Stochastics group), Science Park 123, Amsterdam, Netherlands
[b] Vrije Universiteit Amsterdam, Boelelaan 1111, Amsterdam, Netherlands

ARTICLE INFO

Keywords:
Authorship verification
Cross-domain
Stylometric features
RoBERTa embeddings
Hard negatives
Domain generalization

ABSTRACT

Authorship verification (AV) across communication domains is a challenging task, as models must determine whether two texts are written by the same author despite variations in style, topic, and modality. In this work, we adopt a hybrid approach that combines contextual embeddings from RoBERTa with handcrafted stylometric features to capture both high-level semantic patterns and low-level stylistic cues. We evaluate our method using a cross-domain setup, considering both same-domain and strictly cross-domain pairs, including challenging "hard negatives" where authors write in markedly different styles across domains. To explicitly assess generalization to unseen domains, we further employ a leave-one-domain-out strategy, in which entire domains are excluded during training and used only for testing. Our experiments span multiple domain pairs including emails, text messages, essays, interviews, and speech-to-text data. Results show that the adopted approach consistently outperforms state-of-the-art research on comparable domain pairs, achieving higher overall verification scores while maintaining strong performance across individual metrics, improving the overall score over PAN 2022 from 0.600 to 0.654.

Our findings demonstrate that combining contextual and stylistic representations, together with hard negative sampling, enables robust generalization across heterogeneous text types. Moreover, the leave-one-domain-out evaluation highlights the model's ability to handle previously unseen domains, while analysis of same-domain test pairs shows that these representations also transfer to within-domain verification without explicit training on such pairs. This work provides a comprehensive evaluation of cross-domain AV under realistic conditions and offers insights into building more robust authorship verification systems.

1. Introduction

1.1. Motivation

Authorship verification (AV) is the task of determining whether two texts were written by the same author, independent of their content. This problem arises in a variety of real-world scenarios, including plagiarism detection, forensic linguistics, online fraud investigation, and cross-platform identity analysis. In many of these applications, investigators must determine whether texts originating from different sources or platforms can be attributed to the same individual, often without prior knowledge of the author. For example, forensic analysts may need to link threatening emails to anonymous posts on social media, or connect fraudulent marketplace messages to communication in email or messaging applications. In such settings, evidence is rarely collected from a single platform, meaning that decisions must be made across heterogeneous and often noisy communication channels. As digital communication continues to expand across platforms such as emails, social media, forums, and messaging applications, writing style becomes inherently distributed across domains, making robust cross-domain authorship verification essential for practical applications.

Unlike authorship classification, which assumes a closed set of known authors, AV must generalize to previously unseen authors. This distinction makes AV fundamentally more challenging: instead of learning author-specific classes, models must learn patterns that characterize writing style itself. In addition, texts used for verification may originate from different domains, such as formal essays, informal messages, or spoken transcripts. These domain differences introduce variations in vocabulary, structure, and length, making it difficult for models to distinguish stylistic signals from domain-related characteristics.

* Corresponding author at: Centrum Wiskunde en Informatica (Stochastics group), Science Park 123, Amsterdam, Netherlands.
E-mail addresses: b.e.van.leeuwen@cwi.nl (B. van Leeuwen), s.bhulai@vu.nl (S. Bhulai), rob.van.der.mei@cwi.nl (R. van der Mei).

https://doi.org/10.1016/j.mlwa.2026.100943
Received 15 April 2026; Received in revised form 28 May 2026; Accepted 23 June 2026
Available online 29 June 2026

1.2. Related work

1.2.1. Stylometry and feature-based authorship analysis

Stylometry, the quantitative study of writing style, forms the foundation for many authorship-analysis tasks, including authorship attribution, verification, and profiling (Cammarota et al., 2024; Neal et al., 2017). These tasks aim to capture the unique stylistic fingerprints of authors to identify authorship, detect textual similarities, or infer author characteristics. Traditional approaches represent style using lexical, syntactic, semantic, structural, and domain-specific features. Early work relied on controlled corpora with large training sets and a limited number of candidate authors, achieving good performance under these constrained conditions (He et al., 2024; Misini et al., 2022; Neal et al., 2017; Stamatatos, 2009).

With the proliferation of online texts, authorship analysis has shifted toward more realistic settings involving noisy, heterogeneous, and unbalanced datasets (Luyckx & Daelemans, 2008). Machine learning and natural language processing techniques — including probabilistic models, distance-based methods, and compression approaches — have been increasingly applied to capture stylistic patterns in such complex scenarios (Misini et al., 2022; Neal et al., 2017). However, performance is strongly affected by dataset size, text length, and the number of candidate authors, making reliable analysis challenging in many practical applications.

Feature extraction remains critical for capturing an author's style, yet there is no consensus on an optimal feature set (Misini et al., 2022; Neal et al., 2017). Low-level indicators such as word or character frequencies and syntactic statistics can effectively distinguish authors but fail to fully capture subtler stylistic nuances (Lagutina et al., 2019). Moreover, correlations between stylistic features and contextual factors such as genre or domain are not well understood, motivating research into richer representations that generalize across authors, domains, and textual contexts (Lagutina et al., 2019; Neal et al., 2017).

1.2.2. Neural and representation-based approaches

The advent of pretrained language models has transformed authorship analysis by enabling the extraction of contextualized embeddings that capture semantic and syntactic nuances (Barlas & Stamatatos, 2020; Futrzynski, 2021). Large-scale evaluation frameworks such as the PAN shared tasks (Webis Group, 2024) have further standardized experiments (Bevendorff et al., 2023, 2021; Kestemont et al., 2022). Results from these evaluations show that while neural models can achieve good performance, traditional stylometric methods based on n-grams remain competitive, particularly when training data is limited (Stamatatos et al., 2022; Tyo et al., 2022).

Lightweight convolutional neural networks (CNNs) have also been explored for authorship analysis and text classification tasks (Ammar Aouchiche et al., 2024; Khan et al., 2023). These models can efficiently capture local lexical and syntactic patterns while requiring fewer parameters than large transformer-based architectures. Although computationally efficient, CNN-based methods generally capture less contextual information than modern pretrained language models, which may limit their ability to generalize effectively in cross-domain AV settings.

Several studies use pretrained models for AV by training auxiliary classification tasks and then comparing text embeddings using similarity metrics (Futrzynski, 2021; Petropoulos, 2023; Tyo et al., 2021). These approaches can improve robustness in cross-domain scenarios, but they often depend on sufficient training data and carefully curated normalization corpora (Barlas & Stamatatos, 2020). Pairwise learning frameworks, such as contrastive learning, extend these methods by directly optimizing similarity relationships between text pairs and incorporating hard-negative sampling to distinguish between stylistically similar authors (Guo et al., 2023; Tyo et al., 2022).

Despite these advancements, AV remains challenging under realistic conditions. Cross-domain experiments from PAN show substantial drops in performance when texts originate from different discourse types, such as emails, essays, or text messages (Stamatatos et al., 2022). In some scenarios, simple n-gram baselines outperform sophisticated neural models, particularly when training data is sparse (Guo et al., 2023; Stamatatos et al., 2022). Similar trends are observed in cross-domain attribution, where performance decreases as domain differences increase or candidate sets grow (Kestemont et al., 2018).

1.3. Research gap

Taken together, these observations suggest that current methods, while powerful, struggle to capture stable stylistic signals across domains and discourse types. Traditional stylometric features may overlook subtle patterns, whereas neural embeddings often require abundant domain-specific data to generalize.

Pairwise learning and hard-negative sampling have shown promise in related areas such as authorship attribution in closed-set settings (Tyo et al., 2022) and other areas like computer vision (Guo et al., 2023), yet their integration AV, specifically in cross-domain scenarios, remains underexplored. This gap is partly due to the nature of commonly used benchmarks, such as PAN, where text pairs are predefined and the sampling procedure is not disclosed.

In particular, there is a need for methods that can effectively handle: (i) heterogeneous texts with varying length, style, and communicative purpose, (ii) limited training samples, and (iii) hard-to-distinguish authors whose writing styles are similar. Addressing these challenges motivates the present work.

1.4. Proposed approach

Recent advances in natural language processing enable the extraction of rich contextual embeddings from pretrained language models such as RoBERTa (Liu et al., 2019), capturing high-level syntactic and semantic patterns in text. However, embeddings alone may not fully reflect individual writing style. Traditional stylometric features complement these representations by encoding lexical and structural characteristics, and their combination provides a more comprehensive view of authorship (Hu et al., 2023; van Leeuwen et al., 2025).

In contrast to the usual PAN settings, our setting allows explicit control over pair construction, enabling the incorporation of hard negatives. This provides a novel opportunity to study their impact on cross-domain authorship verification. To better distinguish between similar authors, we incorporate a hard negative sampling strategy, encouraging the model to focus on subtle stylistic differences (Hermans et al., 2017; Tyo et al., 2021). These features are integrated within a feature interaction framework that models pairwise text similarity (van Leeuwen et al., 2025).

To address cross-domain generalization, we adopt a leave-one-domain-out (LODO) setting, where models are evaluated on domains not seen during training. This setup reflects realistic scenarios in which writing styles and text types vary, requiring the model to rely on domain-independent stylistic signals.

Our contributions are as follows:

1. We study authorship verification under a rigorous cross-domain and LODO evaluation protocol, providing a more realistic assessment of model generalization.
2. We introduce a hard negative sampling strategy tailored for cross-domain authorship verification, which improves discrimination between stylistically similar authors.
3. We provide an extensive empirical analysis of feature interaction-based authorship verification models under challenging cross-domain conditions.

1.5. Paper structure

The remainder of this paper is organized as follows. In Section 2, we describe the data and present the analysis that underpins our study. Section 3 details the methodology used, including the model design and feature extraction processes. The experimental setup, including training procedures, evaluation metrics, and implementation details, is presented in Section 4. Results of our experiments are reported in Section 5, followed by a discussion of the findings, their implications, limitations, and directions for future work in Section 6. Finally, Section 7 concludes the paper.

2. Data analysis

2.1. Dataset construction

For the purpose of this research, we use a dataset provided by PAN (PAN Lab, 2022). They created a novel dataset from a subset of the Aston 100 Idiolects Corpus in English. It is specifically designed for cross-domain authorship verification. We selected this dataset because it contains texts from multiple heterogeneous domains, including emails, essays, interviews, speech transcripts, and text messages, making it particularly suitable for evaluating domain generalization and cross-domain robustness. In addition, the PAN datasets are widely used benchmarks in authorship verification research, enabling direct comparison with existing approaches.

Other datasets commonly used in Authorship Verification research include earlier PAN corpora (Webis Group, 2024), the Blog Authorship Corpus (Schler et al., 2006), and Reddit- or review-based datasets. However, many of these primarily focus on single-domain settings or provide less domain diversity. The PAN 2022 dataset is therefore especially appropriate for the objectives of this work.

The reported key statistics from the preprocessed dataset can be found in Table 1. While the data is typically structured as pairs, for our project we received the raw data and construct the pairs ourselves. This is imperative for our second experiment. The raw dataset consists of 25,290 texts written by 112 distinct authors across multiple communication domains. On average, each author contributes 225.8 texts (SD = 25.62), with a median of 224 texts per author. The minimum and maximum number of texts per author are 148 and 298, respectively. This relatively narrow distribution ensures that no single author dominates the dataset while still providing sufficient material to construct robust author profiles.

We included Table 2 with the key statistics for our dataset splits and pairing. Compared to the PAN 2022 cross-discourse type dataset (PAN Lab, 2022), our dataset exhibits a similar balance between positive and negative pairs, with roughly 50% of pairs being authored by the same individual. However, the distribution across discourse types differs notably. While the PAN dataset emphasizes email–text message and essay–email pairs, our dataset contains a higher proportion of short-form communications such as text messages, and a larger number of rare combinations grouped under “other” pairs. Average text lengths also differ: essays in our dataset are slightly shorter, while text messages are comparable in length. These differences underscore the challenges posed by heterogeneous domain distributions and variable text lengths, highlighting the importance of our hard negative sampling strategy.

The distribution of texts across subcorpora is highly imbalanced, as shown in Table 3. Short-form communication types dominate the dataset: text messages constitute nearly half of all texts, followed by emails and interviews. Long-form or less frequent modalities such as essays, handwritten texts, and speech-to-text transcripts are comparatively scarce.

Text length varies substantially across the corpus and differs by subcorpus. Short-form communications, such as text messages, dominate the dataset, whereas longer forms like essays and speech-to-text transcripts are less frequent. Fig. 1 illustrates these patterns across subcorpora, highlighting both the skewness and the variation in text length, which can impact model training and evaluation.

In addition to textual content, the dataset contains metadata fields such as *id*, *thread id*, and *sequence*. These fields allow reconstruction of conversational structure and author-level grouping. Such metadata enables the construction of aggregated author profiles and supports cross-platform verification experiments.

Overall, the dataset is well-suited for cross-domain AV. However, several characteristics must be considered carefully: (i) the strong

Table 1
Key statistics of the dataset for the 2022 cross-discourse type AV task.

Subset	Training	Test
Author match (text pairs)		
positive (same author)	6132 (50.0%)	5239 (50.0%)
negative (different author)	6132 (50.0%)	5239 (50.0%)
Discourse type pairings (text pairs)		
Email–Text message	7484 (61.0%)	6092 (58.1%)
Essay–Email	1618 (13.2%)	1454 (13.9%)
Essay–Text message	1182 (9.6%)	1128 (10.8%)
Business memo–Email	1014 (8.3%)	900 (8.6%)
Business memo–Text message	780 (6.4%)	718 (6.9%)
Essay–Business memo	186 (1.5%)	186 (1.8%)
Discourse type text length (avg. characters)		
Essay	11,098	10,117
Email	2385	2323
Business memo	1255	1042
Text message	611	601

Table 2
Key statistics of the dataset splits for non-holdout.

Subset	Training	Test
Author match (text pairs)		
positive (same author)	6700 (50.0%)	3400 (50.0%)
negative (different author)	6700 (50.0%)	3400 (50.0%)
Discourse type pairings (text pairs)		
email–text message	5730 (42.8%)	2797 (41.1%)
interview–text message	2931 (21.9%)	1402 (20.6%)
email–interview	2156 (16.1%)	1088 (16.0%)
se query–text message	743 (5.5%)	455 (6.7%)
email–se query	656 (4.9%)	428 (6.3%)
interview–se query	368 (2.7%)	201 (3.0%)
essay–text message	123 (0.9%)	69 (1.0%)
email–essay	101 (0.8%)	53 (0.8%)
speech to text–text message	108 (0.8%)	47 (0.7%)
email–speech to text	77 (0.6%)	22 (0.3%)
handwritten–text message	64 (0.5%)	39 (0.6%)
other pairs	886 (6.6%)	773 (11.4%)
Discourse type text length (avg. characters)		
essay	10,986	10,361
speech to text	1979	2405
handwritten	1741	1953
interview	464	428
email	285	286
text message	36	37
se query	28	30

Table 3
Distribution of texts per subcorpus.

Subcorpus	Count
text message	12,398
email	7219
interview	3470
se query	1760
essay	186
speech to text	129
handwritten	128
se query memo	112

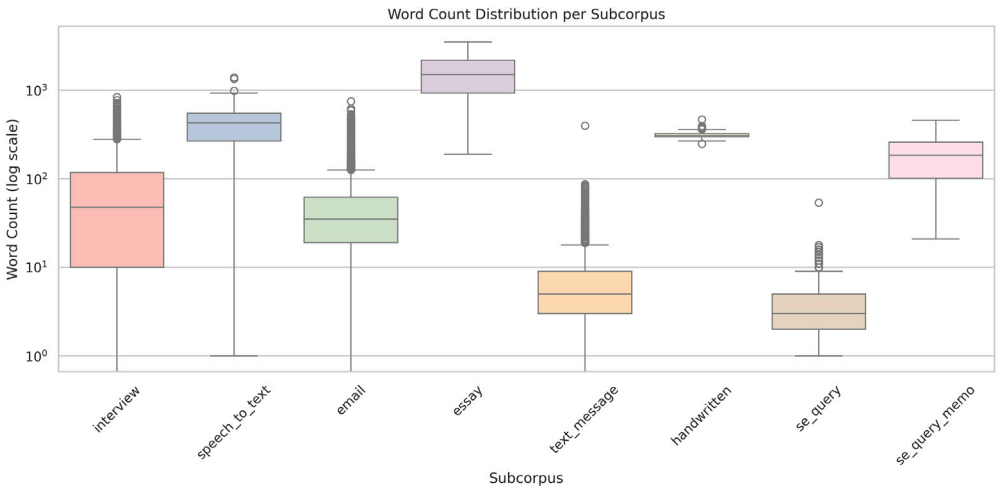

Fig. 1. Word-count distributions for each subcorpus. Each bar shows the range and frequency of text lengths within the domain.

imbalance across subcorpora, (ii) the predominance of short texts, and (iii) the structured thread information. These properties make the task both realistic and challenging.

2.2. Pair generation

To model AV, the corpus is transformed into a pairwise dataset. Each instance consists of two texts and a binary label indicating whether both texts were written by the same author.

2.2.1. Positive pair construction

Positive pairs are created by sampling two texts written by the same author. To encourage cross-domain generalization, pairs are only formed when the texts originate from different subcorpora. This constraint prevents the model from exploiting domain-specific artifacts and encourages learning stylistic patterns that transfer across communication modalities.

2.2.2. Hard negative pair construction

Negative pairs are formed by pairing texts written by different authors. Rather than sampling negatives randomly, we employ a *hard negative sampling* strategy to provide the model with more informative examples.

For each anchor text, we consider negative candidate texts from different domains. We first select approximately the 10% most similar cases. From this subset, one negative pair is randomly sampled. This ensures that selected negatives are challenging, as they are close to the decision boundary in embedding space, while still avoiding only identical or near-duplicate cases.

In addition, a portion of negatives is sampled randomly to maintain variability. All negative pairs are constructed across domains, ensuring that the model cannot rely on domain-specific cues. This combination of hard and random cross-domain negatives encourages the model to learn subtle stylistic distinctions rather than superficial differences.

Each positive pair is paired with one negative pair, resulting in a balanced dataset with equal numbers of positive and negative examples. The proportion of hard negative pairs versus the randomly selected ones is determined by the hard negative ratio, which is explored in Section 4. By using these carefully chosen hard negatives, the model develops a more discriminative representation of authorial style, improving its ability to generalize across unseen authors and domains.

The dataset contains a large pool of potential cross-domain combinations. On average, each author can generate tens of thousands of valid positive pairs, ensuring sufficient diversity when constructing training and evaluation splits.

3. Methodology

We build on our previous research in AV (van Leeuwen et al., 2025), where we demonstrated that a FIN achieves strong performance when combining contextual embeddings with carefully selected stylometric features. A FIN is a neural architecture designed to model interactions between heterogeneous feature types. Rather than treating features as independent inputs, it focuses on learning pairwise and higher-order feature dependencies through learned transformations. In the context of authorship verification, this allows the model to capture interactions between semantic representations (contextual embeddings) and low-level stylistic cues in a unified representation space.

3.1. Feature extraction

We adopt a dual-feature representation similar to our previous study (van Leeuwen et al., 2025). Each text is represented using two complementary feature types: contextual embeddings and stylometric features.

3.1.1. Contextual embeddings

We extract 768-dimensional contextual embeddings using RoBERTa-base. For each text, the final hidden layer is mean-pooled across tokens, producing a single fixed-length vector representation.

These embeddings capture high-level semantic and syntactic information, enabling the model to recognize stylistic similarity beyond surface-level lexical overlap.

3.1.2. Stylometric features

To incorporate traditional authorship cues, we compute a set of numeric style features capturing lexical diversity, character statistics, punctuation usage, and readability-related metrics.

Unlike contextual embeddings, these features explicitly model lower-level structural and stylistic patterns. They are precomputed once and stored to avoid repeated computational overhead.

3.1.3. Feature merging

For each text, the 768-dimensional embedding vector is concatenated with the stylometric feature vector into a unified representation. The resulting pairwise feature matrix contains 1555 columns, including metadata and features for both texts.

No normalization is applied during feature extraction. All scaling is performed exclusively on the training data within each split to prevent information leakage into validation or test sets.

3.2. Preliminary similarity analysis

As a sanity check, we computed cosine similarity between embedding representations of paired texts. Positive pairs exhibit a higher mean similarity (0.0197) compared to negative pairs (0.0066), indicating that same-author texts are, on average, more similar in embedding space.

However, the separation remains limited. A simple logistic regression classifier trained on similarity scores achieves an AUROC of 0.561, only modestly above chance level (0.5). This suggests that linear similarity alone is insufficient and motivates the use of more expressive neural architectures for cross-domain AV.

4. Experimental setup

4.1. Experimental data splits

We evaluate our approach in two experimental scenarios. The first uses a *no-holdout* setup, in which training, validation, and test splits are strictly author-disjoint but drawn from the same set of domains. The second follows a *holdout* protocol, leaving one domain completely unseen during training to assess cross-domain generalization. These experiments provide insights into different aspects of model generalization: author-level and domain-level.

4.1.1. No-holdout setting

The first experiment adopts a protocol similar to the PAN shared tasks. We use solely cross-domain pairs and all communication domains are allowed to appear in training, validation, and testing.

Authors are partitioned into three disjoint sets: training, validation, and testing. Texts written by a given author appear exclusively in one partition, ensuring that test authors are never seen during training.

Pairs are generated independently within each author partition, maintaining both positive (same-author) and negative (different-author) examples. This protocol isolates the challenge of verifying authorship for previously unseen authors, while leveraging information across all available domains. It serves as a baseline that reflects the conventional PAN-style experimental design.

4.1.2. Domain holdout

The second experiment introduces a stricter LODO protocol to evaluate domain generalization. In each run, one subcorpus is designated as the held-out domain and is completely excluded from the training and validation sets.

Training and validation pairs are constructed exclusively from the remaining domains. Test pairs are then generated such that one text always originates from the held-out domain, while the other text comes from a non-held-out domain. Both positive and negative pairs are included.

This setup assesses the model's ability to generalize to a communication domain that was never observed during training, capturing the added challenge of a more realistic cross-domain AV.

Together, these two protocols provide complementary perspectives: the no-holdout experiment evaluates generalization to unseen authors within familiar domains, while the domain holdout experiment evaluates robustness to unseen domains in addition to author verification.

4.2. Feature preprocessing

Prior to model training, all feature vectors are standardized using statistics computed exclusively from the training data. For each feature dimension, the mean and standard deviation are estimated on the training set and subsequently applied to transform the validation and test sets. This procedure ensures consistent feature scaling across splits while preventing information leakage from evaluation data.

All preprocessing operations are performed offline before model training to minimize runtime overhead. Consequently, the neural network receives as input two normalized feature vectors corresponding to the texts in each pair.

As a preliminary sanity check, a simple logistic regression classifier was trained on the difference between paired feature vectors. Although this baseline achieved performance slightly above chance, its limited discriminative ability indicates that linear similarity alone is insufficient to capture the complex stylistic relationships required for cross-domain AV.

4.3. Neural architecture

AV is formulated as a binary classification task over pairs of feature vectors. Let x_a and x_b denote the feature representations of two texts.

To explicitly model relationships between the two representations, the network constructs additional interaction features:

- Element-wise absolute difference: $|x_a - x_b|$
- Element-wise product: $x_a \odot x_b$

These interaction features are concatenated with the original vectors to produce a joint representation:

$$[x_a, x_b, |x_a - x_b|, x_a \odot x_b]$$

The resulting vector is passed through a feed-forward neural network consisting of three fully connected layers:

- Dense layer with 512 units and ReLU activation
- Dropout layer with rate 0.3
- Dense layer with 256 units and ReLU activation
- Dropout layer with rate 0.3
- Dense layer with 128 units and ReLU activation

The final layer is a single sigmoid unit that outputs the probability that the two texts were written by the same author.

Fig. 2 provides a visual overview of the complete AV pipeline, illustrating feature extraction, interaction, and classification.

4.4. Training procedure

The model is implemented using TensorFlow/Keras and optimized with the Adam optimizer (Kingma & Ba, 2017). Training is performed using mini-batch gradient descent with a batch size of 32.

The initial learning rate is set to 5×10^{-5}. During training, a *ReduceLROnPlateau* scheduler (Paszke et al., 2019) monitors the validation loss and reduces the learning rate by a factor of 0.5 when no improvement is observed for two consecutive epochs. The learning rate is bounded below by 10^{-6}.

Training proceeds for a maximum of 20 epochs. To prevent overfitting, early stopping is applied with a patience of four epochs based on validation loss. When triggered, the model restores the parameters corresponding to the best validation performance.

To account for potential imbalance between positive and negative pairs after sampling, class weights are computed from the training labels and incorporated into the loss function during optimization.

4.5. Effect of hard negative sampling

To determine an appropriate balance between random and hard negatives, we evaluate model performance across varying hard negative ratios.

Fig. 3 illustrates the impact of increasing the proportion of hard negatives on multiple evaluation metrics. As the ratio increases, recall improves, indicating that the model becomes more sensitive to

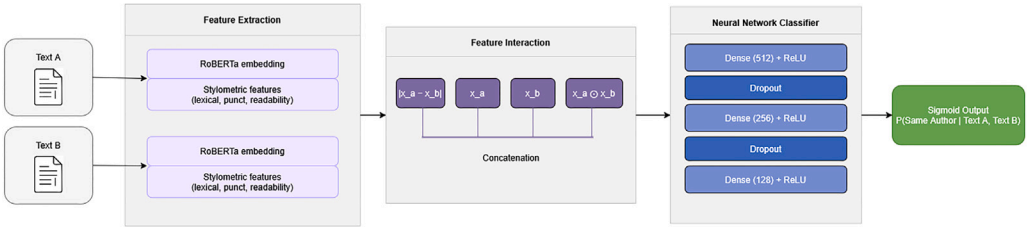

Fig. 2. Flowchart of the proposed AV method, showing feature extraction, feature interaction, concatenation, neural network classifier, and final prediction.

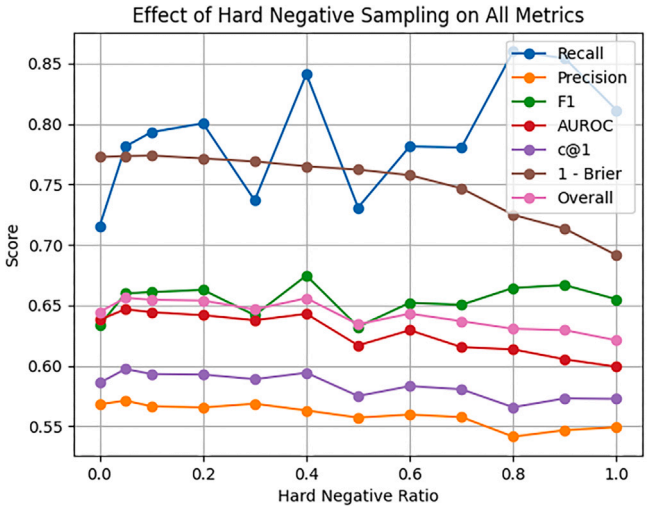

Fig. 3. Effect of hard negative sampling ratio on evaluation metrics. Moderate ratios yield the best Overall and F1 scores, while higher ratios increase false positives and reduce calibration.

detecting same-author pairs. However, this comes at the cost of reduced precision and overall calibration, leading to an increase in false positives.

Performance across combined metrics, such as F1 and the overall score, peaks at a moderate ratio of approximately 0.4. Beyond this point, performance degrades, suggesting that excessive reliance on hard negatives reduces the diversity of training signals.

Based on these observations, a hard negative ratio of 0.4 is used in all subsequent experiments.

4.6. Evaluation

Model performance is evaluated using several standard metrics for binary classification. To facilitate comparison with prior work in AV, results are reported using the PAN evaluation framework, which includes:

- AUROC
- $c@1$ (Peñas & Rodrígo, 2011)
- $F_{0.5u}$
- Brier score
- Overall score: the average of the scores above.

Performance is reported both overall and per domain pair, enabling analysis of how verification accuracy varies across different cross-domain scenarios.

5. Results

To evaluate the stability and convergence of our models, we monitored training and validation performance metrics throughout the optimization process. Across all configurations — including the non-holdout setting and the individual domain holdout experiments — the training loss consistently decreased while training accuracy and recall metrics improved, indicating successful model learning. Validation performance generally stabilized, demonstrating that the models learned to generalize effectively without significant overfitting. Detailed training and validation curves, including loss, accuracy, and recall for each holdout scenario, are provided in the Appendix, Figs. B.1 through B.9.

5.1. No-holdout

Table 4 summarizes overall performance across all domain pairs. The cross-domain-only model outperforms the PAN 2022 reference on most reported metrics, including AUROC, c@1, F0.5u, Brier, and the Overall score, while the PAN system achieves a slightly higher F1 score.

Table 5 presents results for selected domain pairs. For each pair, the cross-domain-only model generally achieves competitive or higher scores than the best PAN 2022 participant across several metrics. In particular, the Overall score is higher for all three domain pairs shown, indicating improved overall verification performance despite variation in individual metrics.

Table 6 presents detailed results for all evaluated cross-domain pairs. Some pairs are omitted because either the corresponding domain was not available in our dataset (e.g., *business memo*) or the number of valid pairs was too small to compute reliable metrics. The bottom three rows report PAN 2022 reference results for comparison, allowing the reader to contextualize our model's performance relative to previously reported benchmarks (see Fig. 4).

Fig. 5 visualizes the model's Overall scores across all domain pairs in a symmetric heatmap, where each cell corresponds to a domain pair regardless of order. Higher scores indicate better AV performance. The heatmap highlights substantial variability across domain combinations. While some pairs, particularly those involving short-form or conversational domains such as *text message*, achieve relatively strong performance, other combinations yield more moderate results.

Notably, certain domain pairs exhibit exceptionally high scores, such as combinations involving *se query* and *se query memo* or *handwritten* text, indicating that stylistic signals can be highly discriminative in specific cross-domain settings. In contrast, domains such as *interview* and some *speech-to-text* pairings tend to show lower performance, reflecting increased difficulty in capturing consistent stylistic patterns.

Similarly, Fig. 6 presents AUROC scores for all domain pairs. The observed patterns largely align with the Overall scores, confirming that some domain combinations allow for more reliable discrimination between same- and different-author pairs. However, AUROC values also show considerable variation across domain pairs, indicating that discriminative performance depends strongly on the specific characteristics of the domains involved rather than belonging to a single category of text types.

Together, these heatmaps provide a comprehensive overview of cross-domain model performance, allowing visual comparison across all combinations of domains in the dataset.

Table 4
Overall AV performance comparing the current cross-domain-only model with the PAN 2022 participant achieving the highest overall score across all domain pairs.

Dataset	AUROC	c@1	F1	F0.5u	Brier	Overall
Overall (current model, cross only)	0.640	0.587	0.654	0.596	0.775	0.651
Overall (PAN 2022 best)	0.598	0.571	0.669	0.542	0.749	0.600

Table 5
AV performance for selected domain pairs, comparing the current cross-domain-only model with the PAN 2022 participant achieving the highest overall score for each pair.

Domain pair	System	AUROC	c@1	F1	F0.5u	Brier	Overall
email–text message	Ours (cross only)	0.584	0.565	0.650	0.585	0.763	0.629
	PAN 2022 best	0.613	0.583	0.672	0.581	0.628	0.614
essay–email	Ours (cross only)	0.688	0.604	0.618	0.531	0.770	0.642
	PAN 2022 best	0.531	0.481	0.659	0.531	0.750	0.590
essay–text message	Ours (cross only)	0.556	0.594	0.725	0.649	0.760	0.657
	PAN 2022 best	0.568	0.553	0.673	0.556	0.595	0.599

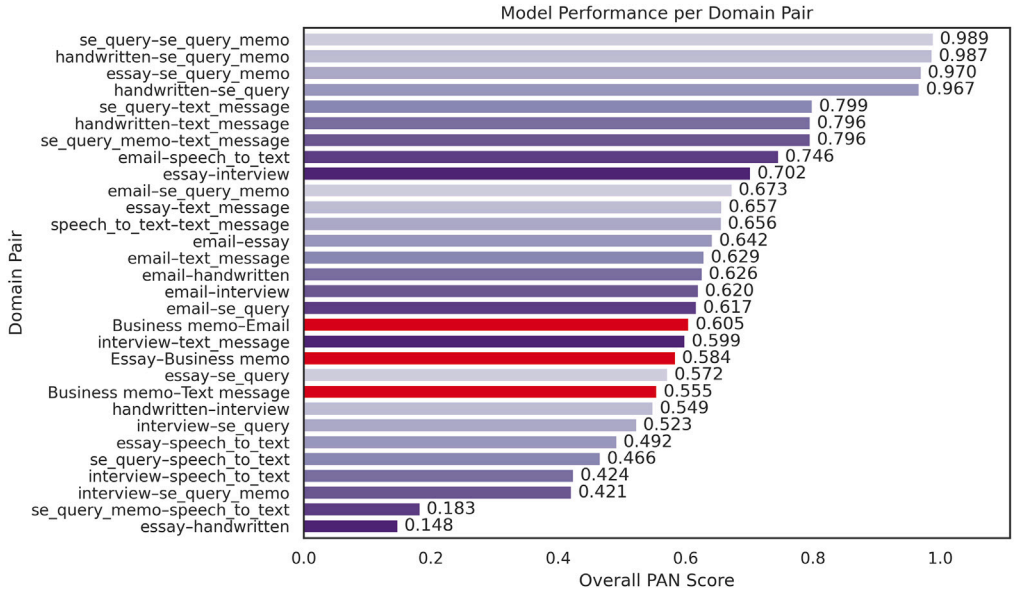

Fig. 4. Overall scores for all cross-domain pairs in our dataset. Red bars indicate PAN 2022 business memo pairs, which are included for reference but not part of our corpus. Only strictly cross-domain pairs are shown.

5.2. Holdout

Table 7 reports results for the author-disjoint cross-domain evaluation, where each held-out domain is not represented in the training set. Results are shown for two test settings: cross-domain pairs only and cross + same-domain pairs. Metrics include Accuracy, Precision, Recall, F1, Balanced Accuracy, AUROC, Brier, c@1, F0.5u, and Overall score.

Performance is reported separately for each domain. Domains such as *Interview*, *Email*, *Text message*, and *SE Query* include results for both evaluation settings, while some domains only contain cross-domain test pairs due to the absence of valid same-domain combinations.

The results demonstrate that the model is able to perform authorship verification on both cross-domain pairs and cross + same-domain pairs, even though same-domain pairs were not included during training. This indicates that the learned representations generalize to both evaluation settings without requiring explicit training on same-domain comparisons.

Table 8 reports the percentage of positive samples in each split. While the training and validation sets remain relatively balanced across domains, the test distributions vary more substantially, particularly for domains such as *SE query* and *SE query-memo*. These differences in class balance arise from the way the test sets are sampled for each domain, with some domains containing proportionally more positive pairs than others, which directly affects metrics such as Precision, Recall, and F1.

Overall, the results highlight the difficulty of AV under strict cross-domain conditions, where the model must generalize to stylistic patterns that were not observed during training.

5.3. Error analysis

The error analysis for the no-holdout setting reveals that the model predominantly produces false positives, with 2281 false positives compared to 533 false negatives. This indicates a bias toward predicting the positive class, suggesting that the model tends to classify text pairs as being written by the same author even when this is incorrect.

Training with hard negative sampling further contributes to this effect. While this strategy improves the model's ability to discriminate between similar authors, it also shifts the decision boundary toward more permissive similarity estimates, increasing the likelihood of false positive predictions in ambiguous cases.

The results show that domain pair errors are generally dominated by false positives, particularly in high-frequency pairs such as *email–text message* (922 FP) and *interview–text message* (492 FP), while false

Table 6
Cross-domain AV results for all evaluated domain pairs using only cross-domain pairs. The bottom three rows show PAN 2022 results for domain pairs not present in our dataset (involving business memo), using the best overall scores from the PAN 2022 task to allow comparison with our model's performance on other pairs.

Domain Pair	AUROC	c@1	F1	F0.5u	Brier	Overall
se query–text message	0.529	0.835	0.910	0.865	0.854	0.799
handwritten–text message	0.719	0.744	0.853	0.824	0.842	0.796
se query memo–text message	0.481	0.833	0.909	0.899	0.855	0.796
email–speech to text	0.810	0.773	0.667	0.641	0.842	0.746
essay–interview	0.726	0.654	0.690	0.625	0.815	0.702
email–se query memo	0.468	0.643	0.783	0.709	0.763	0.673
essay–text message	0.556	0.594	0.725	0.649	0.760	0.657
speech to text–text message	0.633	0.574	0.667	0.629	0.774	0.656
email–essay	0.688	0.604	0.618	0.531	0.770	0.642
email–text message	0.584	0.565	0.650	0.585	0.763	0.629
email–handwritten	0.516	0.556	0.714	0.610	0.733	0.626
email–interview	0.705	0.633	0.480	0.473	0.808	0.620
email–se query	0.528	0.542	0.673	0.595	0.745	0.617
interview–text message	0.557	0.536	0.599	0.542	0.761	0.599
PAN 2022 results for reference						
business memo–email	0.606	0.586	0.612	0.589	0.633	0.605
business memo–text message	0.525	0.376	0.673	0.455	0.748	0.555
essay–business memo	0.496	0.500	0.635	0.547	0.739	0.584

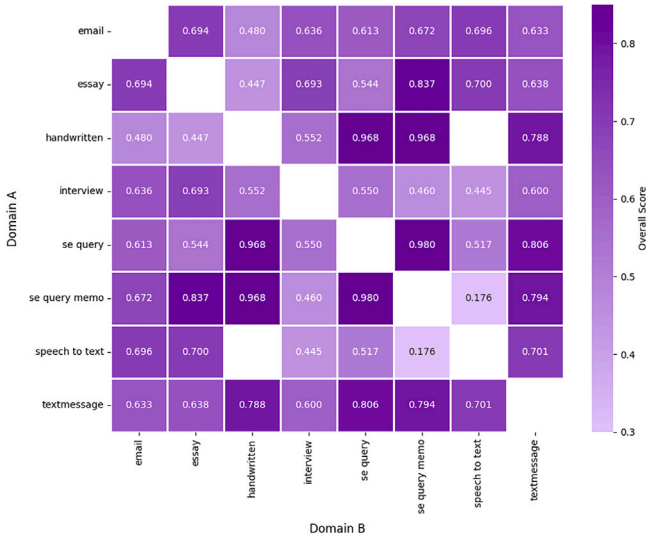

Fig. 5. Heatmap of Overall scores for all domain pairs. Each cell represents the Overall score of the cross-domain AV model for the corresponding pair of domains. The matrix is symmetric, as the order of domains in the pair does not affect the score.

negatives remain comparatively lower across most domain combinations. On author pair level, the false positives are concentrated in a subset of author pairs, while false negatives remain consistently low. In several cases, no false negatives are observed, which aligns with the design choice of prioritizing recall over precision. More extensive analysis on the domain pairs and author pairs error types can be found in the Appendix, Tables A.1 and A.2.

An analysis of errors across cosine similarity bins shows relatively stable error rates across all similarity ranges, indicating that embedding similarity alone is not strongly predictive of model correctness in this setting. This suggests that the model does not rely solely on geometric similarity in the embedding space when making predictions.

At the author level, error rates are relatively consistent across the most affected authors, with no single author dominating the failure cases. This indicates that errors are largely systematic rather than driven by specific outlier authors.

Fig. 6. Heatmap of AUROC for all domain pairs. Each cell represents the AUROC achieved by the model for the corresponding domain pair. The matrix is symmetric, reflecting that domain order does not affect the result.

Overall, the results suggest that the main source of error is an overestimation of similarity between texts from different authors, leading to an increased number of false positives. This behavior is consistent with training under hard negative sampling and highlights a trade-off between discriminative power and calibration in cross-domain authorship verification.

6. Discussion

6.1. Major findings

6.1.1. Non-holdout experiments

In the non-holdout setting, all discourse types are included in the training data. Evaluation is performed exclusively on cross-domain pairs, aligning directly with the PAN 2022 cross-discourse task and providing the fairest basis for comparison. Accordingly, our discussion focuses on results obtained from strictly cross-domain pairs, as reported in Tables 4, 5, and 6.

The model consistently outperforms the PAN 2022 best-performing system across multiple evaluation metrics. For overall pairs (Table 4), the cross-domain configuration achieves an Overall score of 0.654, compared to 0.600 for PAN 2022. These findings indicate that the model effectively captures stylistic signals that generalize across discourse types.

Examining individual domain pairs (Table 5), the cross-domain model shows particularly strong performance on the most frequent and linguistically rich pairs. For instance, *essay–email* achieves an Overall score of 0.694, *essay–text message* reaches 0.638, and *email–text message* achieves 0.633. These scores exceed the corresponding PAN 2022 baselines (Overall scores of 0.590, 0.599, and 0.614, respectively), indicating that the model effectively leverages stylistic patterns shared across domains. Improvements are especially pronounced for *essay–email*, where the Overall score increases by 0.104, highlighting the model's ability to generalize between structured long-form and shorter, less formal texts.

It is worth noting that the PAN 2022 dataset includes *business memo* pairings, which are absent in our corpus. While we cannot directly compare on this specific domain, our dataset contains a broader range of other cross-domain pairings, particularly short-form communications such as *text messages* and *SE queries*, which are highly relevant for evaluating cross-domain generalization. This diversity demonstrates

Table 7
AV results for held-out domains, two test settings: cross-domain pairs only ("cross") and cross + same-domain pairs ("cross + same"). Some domains had no same domain pairs available.

Domain	Setup	Acc	Prec	Rec	AUROC	c@1	F1	F0.5u	Brier	Overall
Interview	cross	0.543	0.430	0.600	0.550	0.543	0.501	0.456	0.691	0.548
	cross + same	0.552	0.456	0.617	0.567	0.552	0.524	0.481	0.696	0.564
Speech-to-text	cross	0.625	0.524	0.846	0.686	0.625	0.647	0.567	0.771	0.659
Email	cross	0.498	0.454	0.928	0.597	0.498	0.610	0.505	0.642	0.570
	cross + same	0.532	0.496	0.937	0.618	0.532	0.649	0.548	0.665	0.602
Essay	cross	0.611	0.507	0.720	0.667	0.611	0.595	0.539	0.757	0.634
Text message	cross	0.546	0.531	0.683	0.575	0.546	0.597	0.556	0.723	0.599
	cross + same	0.587	0.612	0.724	0.607	0.587	0.663	0.631	0.744	0.647
Handwritten	cross	0.444	0.632	0.240	0.531	0.444	0.348	0.476	0.648	0.489
SE query	cross	0.574	0.563	0.957	0.603	0.574	0.709	0.613	0.711	0.642
	cross + same	0.585	0.575	0.959	0.615	0.585	0.719	0.625	0.718	0.652
SE query memo	cross	0.488	0.643	0.346	0.584	0.488	0.450	0.549	0.713	0.557

Table 8
Proportion of positive (same-author) pairs in each domain split. Training and validation distributions are identical for both settings; test distributions are shown for cross-domain only and cross + same-domain pairs.

Domain	Train (%)	Val (%)	Test (cross) (%)	Test (cross + same) (%)
Interview	55.3	56.8	41.8	37.8
Speech-to-text	50.1	50.0	42.0	40.6
Email	55.4	54.1	47.0	46.1
Essay	50.0	50.0	50.6	39.7
Text message	41.9	43.0	53.5	56.1
Handwritten	49.8	50.0	64.9	61.7
SE query	47.2	46.8	64.8	55.4
SE query memo	49.9	49.7	74.4	60.5

that the model's performance gains are not limited to the exact domains present in PAN 2022, but extend to a wider variety of real-world text types.

Error analysis further shows that misclassifications are not uniformly distributed across domain and author pairs. Higher false positive rates are concentrated in several informal cross-domain combinations and a subset of stylistically similar author pairs, while false negatives remain consistently low across most settings. This pattern is consistent with a bias towards recall-oriented behavior, where the model is more inclined to predict positive pairs, thereby reducing missed true author matches at the cost of increased false positives.

Furthermore, the robustness of these findings is supported by our analysis of the training dynamics. In the non-holdout setting, we observed steady convergence across accuracy, recall, and loss metrics, with validation performance stabilizing without evidence of overfitting. This steady learning trajectory provides confidence that the performance gains observed in our cross-domain results are derived from stable stylistic feature extraction rather than artifacts of the training process.

Taken together, these results show that in a non-holdout setting, the model consistently surpasses PAN 2022 baselines on comparable cross-domain pairs. The improvements are most notable for domain pairs that combine longer, structured texts with shorter or informal communications, suggesting effective learning of cross-domain stylistic cues. Overall, the model demonstrates strong generalization to unseen domain pairings within the training set, underscoring its potential applicability to heterogeneous text corpora beyond the original PAN 2022 data.

6.1.2. Holdout experiments

In the holdout setting, the training data is drawn exclusively from non-holdout domains, while the test set contains pairs involving held-out domains. Results are reported for two evaluation settings: cross-domain pairs only and cross + same-domain pairs (when available). This design creates a stringent evaluation scenario, where the model must generalize stylistic signals from known domains to entirely unseen ones, representing a realistic cross-domain verification challenge.

Focusing on strictly cross-domain pairs (Table 7), the model demonstrates consistent performance across several domains. Structured, long-form texts such as *Essay* and *Speech-to-text* achieve Overall scores of 0.634 and 0.659, respectively. Short-form communication domains, including *Text message* and *SE query*, also yield competitive results with Overall scores of 0.599 and 0.642, indicating that stylistic signals can be captured across both formal and informal text types.

Monitoring the training progress for these holdout scenarios reveals that the model maintains consistent convergence patterns even when entire domains are excluded from the training set, suggesting that the model successfully adapts its representations to the stylistic shifts encountered in unseen data.

Some domains show different levels of difficulty under cross-domain evaluation. It should be noted that some domains exhibit imbalanced class distributions (Table 8). While the training data is balanced, test sets vary in their proportion of positive pairs, which can affect threshold-dependent metrics such as F_1-score and c@1. In contrast, AUROC is less sensitive to these differences. Therefore, absolute performance values across domains should be interpreted with this variation in class prior in mind. *Handwritten* obtains an Overall score of 0.489, while *SE query memo* reaches 0.557 in the cross-domain setting, reflecting variation in domain characteristics such as writing style and structure.

For several domains, results are also available when same-domain pairs are included in the test set. In these cases, such as *Interview*, *Email*, *Text message*, and *SE query*, the model is evaluated under both cross-domain and cross + same-domain conditions, allowing comparison of performance across both settings within the same domain. This also shows that the model is able to handle same-domain pairs for domains that were not observed during training, indicating that the learned representations also generalize to same-domain comparisons within unseen domains.

Overall, these results demonstrate strong verification performance in a challenging evaluation scenario, where entire domains are withheld from training. This simulates the practical scenario of encountering completely new text types in real-world applications, such as new communication platforms, emerging text genres, or previously unseen authors. These results suggest that the model can generalize across heterogeneous text corpora, highlighting its applicability to real-world AV tasks.

6.2. Limitations

Several limitations should be noted when interpreting these results.

First, some domain pairs contain relatively few samples. This leads to unstable evaluation metrics and occasionally undefined values such

as NaN AUROC scores. Small sample sizes may also artificially inflate certain metrics, particularly F1 and $F_{0.5u}$, when predictions happen to align with the limited ground truth labels.

Second, the domain distribution in the dataset is highly imbalanced. For example, *text messages* and *emails* contain significantly more samples than domains such as *speech-to-text*, *handwritten*, or *SE query-memo*. As a result, the model may learn domain-specific patterns for the more frequent domains, which could bias evaluation results.

Related to this imbalance, certain domain pairs (e.g., *se query–text message* and *se query memo–text message*) show a strong discrepancy between threshold-based metrics and ranking-based metrics. In these cases, c@1 and F_1 are relatively high, while AUROC is considerably lower. This can be explained by the fact that AUROC evaluates overall separability across all thresholds, whereas c@1 and F_1 reflect performance at a single operating point. As a result, the model can perform well under the chosen decision threshold even when global score separation between classes is weaker. This highlights that different evaluation metrics capture complementary aspects of performance, especially under imbalanced cross-domain conditions.

Third, the error analysis indicates that the model exhibits a bias toward predicting the positive class, resulting in a higher number of false positives compared to false negatives. This behavior is consistent with training using hard negative sampling, which encourages the model to focus on subtle inter-author differences but may also lead to more permissive similarity estimates. As a result, the model may overestimate similarity in ambiguous cross-domain cases, which should be considered when interpreting the reported performance.

The PAN category *business memo* was not present in our data, so we compared our model against all other available domain pairs instead. While this provides a reasonable proxy — and most of our domain pairs outperform the business memo pairs — the stylistic characteristics of these domains may not perfectly align with the original PAN categories.

Finally, the holdout experiment presents additional challenges specific to cross-domain generalization. In this setup, the training set contains only pairs from non-holdout domains, while the test set includes pairs where one text comes from a holdout domain. Some holdout domains are relatively scarce or highly heterogeneous, which can result in greater metric variability and increased difficulty in capturing reliable stylistic patterns. These factors represent inherent limitations of evaluating AV in a realistic, unseen-domain setting.

6.3. Future work

Several directions for future work could further improve cross-domain AV.

One promising direction is the development of domain-invariant representations. Techniques such as adversarial domain adaptation or contrastive learning could encourage the model to focus on author-specific stylistic signals rather than domain-specific features, potentially improving generalization to unseen domains.

Another avenue for improvement is the incorporation of explicit stylistic features. While modern transformer-based embeddings capture many linguistic patterns, combining them with interpretable stylometric features may improve robustness across domains.

A particularly interesting extension involves systematic holdout experiments to quantify the impact of individual domains in the training set. For example, one could train the model on only a subset of the available domains and evaluate it on a specific holdout domain. This would provide insight into how much each training domain contributes to cross-domain performance and whether the model can effectively generalize to new domains based on limited domain diversity. Such experiments could help optimize training data selection and reveal which domain combinations are most informative for AV.

Future research could also investigate more balanced evaluation protocols, ensuring that each domain pair contains sufficient samples to produce stable performance estimates. Expanding the dataset with additional texts from underrepresented domains may significantly improve the reliability of cross-domain evaluation.

This increased data availability per domain would also lead to the possibility of a more extreme LODO setting, where the model is trained on only a small subset of source domains evaluated on unseen target domains. This setting would isolate the minimum domain knowledge required for cross-domain generalization and provide a stricter test of robustness under severe data scarcity. Such an experiment would complement the current LODO setup by evaluating performance under highly constrained training conditions.

Another direction is refining the test set to better reflect real-world conditions. Instead of balancing positive and negative pairs, the test set could contain a majority of negative pairs, mimicking the natural scarcity of same-author examples. Adjusting pair generation in this way may provide a more realistic evaluation of cross-domain AV performance.

Finally, further work could explore profile-based AV approaches, where models compare new texts against aggregated author profiles rather than individual text pairs. Such methods may better capture consistent stylistic characteristics across multiple domains and writing contexts.

7. Conclusion

This study presents a comprehensive approach to cross-domain AV, evaluated across multiple heterogeneous text domains. In non-holdout experiments, where all discourse types are represented in the training data, the proposed model consistently outperforms the best PAN 2022 participant on both overall and per-domain metrics, including AUROC, c@1, F1, F0.5u, Brier, and Overall score. Strongest performance is observed for domain pairs combining longer, structured texts with shorter or informal communications, such as *essay–email*, while shorter or less structured domains, including *handwritten* or *SE query-memo*, remain more challenging, reflecting inherent differences in text style and structure.

Holdout experiments further demonstrate the model's robustness in more realistic scenarios, where training data contains only non-holdout domains while the test set contains pairs in which one text originates from a held-out domain.

Even under these stringent conditions, the model achieves strong performance on strictly cross-domain pairs, confirming its ability to capture domain-invariant stylistic signals that generalize to unseen writing contexts. Interestingly, Including same-domain pairs in the test set provides an additional perspective, showing that the model is also able to handle same-domain comparisons for domains not observed during training. Differences between evaluation settings can be attributed in part to variations in class balance and pair distributions, highlighting the influence of sampling on reported metrics. These results emphasize the importance of holdout evaluations for assessing cross-domain generalization beyond benchmarks such as PAN 2022.

Detailed per-domain analyses reveal variability in performance, emphasizing that domain characteristics — such as text length, structure, and style — significantly influence verification outcomes. While some PAN categories, such as business memos, were not present in our dataset, the model demonstrates consistent performance across a broad range of domain pairings. This indicates that the model effectively captures stylistic similarities across diverse text types, beyond the specific domains represented in PAN 2022.

Overall, the findings indicate that the proposed approach provides a robust and generalizable solution for cross-domain AV, effectively modeling stylistic similarities even for previously unseen domains. Future work could include expanding to larger and more diverse datasets, integrating additional stylometric or semantic features, and further exploring domain-specific contributions to generalization, particularly in low-resource or highly heterogeneous settings.

Table A.1
False positive and false negative analysis for domain pairs included in the main performance evaluation. The table reports the false positive (FP) and false negative (FN) counts and rates for the domain pairs included in the main performance evaluation. For each pair, total comparisons, FP, and FN are shown, along with their normalized rates. The values correspond to the same evaluation setting used in the performance tables, ensuring consistency between metric evaluation and error analysis.

Domain Pair	Total	FP	FN	FP rate	FN rate	Error rate
email–essay	53	13	9	0.245	0.170	0.415
email–handwritten	36	5	11	0.139	0.306	0.444
email–interview	1088	132	254	0.121	0.233	0.355
email–se query	428	192	3	0.449	0.007	0.456
email–se query memo	28	7	5	0.250	0.179	0.429
email–speech to text	22	2	5	0.091	0.227	0.318
email–text message	2797	922	309	0.330	0.110	0.440
essay–interview	26	5	2	0.192	0.077	0.269
essay–text message	69	26	7	0.377	0.101	0.478
handwritten–text message	39	7	5	0.179	0.128	0.308
interview–text message	1402	492	137	0.351	0.098	0.449
se query–text message	455	72	8	0.158	0.018	0.176
se query memo–text message	30	3	6	0.100	0.200	0.300
speech to text–text message	47	13	3	0.277	0.064	0.340

Table A.2
False positive and false negative analysis for selected author pairs with highest error rates. The table reports the false positive (FP) and false negative (FN) counts and rates for the selected author pairs. For each pair, total comparisons, FP, and FN are shown, along with their normalized rates and overall error rate. The values correspond to the same evaluation setting used in the performance tables, ensuring consistency between metric evaluation and error analysis.

Author Pair	Total	FP	FN	FP rate	FN rate	Error rate
(40, 61)	21	18	0	0.857	0.000	0.857
(71, 96)	18	15	0	0.833	0.000	0.833
(44, 98)	41	34	0	0.829	0.000	0.829
(15, 88)	16	13	0	0.813	0.000	0.813
(10, 16)	24	19	0	0.792	0.000	0.792
(15, 40)	43	34	0	0.791	0.000	0.791
(26, 110)	18	14	0	0.778	0.000	0.778
(91, 98)	25	18	0	0.720	0.000	0.720
(18, 81)	42	30	0	0.714	0.000	0.714
(64, 99)	38	27	0	0.711	0.000	0.711

CRediT authorship contribution statement

Britt van Leeuwen: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review and editing. **Sandjai Bhulai:** Conceptualization, Supervision, Writing – review and editing. **Rob van der Mei:** Conceptualization, Supervision, Writing – review and editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Britt van Leeuwen reports financial support was provided by Research Institute for Mathematics and Computer Science in the Netherlands. Britt van Leeuwen reports a relationship with Research Institute for Mathematics and Computer Science in the Netherlands that includes: employment. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. False positive and false negative analysis

See Tables A.1 and A.2.

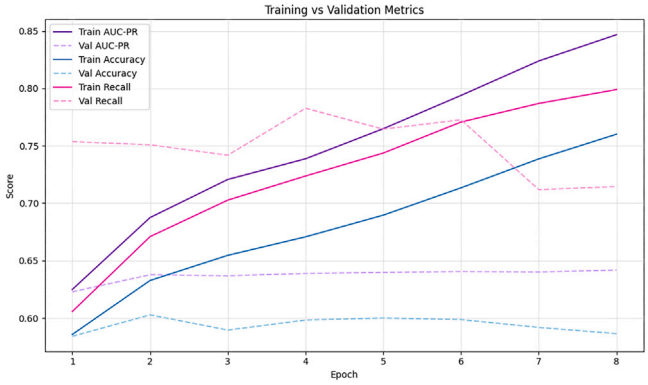

Fig. B.1. Training and validation curves for the non-holdout setting.

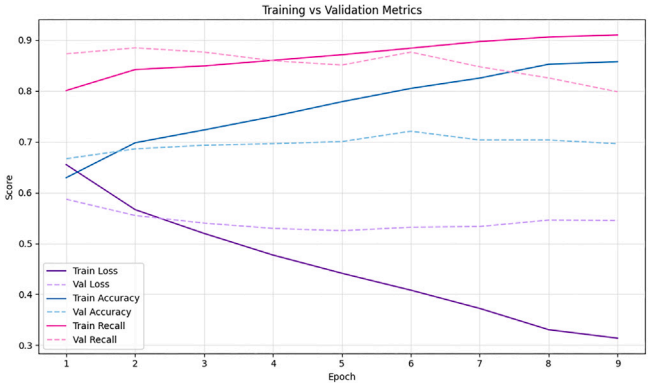

Fig. B.2. Holdout setting: training and validation curves when *email* is used as the held-out domain.

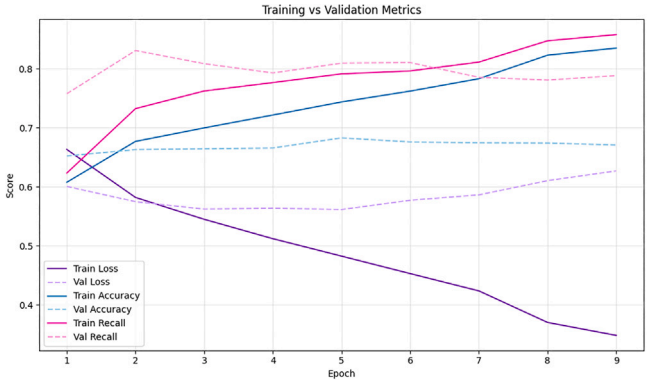

Fig. B.3. Holdout setting: training and validation curves when *essay* is used as the held-out domain.

Appendix B. Training validation graphs

This section presents figures illustrating the training and validation progress for all experimental configurations. Each figure displays the convergence of loss, accuracy, and recall metrics.

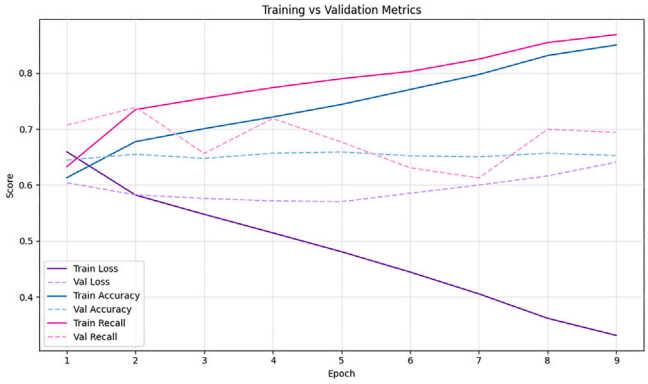

Fig. B.4. Holdout setting: training and validation curves when *handwritten* text is used as the held-out domain.

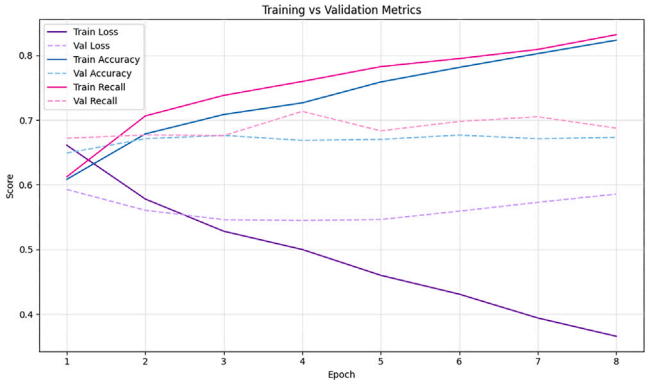

Fig. B.7. Holdout setting: training and validation curves when *SE query* data is used as the held-out domain.

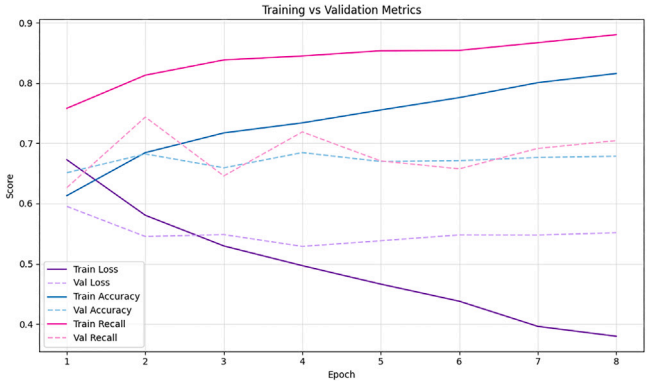

Fig. B.5. Holdout setting: training and validation curves when *interview* transcripts are used as the held-out domain.

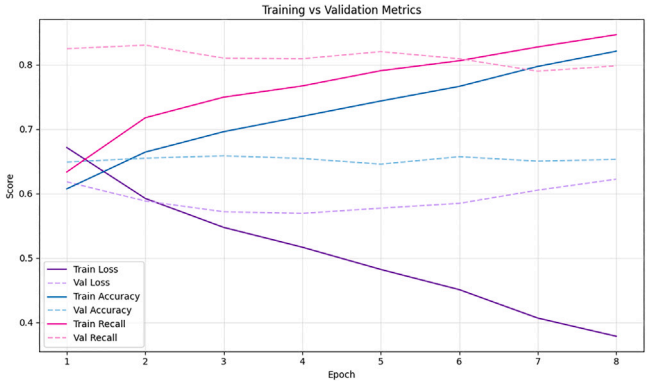

Fig. B.8. Holdout setting: training and validation curves when *speech-to-text* data is used as the held-out domain.

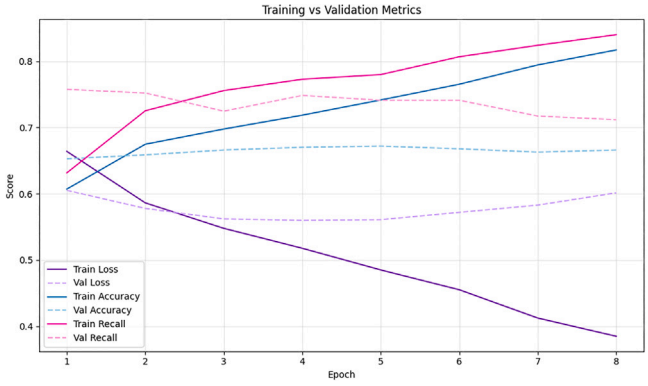

Fig. B.6. Holdout setting: training and validation curves when *SE query memo* data is used as the held-out domain.

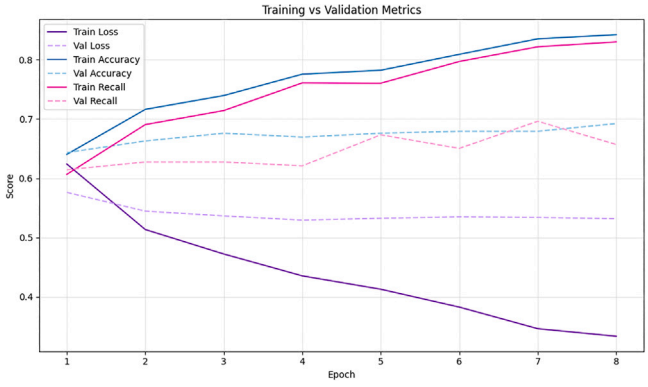

Fig. B.9. Holdout setting: training and validation curves when *text messages* are used as the held-out domain.

Data availability

The authors do not have permission to share data.

References

Ammar Aouchiche, R. I., Boumahdi, F., Remmide, M. A., Hemina, K., & Guendouz, A. (2024). Enhanced authorship verification for textual similarity with siamese deep learning. *Journal of Mobile Multimedia*, 821–844. http://dx.doi.org/10.13052/jmm1550-4646.2043.

Barlas, G., & Stamatatos, E. (2020). Cross-domain authorship attribution using pre-trained language models. In *IFIP international conference on artificial intelligence applications and innovations* (pp. 255–266). Springer.

Bevendorff, J., Borrego-Obrador, I., Chinea-Ríos, M., Franco-Salvador, M., Fröbe, M., Heini, A., Kredens, K., Mayerl, M., Pęzik, P., Potthast, M., et al. (2023). Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview. In *International conference of the cross-language evaluation forum for European languages* (pp. 459–481). Springer.

Bevendorff, J., Chulvi, B., De La Peña Sarracén, G. L., Kestemont, M., Manjavacas, E., Markov, I., Mayerl, M., Potthast, M., Rangel, F., Rosso, P., et al. (2021). Overview of PAN 2021: authorship verification, profiling hate speech spreaders on twitter, and style change detection. In *International conference of the cross-language evaluation forum for European languages* (pp. 419–431). Springer.

Cammarota, V., Bozza, S., Roten, C. A., & Taroni, F. (2024). Stylometry and forensic science: A literature review. *Forensic Science International: Synergy*, *9*, Article 100481. http://dx.doi.org/10.1016/j.fsisyn.2024.100481, URL https://www.sciencedirect.com/science/article/pii/S2589871X24000287.

Futrzynski, R. (2021). Author classification as pre-training for pairwise authorship verification. In *CLEF (working notes)* (pp. 1945–1952).

Guo, M., Han, Z., Chen, H., & Qi, H. (2023). A contrastive learning of sample pairs for authorship verification. In *CLEF (working notes)* (pp. 2608–2612).

He, X., Lashkari, A. H., Vombatkere, N., & Sharma, D. P. (2024). Authorship attribution methods, challenges, and future research directions: A comprehensive survey. *Information*, *15*(3), 131.

Hermans, A., Beyer, L., & Leibe, B. (2017). In defense of the triplet loss for person re-identification. URL https://arxiv.org/abs/1703.07737, arXiv:1703.07737.

Hu, X., Ou, W., Acharya, S., Ding, S. H., D'Gama, R., & Yu, H. (2023). TDRLM: Stylometric learning for authorship verification by topic-debiasing. *Expert Systems with Applications*, *233*, Article 120745. http://dx.doi.org/10.1016/j.eswa.2023.120745, URL https://www.sciencedirect.com/science/article/pii/S0957417423012472.

Kestemont, M., Stamatatos, E., Modaresi, P., Manjavacas, E., Potthast, M., Ghanem, B., Specht, G., & Stein, B. (2022). Overview of the authorship verification task at PAN 2022. In *CLEF 2022 working notes*. CEUR-WS.org, URL https://ceur-ws.org/Vol-3180/paper-122.pdf.

Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2018). Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection. In *Working notes papers of the CLEF 2018 evaluation labs* (pp. 1–25).

Khan, T. F., Anwar, W., Arshad, H., & Abbas, S. N. (2023). An empirical study on authorship verification for low resource language using hyper-tuned CNN approach. *IEEE Access*, *11*, 80403–80415. http://dx.doi.org/10.1109/ACCESS.2023.3299565.

Kingma, D. P., & Ba, J. (2017). Adam: A method for stochastic optimization. URL https://arxiv.org/abs/1412.6980, arXiv:1412.6980.

Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., & Demidov, P. (2019). A survey on stylometric text features. In *25th conference of open innovations association* (pp. 184–195). IEEE.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. http://dx.doi.org/10.48550/arXiv.1907.11692.

Luyckx, K., & Daelemans, W. (2008). Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd international conference on computational linguistics* (pp. 513–520).

Misini, A., Kadriu, A., & Canhasi, E. (2022). A survey on authorship analysis tasks and techniques. *Seeu Review*, *17*(2), 153–167.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys (CSuR)*, *50*(6), 1–36.

PAN Lab (2022). PAN 2022 authorship verification dataset. https://pan.webis.de/data.html. (Accessed 06 March 2026).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems: vol. 32*, (pp. 8024–8035). Curran Associates, Inc., URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Peñas, R., & Rodrígo, P. (2011). Evaluating systems for authorship verification. In *CLEF (online working notes/labs/workshop)*. Introduced the c@1 score for authorship verification evaluations.

Petropoulos, P. (2023). Contrastive learning for authorship verification using BERT and bi-LSTM in a siamese architecture.. In *CLEF (working notes)* (pp. 2734–2741).

Schler, J., Koppel, M., Argamon, S., & Pennebaker, J. (2006). Effects of age and gender on blogging. (pp. 199–205).

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, *60*(3), 538–556.

Stamatatos, E., Kestemont, M., Kredens, K., Pezik, P., Heini, A., Bevendorff, J., Stein, B., & Potthast, M. (2022). Overview of the authorship verification task at PAN 2022. In *CEUR workshop proceedings: vol. 3180*, (pp. 2301–2313).

Tyo, J., Dhingra, B., & Lipton, Z. C. (2021). Siamese bert for authorship verification. In *CLEF (working notes)* (pp. 2169–2177).

Tyo, J., Dhingra, B., & Lipton, Z. C. (2022). On the state of the art in authorship attribution and authorship verification. URL https://arxiv.org/abs/2209.06869, arXiv:2209.06869.

van Leeuwen, B., Bhulai, S., & van der Mei, R. D. (2025). Combining style and semantics for robust authorship verification. *Machine Learning with Applications*, *22*, Article 100732. http://dx.doi.org/10.1016/j.mlwa.2025.100732, URL https://www.sciencedirect.com/science/article/pii/S266682702500115X.

Webis Group (2024). PAN @ CLEF shared tasks on authorship analysis, computational ethics, and originality. https://pan.webis.de/shared-tasks.html. (Accessed 09 March 2026).