O2D2: Out-Of-Distribution Detector to Capture Undecidable Trials in Authorship Verification

Notebook for PAN at CLEF 2021

Benedikt Boenninghoff[1], Robert M. Nickel[2] and Dorothea Kolossa[1]

[1]Ruhr University Bochum, Germay
[2]Bucknell University, USA

Abstract

The PAN 2021 authorship verification (AV) challenge is part of a three-year strategy, moving from a cross-topic/closed-set AV task to a cross-topic/open-set AV task over a collection of fanfiction texts. In this work, we present a novel hybrid neural-probabilistic framework that is designed to tackle the challenges of the 2021 task. Our system is based on our 2020 winning submission, with updates to significantly reduce sensitivities to topical variations and to further improve the system's calibration by means of an uncertainty adaptation layer. Our framework additionally includes an *out-of-distribution* detector (O2D2) for defining non-responses. Our proposed system outperformed all other systems that participated in the PAN 2021 AV task.

Keywords

Authorship Verification, Out-Of-Distribution Detection, Open-Set,

1. Introduction

In this paper we are proposing a significant extension to the authorship verification (AV) system presented in [1]. The work is part of the PAN 2021 AV shared task[2], for which the PAN organizers provided the challenge participants with a publicly available dataset of fanfiction.

Fanfiction texts are fan-written extensions of well-known story lines, in which the so-called fandom topic describes the principal subject of the literary document (e.g. *Harry Potter*). The use of fanfiction as a *genre* has three major advantages. Firstly, the abundance of texts written in this genre makes it feasible to collect a large training dataset and, therefore, to build more complex authorship verification (AV) systems based on modern deep learning techniques, which will hopefully boost progress in this research area. Additionally, fanfictional documents also come with meaningful meta-data like topical information, which can be used to investigate the topical interference in authorship analysis. Lastly, although the documents are usually produced by non-professional writers, contrary to social media messages, they usually follow standard grammatical and spelling conventions. This allows participants to incorporate pretrained models for, e.g., part-of-speech tagging, and to reliably extract traditional stylometric features [3].

The previous edition of the PAN AV task dealt with cross-fandom/closed-set AV [4]. The objective of the *cross-fandom* AV task is to automatically decide whether two fanfictional documents covering different fandoms belong to the same author. The term *closed-set* refers to

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

benedikt.boenninghoff@rub.de (B. Boenninghoff); rmn009@bucknell.edu (R. M. Nickel); dorothea.kolossa@rub.de (D. Kolossa)

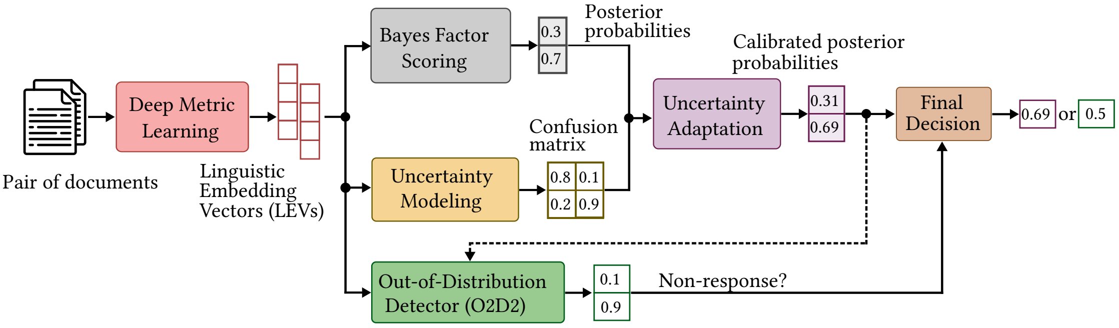


Figure 1: Our proposed hybrid neural-probabilistic framework for the PAN 2021 cross-fandom open-set authorship verification task.

the fact that the test dataset, which is not publicly available, only contains trials from a subset of the authors and fandoms provided in the training data.

To increase the level of difficulty, the current PAN AV challenge moved from a *closed-set* task to an *open-set* task in 2021, while the training dataset is identical to that of the previous year [5]. In this scenario, the new test data contains *only* authors and fandoms that were *not* included in the training data. We thus expect a *covariate shift* between training and testing data, i.e. the distribution of our neural stylometric representations extracted from the training data is expected to be different from the distribution of the test data representations. It was implicitly shown in [4], and our experiments confirm this analysis, that such a covariate shift, due to topic variability, is a major cause of errors.

2. System Overview

The overall structure of our revised system[1] is shown in Fig. 1. It expands our winning system from 2020 as follows: Suppose we have a pair of documents $\mathcal{D}_1$ and $\mathcal{D}_2$ with an associated ground-truth hypothesis $\mathcal{H}_a$ for $a \in \{0, 1\}$. The value of a indicates, whether the two documents were written by the same author ($a = 1$) or by different authors ($a = 0$). Our task can formally be expressed as a mapping $f{:}\{\mathcal{D}_1, \mathcal{D}_2\} \longrightarrow p \in [0, 1]$. The estimated label $\widehat{a}$ is obtained from a threshold test applied to the output prediction p. In our case, we choose $\widehat{a} = 1$ if $p > 0.5$ and $\widehat{a} = 0$ if $p < 0.5$. The PAN 2020/21 shared tasks also permit the return of a *non-response* (in addition to $\widehat{a} = 1$ and $\widehat{a} = 0$) in cases of high uncertainty [4], e.g. when p is close to 0.5. In this work, we therefore define three hypotheses:

$$\begin{aligned}
\mathcal{H}_0 &: \text{ The two documents were written by two different persons,}\\
\mathcal{H}_1 &: \text{ The two documents were written by the same person,}\\
\mathcal{H}_2 &: \text{ Undecidable, trial does not suffice to establish authorship.}
\end{aligned}$$

In [1], we introduced the concept of *linguistic embedding vectors* (LEVs). To obtain these, we perform neural feature extraction followed by *deep metric learning* (DML) to encode the stylistic characteristics of a pair of documents into a pair of fixed-length and topic-invariant stylometric representations. Given the LEVs, a *Bayes factor scoring* (BFS) layer computes the

[1] The source code is accessible online: https://github.com/boenninghoff/pan_2020_2021_authorship_verification

posterior probability for a trial. This discriminative two-covariance model was introduced in [6]. As a new component, we propose an *uncertainty adaptation layer* (UAL). This idea is adopted from [7], aiming to find and correct wrongly classified trials of the BFS layer, to model its noise behavior, and to return re-calibrated posteriors.

For the decision whether to accept $\mathcal{H}_0/\mathcal{H}_1$, or to return a non-response, i.e. $\mathcal{H}_2$, it is desirable that the value of the posterior p reliably reflects the uncertainty of the decision-making process. We may roughly distinguish two different types of uncertainty [8]: In AV, *aleatoric* or data uncertainty is associated with properties of the document pairs. Examples are topical variations or the intra- and inter-author variabilities. Aleatoric uncertainty generally can not be reduced, but it can be addressed (to a certain extent) by returning a non-response (i.e. hypothesis $\mathcal{H}_2$) if it is too large to allow for a reliable decision. To accomplish this, and inspired by [9], we incorporate a feed-forward network for *out-of-distribution detection* (O2D2), which is trained on a dataset that is different, i.e. disjoint w.r.t. authors and fandoms, from the training set used to optimize the DML, BFS and UAL components.

Additionally, *epistemic* or model uncertainty characterizes uncertainty in the model parameters. Examples are *unseen* authors or topics. Epistemic uncertainty can be reduced through a substantial increase in the amount of training data, i.e. an increase in the number of training pairs. We capture epistemic uncertainty in our work through the proposed O2D2 approach and also by extending our model to an ensemble. We expect all models to behave similarly for known authors or topics, but the output predictions may be widely dispersed for pairs under covariate shift [10].

The training procedure consists of two stages: In the first stage, we simultaneously train the DML, BFS and UAL components. In the second stage, we learn the parameters of the O2D2 model.

3. Dataset Splits for the PAN 2021 AV Task

The text preprocessing strategies, including tokenization and pair re-sampling, are comprehensively described in [11]. The *fanfictional* dataset for the PAN 2020/21 AV tasks are described in [4, 5]. In the following, we report on the various dataset splits that we employed for our PAN 2021 submission.

Each document pair is characterized by a tuple (a, f), where $a \in \{0, 1\}$ denotes the *authorship similarity label* and $f \in \{0, 1\}$ describes the equivalent for the fandom. We assign each document pair to one of the following author-fandom subsets[2] SA_SF, SA_DF, DA_SF, and DA_DF given its label tuple (a, f).

As shown in [11], one of the difficulties working with the provided small/large PAN datasets is that each author generally contributes only with a small number of documents. As a result, we observe a high degree of overlap in the re-sampled subsets of same-author trials. We decided to work only with the large dataset this year and split the documents into three disjoint (w.r.t. authorship and fandom) sets. Overlapping documents, where author and fandom belong to different sets, are removed. The splits are summarized in Fig. 2 and Table 1. Altogether, the following datasets have been involved in the PAN 2021 shared task, to train the model components, tune the hyper-parameter and for testing:

[2] SA=same author, DA=different authors, SF=same fandom, DF=different fandoms

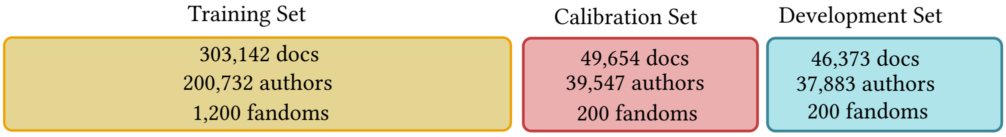


Figure 2: Disjoint splits of the large PAN 2020/21 training set.

Table 1: Numbers of (re-)sampled pairs for all datasets.

Dataset	SA_SF	SA_DF	DA_SF	DA_DF
Training set	16,045	28,500	64,300	42,730
Calibration set	2,100	2,715	4,075	4,075
Validation set	0	2,280	3075	0
Development set	0	5,215	7,040	0

- The **training set** is identical to the one used in [11] and was employed for the first stage, i.e., to train the DML, BFS and UAL components simultaneously. During training we re-sampled the pairs epoch-wise such that all documents contribute equally to the neural network training in each epoch. The numbers of training pairs provided in Table 1 therefore vary in each epoch.
- The **calibration set** has been used for the second stage, i.e., to train (calibrate) the O2D2 model. During training, we again re-sampled the pairs in each epoch and limited the total number of pairs in the different-authors subsets to partly balance the dataset.
- The purpose of the **validation set** is to tune the hyper-parameters of the O2D2 stage and to report the final evaluation metrics for all stages in Section 5.
- The **development set** is identical to the evaluation set in[11] and was used to tune the hyper-parameters during the training of the first stage. This dataset contains pairs from the calibration and validation sets. However, due to the pair re-sampling strategy in [11], documents may appear in different subsets and varied document pairs may be sampled. It thus does not represent a union of the calibration and validation sets.
- Finally, the **PAN 2021 evaluation set**, which is not publicly available, has been used to test our submission and to compare it with the proposed frameworks of all other participants.

Note that both, the validation and development set in Table 1 only contain SA_DF and DA_SF pairs, for reasons discussed in Section 5. The pairs of these sets are sampled once and then kept fixed.

4. Methodologies

In this section, we briefly describe all components of our neural-probabilistic model. Sections 4.1 through 4.4 repeat information that is already provided in [11] to provide proper context.

4.1. Neural Feature Extraction and Deep Metric Learning

Feature extraction and deep metric learning are realized in the form of a *Siamese* network, feeding both input documents through exactly the same function.

4.1.1. Neural Feature Extraction:

The system passes token and character embeddings into a two-tiered bidirectional LSTM network with attentions,

$$\boldsymbol{x}_i = \text{NeuralFeatureExtraction}_{\boldsymbol{\theta}}\big(\boldsymbol{E}_i^w, \boldsymbol{E}_i^c\big), \tag{1}$$

where $\boldsymbol{\theta}$ contains all trainable parameters, $\boldsymbol{E}_i^w$ represents word embeddings and $\boldsymbol{E}_i^c$ represents character embeddings. A comprehensive description is given in [12].

4.1.2. Deep Metric Learning:

We feed the document embeddings $\boldsymbol{x}_i$ in Eq. (1) into a metric learning layer, $\boldsymbol{y}_i = \tanh\big(\boldsymbol{W}^{\text{DML}}\boldsymbol{x}_i + \boldsymbol{b}^{\text{DML}}\big)$, which yields the two LEVs $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ via the trainable parameters $\boldsymbol{\psi} = \{\boldsymbol{W}^{\text{DML}}, \boldsymbol{b}^{\text{DML}}\}$. We then compute the Euclidean distance between both LEVs, $d(\boldsymbol{y}_1, \boldsymbol{y}_2) = \|\boldsymbol{y}_1 - \boldsymbol{y}_2\|_2^2$. In [11], we introduced a new *probabilistic* version of the contrastive loss: Given the Euclidean distance of the LEVs, we apply a kernel function

$$p_{\text{DML}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2) = \exp\big(-\gamma\, d(\boldsymbol{y}_1, \boldsymbol{y}_2)^{\alpha}\big), \tag{2}$$

where γ and α can be seen as both, hyper-parameters or trainable variables. The loss then is given by

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\psi}}^{\text{DML}} = a \cdot \max\big\{\tau_s - p_{\text{DML}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2), 0\big\}^2 + (1-a) \cdot \max\{p_{\text{DML}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2) - \tau_d, 0\}^2\,, \tag{3}$$

where we set $\tau_s = 0.91$ and $\tau_d = 0.09$.

4.2. Deep Bayes Factor Scoring

We assume that the LEVs stem from a Gaussian generative model that can be decomposed as $\boldsymbol{y} = \boldsymbol{s} + \boldsymbol{n}$, where $\boldsymbol{n}$ characterizes a noise term. We assume that the writing characteristics of the author lie in a latent stylistic variable $\boldsymbol{s}$. The probability density functions for $\boldsymbol{s}$ and $\boldsymbol{n}$ are modeled as Gaussian distributions. We outlined in [1] how to compute the likelihoods for both hypotheses. The verification score for a trial is then given by the log-likelihood ratio: $\text{score}(\boldsymbol{y}_1, \boldsymbol{y}_2) = \log p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_1) - \log p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_0)$. Assuming $p(\mathcal{H}_1) = p(\mathcal{H}_0) = \frac{1}{2}$, the probability for a same-author trial is calculated as [1]:

$$p_{\text{BFS}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_1)}{p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_1) + p(\boldsymbol{y}_1, \boldsymbol{y}_2|\mathcal{H}_0)} = \text{Sigmoid}\big(\text{score}(\boldsymbol{y}_1, \boldsymbol{y}_2)\big) \tag{4}$$

We reduce the dimension of the LEVs via $\boldsymbol{y}_i^{\text{BFS}} = \tanh\big(\boldsymbol{W}^{\text{BFS}}\boldsymbol{y}_i + \boldsymbol{b}^{\text{BFS}}\big)$ to ensure numerically stable inversions of the matrices [1]. We rewrite Eq. (4) as

$$p_{\text{BFS}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2) = \text{Sigmoid}\big(\text{score}(\boldsymbol{y}_1^{\text{BFS}}, \boldsymbol{y}_2^{\text{BFS}})\big) \tag{5}$$

and incorporate Eq. (5) into the binary cross entropy,

$$\mathcal{L}_{\boldsymbol{\phi}}^{\text{BFS}} = a \cdot \log\{p_{\text{BFS}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2)\} + (1-a) \cdot \log\{1 - p_{\text{BFS}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2)\}\,, \tag{6}$$

where all trainable parameters are denoted with $\boldsymbol{\phi} = \big\{\boldsymbol{W}^{\text{BFS}}, \boldsymbol{b}^{\text{BFS}}, \boldsymbol{W}, \boldsymbol{B}, \boldsymbol{\mu}\big\}$.

4.3. Uncertainty Modeling and Adaptation

Now, we treat the posteriors of the BFS component as noisy outcomes and rewrite Eq. (5) as $p_{\text{BFS}}(\widehat{\mathcal{H}}_1|\boldsymbol{y}_1, \boldsymbol{y}_2)$ to emphasize that this represents an estimated posterior. We firstly have to find a single representation for both LEVs, which is done by $\boldsymbol{y}^{\text{UAL}} = \tanh\left(\boldsymbol{W}^{\text{UAL}}\left(\boldsymbol{y}_1 - \boldsymbol{y}_2\right)^{\circ 2} + \boldsymbol{b}^{\text{UAL}}\right)$, where $(\cdot)^{\circ 2}$ denotes the element-wise square. Next, we compute a 2×2 confusion matrix as follows

$$p(\mathcal{H}_j|\widehat{\mathcal{H}}_i, \boldsymbol{y}_1, \boldsymbol{y}_2) = \frac{\exp\left(\boldsymbol{w}_{ji}^T\, \boldsymbol{y}^{\text{BFS}} + b_{ji}\right)}{\sum\limits_{i' \in \{0,1\}} \exp\left(\boldsymbol{w}_{ji'}^T\, \boldsymbol{y}^{\text{BFS}} + b_{ji'}\right)} \quad \text{for } i, j \in \{0, 1\}. \tag{7}$$

The term $p(\mathcal{H}_j|\widehat{\mathcal{H}}_i, \boldsymbol{y}_1, \boldsymbol{y}_2)$ defines the conditional probability of the true hypothesis $\mathcal{H}_j$ given the hypothesis $\widehat{\mathcal{H}}_i$ assigned by the BFS. We can then define the final output predictions as:

$$p_{\text{UAL}}(\mathcal{H}_j|\boldsymbol{y}_1, \boldsymbol{y}_2) = \sum_{i \in \{0,1\}} p(\mathcal{H}_j|\widehat{\mathcal{H}}_i, \boldsymbol{y}_1, \boldsymbol{y}_2) \cdot p_{\text{BFS}}(\widehat{\mathcal{H}}_i|\boldsymbol{y}_1, \boldsymbol{y}_2). \tag{8}$$

The loss consists of two terms, the negative log-likelihood of the ground-truth hypothesis and a regularization term,

$$\mathcal{L}_{\boldsymbol{\lambda}}^{\text{UAL}} = -\log p_{\text{UAL}}(\mathcal{H}_j|\boldsymbol{y}_1, \boldsymbol{y}_2) + \beta \sum_{i \in \{0,1\}} \sum_{j \in \{0,1\}} p(\mathcal{H}_j|\widehat{\mathcal{H}}_i, \boldsymbol{y}_1, \boldsymbol{y}_2) \cdot \log p(\mathcal{H}_j|\widehat{\mathcal{H}}_i, \boldsymbol{y}_1, \boldsymbol{y}_2), \tag{9}$$

with trainable parameters denoted by $\boldsymbol{\lambda} = \left\{\boldsymbol{W}^{\text{UAL}}, \boldsymbol{bf}^{\text{UAL}}, \boldsymbol{w}_{ji}, \boldsymbol{b}_{ji}|j, i \in \{0, 1\}\right\}$. The regularization term, controlled by β, follows the maximum entropy principle to penalize the confusion matrix for returning over-confident posteriors [13].

4.4. Combined Loss Function:

All components are optimized independently w.r.t. the following combined loss:

$$\mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\psi},\boldsymbol{\phi},\boldsymbol{\lambda}} = \mathcal{L}_{\boldsymbol{\theta},\boldsymbol{\psi}}^{\text{DML}} + \mathcal{L}_{\boldsymbol{\phi}}^{\text{BFS}} + \mathcal{L}_{\boldsymbol{\lambda}}^{\text{UAL}}. \tag{10}$$

4.5. Out-of-Distribution Detector (O2D2)

Following [9], we incorporate a second neural network to detect undecidable trials. We treat the training procedure as a binary verification task. Given the learned DML, BFS and UAL components, the estimated authorship labels are obtained via

$$\widehat{a} = \arg\max\left[p_{\text{UAL}}(\mathcal{H}_0|\boldsymbol{y}_1, \boldsymbol{y}_2),\ p_{\text{UAL}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2)\right]. \tag{11}$$

Now, we can define the binary O2D2 labels as follows:

$$l^{\text{O2D2}} = \begin{cases} 1, & \text{if } a \neq \widehat{a} \quad \text{or} \quad 0.5 - \epsilon \leq p_{\text{UAL}}(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2) \leq 0.5 + \epsilon, \\ 0, & \text{otherwise.} \end{cases} \tag{12}$$

The model-dependent hyper-parameter $\epsilon \in [0.05, 0.15]$ is optimized on the validation set w.r.t the PAN 2021 metrics. The input of O2D2, noted as $\boldsymbol{y}^{\text{O2D2}}$, is a concatenated vector of the LEVs, i.e. $\left(\boldsymbol{y}_1 - \boldsymbol{y}_2\right)^{\circ 2}$ and $\left(\boldsymbol{y}_1 + \boldsymbol{y}_2\right)^{\circ 2}$, and the confusion matrix. This vector is fed into a three-layer

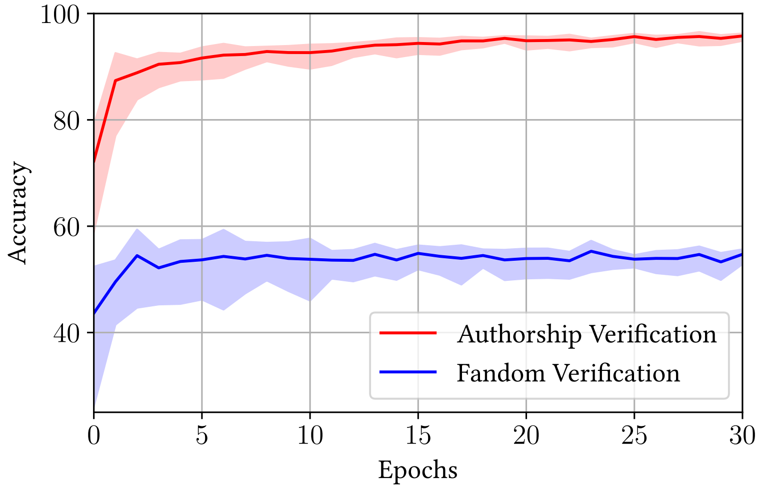


Figure 3: Averaged accuracy curves (including mean and standard deviation) for the authorship and fandom verification outputs during training.

architecture,

$$\begin{aligned} \boldsymbol{h}_1 &= \tanh\left(\boldsymbol{W}_1^{\mathrm{O2D2}}\boldsymbol{y}^{\mathrm{O2D2}} + \boldsymbol{b}_1^{\mathrm{O2D2}}\right), \\ \boldsymbol{h}_2 &= \tanh\left(\boldsymbol{W}_2^{\mathrm{O2D2}}\boldsymbol{h}_1 + \boldsymbol{b}_2^{\mathrm{O2D2}}\right), \\ p_{\mathrm{O2D2}}(\mathcal{H}_2|\boldsymbol{y}_1,\boldsymbol{y}_2) &= \mathrm{Sigmoid}\left(\boldsymbol{W}_3^{\mathrm{O2D2}}\boldsymbol{h}_2 + \boldsymbol{b}_3^{\mathrm{O2D2}}\right). \end{aligned} \tag{13}$$

All trainable parameters are summarized in $\boldsymbol{\Gamma} = \left\{\boldsymbol{W}_i^{\mathrm{O2D2}}, \boldsymbol{b}_i^{\mathrm{O2D2}} | i \in \{1,2,3\}\right\}$. The obtained prediction for hypothesis $\mathcal{H}_2$ is inserted into the cross-entropy loss,

$$\mathcal{L}_{\boldsymbol{\Gamma}}^{\mathrm{O2D2}} = l^{\mathrm{O2D2}} \cdot \log\left\{p_{\mathrm{O2D2}}(\mathcal{H}_2|\boldsymbol{y}_1,\boldsymbol{y}_2)\right\} + (1 - l^{\mathrm{O2D2}}) \cdot \log\left\{1 - p_{\mathrm{O2D2}}(\mathcal{H}_2|\boldsymbol{y}_1,\boldsymbol{y}_2)\right\}. \tag{14}$$

4.6. Ensemble Inference

As a last step, an ensemble is constructed from M trained models, $\mathcal{M}_1, \ldots, \mathcal{M}_M$, with M being an odd number. Since all models are randomly initialized and trained on different re-sampled pairs in each epoch, we expect to obtain a slightly different set of weights/biases, which in turn produces different posteriors, especially for pairs under covariate shift. We propose a majority voting for the non-responses. More precisely, the ensemble returns a non-response, if

$$\sum_{m=1}^{M} \mathbb{1}\left[p_{\mathrm{O2D2}}(\mathcal{H}_2|\boldsymbol{y}_1,\boldsymbol{y}_2,\mathcal{M}_m) \geq 0.5\right] > \left\lfloor \frac{M}{2} \right\rfloor, \tag{15}$$

where $\mathbb{1}[\cdot]$ denotes the indicator function. Otherwise, we define a subset of confident models, $\mathcal{M}_c = \{\mathcal{M} |\ p_{\mathrm{O2D2}}(\mathcal{H}_2|\boldsymbol{y}_1,\boldsymbol{y}_2,\mathcal{M}) < 0.5\}$, and return the averaged posteriors of its elements,

$$\mathbb{E}\left[p_{\mathrm{UAL}}(\mathcal{H}_1|\boldsymbol{y}_1,\boldsymbol{y}_2)\right] = \frac{1}{|\mathcal{M}_c|} \sum_{\mathcal{M} \in \mathcal{M}_c} p_{\mathrm{UAL}}(\mathcal{H}_1|\boldsymbol{y}_1,\boldsymbol{y}_2,\mathcal{M}). \tag{16}$$

Our submitted system consisted of an ensemble with $M = 21$ trained models.

Table 2: Averaged results (including mean and standard deviation) of the UAL framework for different subset combinations on the calibration dataset.

Model	PAN 2021 Evaluation Metrics					
	AUC	c@1	f_05_u	F1	Brier	overall
SA_SF + DA_DF	99.8 ± 0.0	97.5 ± 0.2	97.2 ± 0.3	97.5 ± 0.2	98.1 ± 0.1	98.0 ± 0.2
SA_SF + DA_SF	99.6 ± 0.1	95.9 ± 0.4	94.8 ± 0.6	96.0 ± 0.4	97.1 ± 0.2	96.7 ± 0.4
SA_DF + DA_DF	98.1 ± 0.1	92.4 ± 0.3	94.8 ± 0.3	92.1 ± 0.3	94.2 ± 0.2	94.3 ± 0.2
SA_DF + DA_SF	97.1 ± 0.1	90.9 ± 0.3	92.3 ± 0.6	90.6 ± 0.3	93.1 ± 0.2	92.8 ± 0.3

Table 3: Averaged calibration results (including mean and standard deviation) of the UAL framework for different subsets on the calibration dataset.

Model	Calibration Metrics			
	acc	conf	ECE	MCE
SA_SF + DA_DF	98.4 ± 0.3	97.4 ± 1.0	1.1 ± 0.9	5.7 ± 3.6
SA_SF + DA_SF	96.1 ± 0.6	95.6 ± 1.0	1.3 ± 0.8	9.9 ± 4.0
SA_DF + DA_DF	92.4 ± 0.3	93.4 ± 1.2	1.6 ± 0.6	6.2 ± 2.5
SA_DF + DA_SF	90.9 ± 0.3	92.4 ± 1.2	2.0 ± 0.7	7.4 ± 2.9

5. Experiments

The PAN evaluation metrics and procedure are described in [4, 5, 14]. To capture the calibration capacity, we also provide the *accuracy* (acc), *confidence score* (conf), *expected calibration error* (ECE) and *maximum calibration error* (MCE) [15]. All confidence values lie within the interval $[0.5, 1]$, since we are solving a binary classification task. Hence, to obtain confidence scores, the posterior values are transformed w.r.t. the estimated authorship label, showing $p(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2)$ if $\widehat{a} = 1$ and $1 - p(\mathcal{H}_1|\boldsymbol{y}_1, \boldsymbol{y}_2)$ if $\widehat{a} = 0$. For both metrics, the confidence interval is discretized into a fixed number of bins. The ECE then reflects the average absolute error between confidence and accuracy of all bins, while the MCE returns the maximum absolute error. For acc and conf, we perform weighted macro-averaging w.r.t. the number of trials in each bin.

Inspired by the promising results in domain-adversarial training of neural networks in [16, 17], we also experimented with an adversarial *fandom verifier*: Starting with the document embeddings in Eq. (1), we fed this vector into the author verification system (including DML, BFS and UAL) and into an additional *fandom verifier*, which is placed parallel to the author verification system. It has the same architecture but includes a gradient reversal layer and different trainable parameters. However, in these experiments, we did not achieve any significant improvements by domain-adversarial training. Therefore, we independently optimized the fandom verifier by stopping the flow of the gradients from the fandom verifier to the authorship verification components, so that the training of the fandom verifier does not affect the target system at all. Fig. 3 shows the obtained epoch-wise accuracies during training. It can be seen that the fandom accuracy stays around 55%, which indicates that the training strategy yields nearly topic-invariant stylometric representations, even without domain-adversarial training.

5.1. Results on the Calibration Dataset

We first evaluated the UAL component on the calibration set (without non-responses) and calculated the respective PAN metrics for different combinations of the author-fandom subsets.

Table 4: Results for PAN 2021 evaluation metrics on the validation datset.

Model		PAN 2021 Evaluation Metrics					
		AUC	c@1	f_05_u	F1	Brier	overall
single	DML	97.2 ± 0.1	91.3 ± 0.3	90.5 ± 0.6	89.6 ± 0.4	93.2 ± 0.4	92.4 ± 0.2
	BFS	97.1 ± 0.1	91.0 ± 0.3	90.7 ± 0.8	89.2 ± 0.5	93.2 ± 0.1	92.3 ± 0.2
	UAL	97.2 ± 0.1	91.3 ± 0.3	90.7 ± 0.5	89.6 ± 0.4	93.5 ± 0.2	92.5 ± 0.2
	O2D2	97.1 ± 0.1	93.8 ± 0.2	88.1 ± 0.6	93.5 ± 0.3	93.4 ± 0.1	93.2 ± 0.2
ensemble		97.8	92.5	92.1	90.9	94.3	93.5
ensemble + O2D2		97.7	94.8	90.0	94.5	94.2	94.2

Table 5: Results for the calibration metrics on the validation dataset.

Model		Calibration Metrics			
		acc	conf	ECE	MCE
single	DML	91.3 ± 0.3	87.9 ± 2.7	3.4 ± 2.7	9.0 ± 3.6
	BFS	91.0 ± 0.3	90.0 ± 2.3	2.3 ± 1.5	6.2 ± 3.0
	UAL	91.3 ± 0.3	92.3 ± 1.2	1.6 ± 0.6	5.8 ± 2.2
	O2D2	91.4 ± 0.3	90.9 ± 1.1	2.3 ± 0.5	10.7 ± 2.7
ensemble		92.5	91.2	1.2	2.9
ensemble + O2D2		92.6	91.8	1.5	10.2

Results are shown in Table 2. To guarantee that the calculated metrics are not biased by an imbalanced dataset, we reduced the number of pairs to the smallest number of pairs of all subsets. Thus, all results in Table 2 were computed from $2 \times 2,100$ pairs. Unsurprisingly, best performance was obtained for the least challenging SA_SF + DA_DF pairs and the worst performance was seen for the most challenging SA_DF + DA_SF pairs. We continued to optimize our system w.r.t this most challenging subset combination in particular, even though we specifically expect to see SA_DF + DA_DF pairs in the PAN 2021 evaluation set.

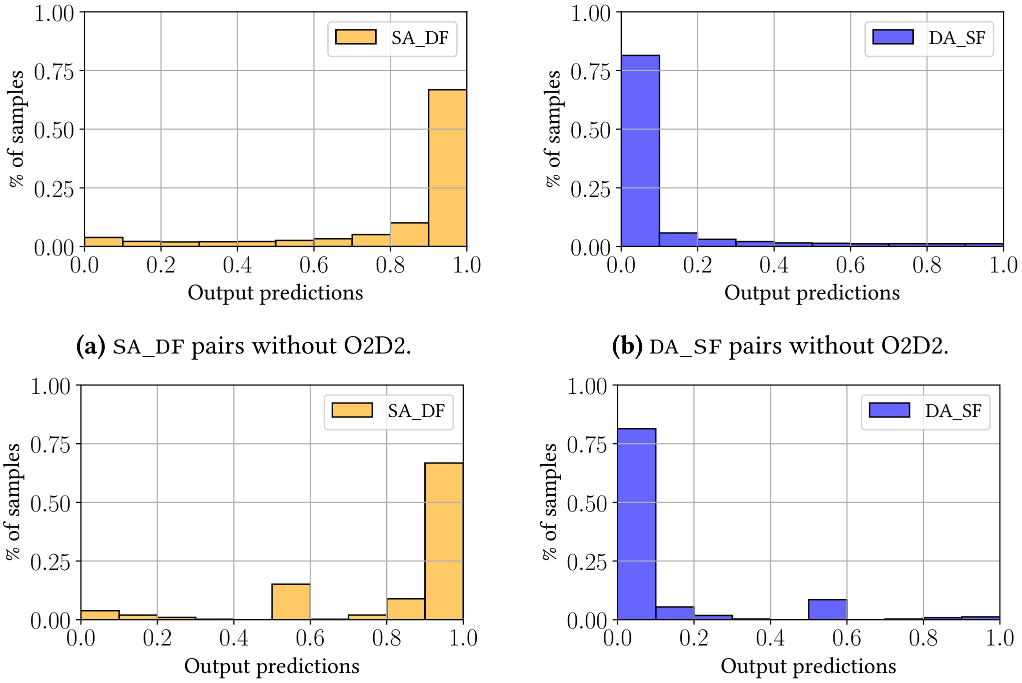
Table 3 additionally provides the corresponding calibration metrics. Analogously to the PAN metrics, the ECE consistently increases from the least to the most challenging data scenarios. Interestingly, our system is *under-confident* for SA_SF pairs, i.e. $\mathtt{conf} < \mathtt{acc}$. The predictions then change to be *over-confident* ($\mathtt{conf} > \mathtt{acc}$) for SA_DF pairs.

5.2. Results on the Validation Dataset

Next, we separately provide experimental results for all system components on the validation dataset, since O2D2 has been trained on the calibration dataset. The first four rows in Tables 4 and 5 summarize the PAN metrics and the corresponding calibration measures averaged over all ensembles models.

The overall score of the UAL component in the third row of Table 4 is on par with the DML and BFS components and slightly lower compared to the corresponding UAL score measured on the calibration dataset in Table 2. Nevertheless, we do not observe significant differences in the metrics for both datasets, which shows the robustness and generalization of our system.

Going from the third to the fourth row in Table 4, it can be observed that the overall score, boosted by c@1 and F1, significantly increases from 92.5 to 93.2. Hence, the model performs better if we take undecidable trials into account. However, the f_05_u score decreases, since it treats non-responses as false negatives. The percentage of undecidable trials generally ranges



(a) SA_DF pairs without O2D2.

(b) DA_SF pairs without O2D2.

(c) SA_DF pairs including O2D2.

(d) DA_SF pairs including O2D2.

Figure 4: Posterior histograms on the validation dataset.

from 8% to 11%.

In Table 5, we see that both, the BFS and UAL components notably improve the ECE and MCE metrics. However, an insertion of non-responses via O2D2 significantly increases the MCE. This can be explained by the posterior histograms in Fig. 4. The plots (a) and (b) show the histograms for SA_DF and DA_SF pairs without applying O2D2 to define non-responses. In contrast, plots (c) and (d) present the corresponding histograms including the 0.5-values of non-responses. The effect of O2D2 is that most of the trials, whose posteriors fall within the interval $[0.3, 0.7]$, are eventually declared as undecidable. Hence, the system correctly predicts nearly all of the remaining as confidently assigned trials around $0.7/0.8$ for same-author pairs or $0.2/0.3$ for different-author pairs. As a result, we see a large gap (i.e. $\texttt{conf} << \texttt{acc}$) between the confidence score and the averaged accuracy in these bins.

The last two rows in Tables 4 and 5 show the performance of the ensemble, first without and then with non-responses, to show the effect of O2D2. On the validation set, our ensemble with O2D2 returns non-responses in 9% of the test cases. Comparing the last two rows, we obtain the highest overall score with our proposed framework, which ultimately presents our final submission.

5.3. Results on the PAN 2021 Evaluation Dataset

To conclude this section, we present our results on the official PAN 2021 evaluation set. The performance for both, the early-bird and the final submission, can be found in Table 6. We also provide the reported result on the PAN 2020 evaluation set for the predecessor model.

Table 6: Results of the early-bird (first two rows) and the final submission runs.

Dataset	Model type	AUC	c@1	f_05_u	F1	brier	overall
Validation dataset	single-21	97.2	93.6	89.3	92.9	93.3	93.3
PAN 21 evaluation dataset	single-21	98.3	92.6	94.6	92.1	92.7	94.0
Validation dataset	ensemble-21	97.7	94.8	90.1	94.4	94.2	94.2
PAN 21 evaluation dataset	ensemble-21	98.7	95.0	93.8	95.2	94.5	95.5
PAN 20 evaluation dataset	ensemble-20	96.9	92.8	90.7	93.6	-	93.5

Unsurprisingly, the early-bird overall score (single model) on the PAN 2021 evaluation set is slightly higher, since it contains DA_DF pairs instead of DA_SF pairs. The main difference is, unexpectedly, given by the f_05_u score, which increases from 89.3% to 94.6%. In our opinion, this is caused by returning a lower number of non-responses, which would also explain the lower values for c@1 and F1.

Comparing the early-bird (2^{nd} row) with the final submission (4^{th} row), we can further significantly increase the overall score by 1.5%. We assume that the ensemble now returns a higher number of non-responses, which results in a slightly lower f_05_u score. Conversely, we can observe improved values for the c@1, F1 and brier scores.

The last row displays the achieved PAN 2020 results. As can be seen, our final submission ends up with a higher overall score (plus 2%) by significantly improving all single metrics, although the PAN competition moved from a closed-set to open-set shared task, illustrating the efficiency of the proposed extensions.

6. Conclusion

In this work, we presented O2D2, which captures undecidable trials and supports our hybrid neural-probabilistic end-to-end framework for authorship verification. We made use of the early-bird submission to receive a preliminary assessment of how the framework behaves on the novel open-set evaluation. Finally, based on the presented results, we submitted an O2D2-supported ensemble to the shared task, which clearly outperformed our own system from 2020 as well as the new submissions to the PAN 2021 AV task.

These results support our hypothesis that modeling aleatoric and epistemic uncertainty and using them for decision support is a beneficial strategy—not just for responsible ML, which needs to be aware of the reliability of its proposed decisions, but also, importantly, for achieving optimal performance in real-life settings, where distributional shift is almost always hard to avoid.

Acknowledgments

This work was in significant parts performed on an HPC cluster at Bucknell University through the support of the National Science Foundation, Grant Number 1659397. Project funding was provided by the state of North Rhine-Westphalia within the Research Training Group "SecHuman - Security for Humans in Cyberspace" and by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy - EXC2092CaSa- 390781972.

References

[1] B. Boenninghoff, J. Rupp, R. Nickel, D. Kolossa, Deep Bayes Factor Scoring for Authorship Verification, in: CLEF 2020, Notebook Papers, 2020.

[2] J. Bevendorff, B. Chulvi, G. L. D. L. P. Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2021: Authorship Verification,Profiling Hate Speech Spreaders on Twitter,and Style Change Detection, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.

[3] E. Stamatatos, A Survey of Modern Authorship Attribution Methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556.

[4] M. Kestemont, E. Manjavacas, I. Markov, J. Bevendorff, M. Wiegmann, E. Stamatatos, M. Potthast, B. Stein, Overview of the Cross-Domain Authorship Verification Task at PAN 2020, in: CLEF 2020, Notebook Papers, 2020.

[5] M. Kestemont, E. Stamatatos, E. Manjavacas, J. Bevendorff, M. Potthast, B. Stein, Overview of the Authorship Verification Task at PAN 2021, in: CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021.

[6] S. Cumani, N. Brümmer, L. Burget, P. Laface, O. Plchot, V. Vasilakakis, Pairwise Discriminative Speaker Verification in the I-Vector Space, IEEE Trans. Audio, Speech, Lang. Process. (2013).

[7] B. Luo, Y. Feng, Z. Wang, Z. Zhu, S. Huang, R. Yan, D. Zhao, Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix, in: 55th Annual Meeting of the ACL, 2017, pp. 430–439.

[8] A. Kendall, Y. Gal, What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision?, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017.

[9] Z. Shao, J. Yang, S. Ren, Calibrating Deep Neural Network Classifiers on Out-of-Distribution Datasets, ArXiv abs/2006.08914 (2020).

[10] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles, in: 31st NeurIPS, 2017, p. 6405–6416.

[11] B. Boenninghoff, D. Kolossa, Robert M. Nickel, Self-Calibrating Neural-Probabilistic Model for Authorship Verification Under Covariate Shift, in: 12th International Conference of the CLEF Association (CLEF 2021), Springer, 2021.

[12] B. Boenninghoff, S. Hessler, D. Kolossa, R. M. Nickel, Explainable Authorship Verification in Social Media via Attention-based Similarity Learning, in: IEEE International Conference on Big Data, 2019, pp. 36–45.

[13] G. Pereyra, G. Tucker, J. Chorowski, Łukasz Kaiser, G. Hinton, Regularizing Neural Networks by Penalizing Confident Output Distributions, 2017. arXiv:1701.06548.

[14] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019.

[15] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: 34th International Conference on Machine Learning, volume 70, PMLR, 2017, pp. 1321–1330.

[16] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, V. Lempitsky, Domain-Adversarial Training of Neural Networks, J. Mach. Learn. Res. 17 (2016) 2096–2030.

[17] S. Bischoff, N. Deckers, M. Schliebs, B. Thies, M. Hagen, E. Stamatatos, B. Stein, M. Potthast, The Importance of Suppressing Domain Style in Authorship Analysis, CoRR abs/2005.14714 (2020).