Final Project

Course: Computational Semantics for Natural Language Processing FS2025

Authors: Maja Gwóźdź (ETH-Z) and Gefei Wang (UZH)

Date: June 22, 2025

Detailed author contribution

Gefei Wang: feature extraction, dataset handling, the implementation of the FA model, outline of the initial proposal, full results of FA ablation experiments (including tables), tests of the FA model on the other dataset (StackOverflow), table in Appendix G, a low-resolution diagram outline of the approach and low-resolution plots of the FA ablation results, initial idea to solve the 2025 PAN shared task, the *Related Works* section (Style Change Detection, Factorized Attention), descriptions of: the PAN dataset, the FA model, feature extraction, the transformer encoder, and the *Limitations* section based on the other author's bullet points

Maja Gwóźdź: everything else (including the idea to test on a different dataset, idea to include the results from Appendix G, bullet points for the *Limitations* section), plus: full report formatting, proof-reading, TikZ diagrams, extensive editing, proposal rewriting and formatting, additional references for the *Feature Extraction* section, and task coordination (including GitHub repository management and meeting organisation)

Stylistic Shift Detection: A Hybrid Framework

Maja Gwóźdź
ETH Zürich
mgwozdz@ethz.ch

Gefei Wang
University of Zurich
gefei.wang@uzh.ch

Abstract

Authorship analysis is an important NLP task with applications in verification, plagiarism detection, and forensic linguistics. Style Change Detection (SCD) aims to identify points in a document where the author changes, which is especially challenging in multi-author writing. This project proposes the combination of three methods and their subvariants to improve SCD, namely: Factorized Attention, Optimal Transport, and Contrastive Learning. We identify the superior method and offer practical guidelines for future research.

1 Introduction

The paper contributes to the ongoing research on style shift detection. More precisely, we attempt to solve the shared task on Multi-Author Writing Style Analysis posed by PAN in 2025 (Zangerle et al., 2025). To this end, we explore three main approaches and also test their variants and combinations. We implement previously known orthogonal approaches, such as Factorized Attention (FA), Optimal Transport (OT), Contrastive Learning (CL), and perform multiple ablation experiments to identify the most promising method. Even though the algorithms are already known in the literature, the novelty of our contribution lies in the adaptation of those methods and the implementation of their combinations to the specific domain of the shared task's dataset.

We emphasize that the inclusion of multiple approaches and their combinations is not haphazard. On the contrary, each method is introduced to test our initial hypotheses, namely: we implement the FA model based on the clear evidence in the literature that the orthogonal split between content and style improves SCD accuracy (John et al., 2019). We also test OT and its derivative methods because we work with token-distribution metrics that are indifferent to word order. Again, there is enough evidence in the theoretical and empirical literature to justify this implementation (Bhardwaj et al., 2022; Montariol et al., 2021; Nouri, 2022; Shen et al., 2024).

The paper is organized as follows. We first briefly discuss the development of style shift detection research, and then cite the most important works related to the main orthogonal methods. In the next section, we provide the theoretical details of each implementation, and then analyze the experimental results. Finally, we focus on several limitations and propose concrete directions for future research. The code and pre-trained models can be found here: https://github.com/mkg33/CSNLP-ETH.

2 Related Works

2.1 Style Change Detection

In the early stages, authorship analysis and style change detection relied on stylometric techniques, which emphasized features such as function word frequency, sentence structure, and part of speech patterns to characterize stylistic differences (Stamatatos, 2009). Subsequent research proposed methods to model consistent style characteristics between documents, which provided the foundation for modern authorship attribution (Koppel et al., 2009).

Transformer-based models such as BERT and DeBERTa (He et al., 2020) have been successfully adapted to the detection of sentence-level style changes, often enhanced with task-specific fine-tuning or stylometric features (Księżniak et al., 2024; Fabien et al., 2020). Meanwhile, topic-aware debiasing techniques have been introduced to reduce semantic leakage and better isolate stylistic signals (Hu et al., 2023). Some research efforts have focused on model transparency and interpretability by using architecture-level mechanisms to identify style-relevant components (Boenninghoff et al., 2019). Others have emphasized

temporal and causal context modeling to detect gradual or contextual style drifts in longer texts (Boenninghoff et al., 2024).

The PAN shared tasks played a central role in formalizing the benchmark for multi-author style change detection. They also standardized evaluation and compared a wide range of modeling paradigms, from classical stylometry to transformer-based architectures (Zangerle et al., 2023; Księżniak et al., 2024). A recent survey of the task can be found in (Hashemi and Shi, 2025).

2.2 Factorized Attention

The idea of decoupling content from style has recently become popular in tasks such as style transfer. One effective approach (John et al., 2019) combines orthogonal constraints with adversarial training to learn independent latent spaces of style and content. Another method introduces cross-aligned autoencoders to separate style and content in non-parallel corpora (Shen et al., 2017). Recent work has explored structured attention decomposition to allocate different attention subspaces to domain-specific signals (Deng et al., 2020). This design also supports more efficient parameter sharing and allows for interpretable specialization.

2.3 Optimal Transport

Recent interest in computational[1] OT (Peyré and Cuturi, 2019; Cuturi, 2013) has led to the development of the fast Sinkhorn solver (Cuturi, 2013), which has very efficient GPU implementations. OT has proved particularly effective in computer vision (Bonneel and Digne, 2023), computer graphics (Solomon, 2018; Solomon et al., 2015; Mérigot, 2011), and even document representation (Yurochkin et al., 2019).

In NLP research, OT-based techniques (Bhardwaj et al., 2022) have been successfully applied to style transfer analysis (Nouri, 2022), authorship attribution on Weibo (Tang et al., 2019), hate speech transfer (Bose et al., 2022) (with neighborhood-aware OT), and sentiment analysis (Lazić et al., 2023), to mention but a few domains. In crude terms, OT compares two finite probability measures and constructs a minimum-cost plan of moving one measure to another, so it is not surprising that this method has also been used in diachronic semantics (Montariol et al., 2021). For instance, (Pranjić et al., 2024) study semantic shift in Slovenian and achieve satisfactory results with entropy-regularized OT, while (Kishino et al., 2024) analyze lexical shift via unbalanced OT. We also test these variants as part of the combined method.

Recently, statistical OT has also been extensively studied by the Machine Learning community (Niles-Weed and Rigollet, 2022; Chewi et al., 2024). Algorithmic advances, such as fast 1-D projections (Wu et al., 2019; Lee et al., 2019), multi-scale variants (Mérigot, 2011), or tree-based linear approximations (Otao and Yamada, 2023), provide concrete theoretical guarantees, which are required for wider adoption of OT in applied domains. These insights form the basis for the experiments in our paper. The combination of OT with CL is by no means novel. For example, (Chen et al., 2024; Shi et al., 2023; Shen et al., 2024) have already noticed that the two approaches work well in tandem. To the best of our knowledge, there are no studies that apply OT with CL to style shift detection.

3 Methods

3.1 Datasets

We used the third version[2] of the dataset provided by the authors of the Multi-Author Writing Style Analysis 2025 PAN shared task (Zangerle et al., 2025). This dataset consists of a training set and a validation set. However, since the test set was not provided, we used the validation set for testing instead. The training set contains 12600 files (with 158280 labels), while the validation set has 2700 documents (with 33654 labels). Let $\mathcal{V}$ be a finite vocabulary. A *sentence* is a finite sequence $s = (w_1, \ldots, w_\ell) \in \mathcal{V}^+$. The empirical training set is: $\mathcal{D} = \{(s_i^{(1)}, s_i^{(2)}, y_i)\}_{i=1}^N, y_i \in \{0, 1\}$, where $y_i = 1$ if and only if an author switch occurs between $s_i^{(1)}$ and $s_i^{(2)}$. The dataset presents three difficulty levels:

1. **Easy**: The sentences in the document cover various topics, which allows for the use of topic information to detect authorship changes.

2. **Medium**: The topic changes in the document are subtle, which forces the method to focus more on style to effectively solve the detection task.

[1] For a comprehensive mathematical treatment, consult (Villani, 2008). See (Bhardwaj et al., 2022) for a discussion of OT-based loss functions in NLP.

[2] Available here: https://zenodo.org/records/15053260.

3. **Hard**: All sentences in the document involve the same topic.

3.1.1 Additional Data

We also use *dataset 3* provided by the authors of the 2022 PAN shared task (Zangerle et al., 2022) to evaluate the generalization ability of our model. This dataset is based on user posts from various sites of the StackExchange network, covering different topics. We used the validation set for testing (since ground-truth labels were not provided for the original test set), which includes 1500 documents.

3.2 Baseline

(Zangerle et al., 2024) provide naïve baseline methods. One always predicts the result as 1, another always predicts 0, and the third randomly predicts either 0 or 1. We tested those baselines on the validation set:

Method	Easy	Medium	Hard
Baseline 1	0.1768	0.1773	0.1518
Baseline 0	0.4398	0.4396	**0.4509**
Baseline Random	**0.4554**	**0.4591**	0.4454

Table 1: Macro F1 scores of baseline methods across difficulty levels.

3.3 Factorized Attention

The FA model takes sentence pairs as input and outputs a binary label indicating whether the style has changed. As shown in Figure 5 the entire model can be divided into the following parts: handcrafted feature extraction, transformer encoder, content/style projections, handcrafted feature projection, and a style-aware classifier. All input sentence pairs are first used to extract features, and then the transformer encoder extracts the context-related `[CLS]` representation. These `[CLS]` representations are then projected separately into two subspaces to separate content embedding and style embedding. The previously extracted features are also mapped to a space of the same dimension as the style embedding for style supervision. The final classifier combines the style embedding with the handcrafted feature embedding to determine whether the sentence pair comes from the same author.

3.3.1 Feature Extraction

For the input sentences, we selected some features that might reflect the author's style based on (Stamatatos, 2009). These features include: average word length, number of function words, number of nouns, number of verbs, average sentence length, number of punctuation marks, and readability score. We also explored other feature combinations, but their effects were worse than those of this combination – we discuss this issue later. We extract these features for two main purposes. Firstly, they can be used to assist the final classifier in making decisions. Secondly, they can be used for style supervision. We concatenate handcrafted feature (Lee and Lee, 2023) vectors of sentence pairs and project them into the style subspace: $\mathbf{f} = \text{concat}(\mathbf{f}_1, \mathbf{f}_2) \in \mathbb{R}^{2d_{\text{feat}}}$, $\mathbf{f}' = \mathbf{W}_{\text{feat}}\mathbf{f} \in \mathbb{R}^{d_{\text{style}}}$.

3.3.2 Transformer Encoder

Given a sentence pair, we obtain the `[CLS]` embedding from a pre-trained transformer encoder: $\mathbf{h} = \text{Transformer}(x)_{\texttt{[CLS]}} \in \mathbb{R}^{d_{\text{model}}}$. To account for the trade-off between efficiency and computational resources, we used DeBERTa-v3-base (He et al., 2020) as the encoder part of the model[3]. The parameters of this model were kept frozen and were only used to extract the context-related `[CLS]` representation from sentence pairs.

3.3.3 Content/Style Projection

We project $\mathbf{h}$ into two disentangled subspaces: content and style. Formally, $\mathbf{c} = \mathbf{W}_{\text{content}}\mathbf{h} \in \mathbb{R}^{d_{\text{content}}}$ and $\mathbf{s} = \mathbf{W}_{\text{style}}\mathbf{h} \in \mathbb{R}^{d_{\text{style}}}$. The purpose of this step is to extract the required style embedding. We incorporated an orthogonality loss to obtain the disentanglement between the content and the style: $\mathcal{L}_{\text{ortho}} = \left\|\text{normalize}(\tilde{\mathbf{c}})^\top \cdot \text{normalize}(\tilde{\mathbf{s}})\right\|^2$. To ensure the alignment of the style embedding with the stylistic information that we require, we introduced the following loss: $\mathcal{L}_{\text{align}} = 1 - \cos\left(\tilde{\mathbf{s}}, \text{normalize}(\mathbf{f}')\right)$.

3.3.4 Style-aware Classifier

We define a simple classifier to combine the style embedding with the handcrafted feature embedding. The prediction is calculated from the concatenation of $\tilde{\mathbf{s}}$ and $\mathbf{f}'$: $\mathbf{z} = \text{concat}(\tilde{\mathbf{s}}, \mathbf{f}') \in \mathbb{R}^{2d_{\text{style}}}$ and $\hat{y} = \sigma(\mathbf{W}_{\text{cls}}\mathbf{z} + b)$. We train the model using a weighted sum of three losses: $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda_{\text{ortho}}\mathcal{L}_{\text{ortho}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}}$, where the classification loss is binary cross-entropy:
$\mathcal{L}_{\text{cls}} = -\left[y \log \hat{y} + (1-y)\log(1-\hat{y})\right]$.

[3]We also tested smaller models such as DeBERTa-v3-small, but their performance slightly declined compared to DeBERTa-v3-base.

3.4 Optimal Transport

We briefly discuss the basic theoretical aspects of OT and the Sinkhorn algorithm in Appendices A and B. We also relegate the details of derivative OT-based approaches to C, E, D, F, as they are not strictly required for understanding the main findings. All theoretical details are based on the following standard references: (Villani, 2008; Niles-Weed and Rigollet, 2022; Chewi et al., 2024; Wu et al., 2019; Solomon, 2018; Mérigot, 2011; Zhang et al., 2023; Cuturi, 2013; Peyré and Cuturi, 2019).

The crucial aspect of the OT-based experiments is the fact that some approaches rely on feature extraction (one variant of balanced OT, unbalanced OT, and the combination of CT with OT), while others do not require features at all. This experimental setup allows us to test the importance of feature extraction. We also implement a simple majority voting classifier to test the combination of those methods. In case of a tie, the classifier predicts the label 1.

3.4.1 Unbalanced Optimal Transport

In this approach, we inject a τ-unbalanced entropic Sinkhorn divergence between the two token clouds, and apply a vector-valued gate that independently weights each style coordinate. For every batch item b, we split the encoder output $H^{(b)} \in \mathbb{R}^{L\times h}$ by token type IDs into:

$$X^{(b)} = \{x_i^{(b)}\}_{i=1}^{m_b} \tag{1}$$

$$Y^{(b)} = \{y_j^{(b)}\}_{j=1}^{n_b} \tag{2}$$

with default uniform weights $w_X = m_b^{-1}\mathbf{1}_{m_b}$, $w_Y = n_b^{-1}\mathbf{1}_{n_b}$. If a segment is empty, we insert a single zero-vector[4].

The τ-unbalanced entropic Sinkhorn functional is defined as follows. Let $C_{ij} = \|x_i - y_j\|_2$ $(p = 1)$ and set the entropic scale to 0.03. For $\tau \geq 0$, define $\mathrm{OT}_{\varepsilon,\tau}(w_X, X;\, w_Y, Y) = \min_{\pi\geq 0}\langle C, \pi\rangle - \varepsilon H(\pi) + \tau\Big(\mathrm{KL}\big(\pi\mathbf{1}\|w_X\big) + \mathrm{KL}\big(\pi^\top\mathbf{1}\|w_Y\big)\Big)$, where:

$$H(\pi) = \sum_{ij}\pi_{ij}(\log\pi_{ij} - 1) \tag{3}$$

$$\mathrm{KL}(u\|v) = \sum_i u_i \log\frac{u_i}{v_i} - u_i + v_i. \tag{4}$$

Finally, we obtain the Sinkhorn divergence by computing:

$$\mathrm{SD}_{\varepsilon,\tau} = \mathrm{OT}_{\varepsilon,\tau}(w_X, X;\, w_Y, Y) \tag{5}$$
$$-\tfrac{1}{2}\,\mathrm{OT}_{\varepsilon,\tau}(w_X, X;\, w_X, X)$$
$$-\tfrac{1}{2}\,\mathrm{OT}_{\varepsilon,\tau}(w_Y, Y;\, w_Y, Y).$$

The model compresses this quantity via $s = \log(1+\mathrm{SD}_{\varepsilon,\tau})$.

To construct a learnable OT embedding, we compute a style-space as follows:

$$e_{\mathrm{OT}} = \mathrm{LN}\big(W_{\mathrm{OT}}\, s\big) \in \mathbb{R}^{d_s},\ W_{\mathrm{OT}} \in \mathbb{R}^{d_s\times 1}. \tag{6}$$

We also obtain the following vector-valued gate by concatenating projected features f, content c, style S, and e_{OT}: $c = W_c\, h_{[\mathrm{CLS}]}$, $S = W_s\, h_{[\mathrm{CLS}]} \in \mathbb{R}^{d_s}$. We further compute:

$$g = \sigma\Big(G\,[\,f\,\|\,c\,\|\,S\,\|\,e_{\mathrm{OT}}\,]\Big) \in (0,1)^{d_s},$$
$$m = (1-g)\odot c + g\odot S. \tag{7}$$

The final representation $r = [\,m\|f\|e_{\mathrm{OT}}\,] \in \mathbb{R}^{3d_s}$ $(= 384$ for $d_s = 128)$ is mapped by a linear head to a logit and trained via binary cross-entropy with logits.

3.4.2 Sliced Variant

Let $d \geq 1$ and fix a unit vector $v \in \mathbb{S}^{d-1} \subset \mathbb{R}^d$. Then the orthogonal projection onto the line spanned by v is given by:

$$\pi_v : \mathbb{R}^d \longrightarrow \mathbb{R}, \qquad \pi_v(z) = \langle v, z\rangle. \tag{8}$$

For probability measures α, β on $\mathbb{R}^d$ with

$$\int_{\mathbb{R}^d}\|z\|_2\, d\alpha(z) + \int_{\mathbb{R}^d}\|z\|_2\, d\beta(z) < \infty, \tag{9}$$

the one-dimensional push-forwards $\pi_{v\#}\alpha, \pi_{v\#}\beta \in \mathcal{P}(\mathbb{R})$ also have finite first moment because $\big|\pi_v(z)\big| \leq \|z\|_2$. Let us now define the **single-direction sliced** Wasserstein-1 distance by: $W_1^{(v)}(\alpha,\beta)$, which can alternatively be written as:

$$W_1\big(\pi_{v\#}\alpha,\ \pi_{v\#}\beta\big) = \int_0^1 \big|F^{-1}_{\pi_{v\#}\alpha}(t) - F^{-1}_{\pi_{v\#}\beta}(t)\big|\, dt, \tag{10}$$

where $F_\mu^{-1}(t) = \inf\{x \in \mathbb{R} : F_\mu(x) \geq t\}$ is the generalized inverse of the cumulative distribution function F_μ.

In the sliced approach, the task is to map two token clouds $X = \{x_i\}_{i=1}^m$ and $Y = \{y_j\}_{j=1}^n \subset \mathbb{R}^d$ to a single scalar. To solve this problem, we first draw random directions. We fix $n_\pi \geq 1$ and sample i.i.d. $v_1, \ldots, v_{n_\pi} \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ by taking a Gaussian matrix $V \sim \mathcal{N}(0,1)^{n_\pi\times d}$ and normalizing each row to unit Euclidean length. In the next

[4]Note that the weight of this vector is 1, so sum normalization is preserved.

step, we compute the one-dimensional projections for every $k \in \{1, \dots, n_\pi\}$:

$$\pi_{v_k}(x_i) = \langle v_k, x_i \rangle,\ \pi_{v_k}(y_j) = \langle v_k, y_j \rangle. \quad (11)$$

These projections are then sorted and trimmed, as described below. Given finite token clouds: $X = \{x_1, \dots, x_m\} \subset \mathbb{R}^d$ and $Y = \{y_1, \dots, y_n\} \subset \mathbb{R}^d$, let

$$x^{\uparrow,k} = \big(\pi_{v_k}(x_1), \dots, \pi_{v_k}(x_m)\big)^{\uparrow} \quad (12)$$

$$y^{\uparrow,k} = \big(\pi_{v_k}(y_1), \dots, \pi_{v_k}(y_n)\big)^{\uparrow}, \quad (13)$$

where $\big(\cdot\big)^{\uparrow}$ denotes the non-decreasing sort with ties broken arbitrarily (since the final mean is permutation-invariant). We keep $m_* := \min\{m, n\}$ elements of each sorted list and average their absolute differences as follows[5]:

$$\widehat{\mathrm{SW}}_1^{(n_\pi)}(X, Y) = \frac{1}{n_\pi\, m_*} \sum_{k=1}^{n_\pi} \sum_{i=1}^{m_*} \big| x_i^{\uparrow,k} - y_i^{\uparrow,k} \big|. \quad (14)$$

We note in passing that for $m = n$, the inner sum recovers the usual one-dimensional Wasserstein distance:

$$\frac{1}{m} \sum_{i=1}^{m} \big| x_i^{\uparrow,k} - y_i^{\uparrow,k} \big| = W_1\big(\pi_{v_k\#}\hat{\alpha}_m, \pi_{v_k\#}\hat{\beta}_m\big), \quad (15)$$

which implies that $\widehat{\mathrm{SW}}_1^{(n_\pi)}$ is precisely the Monte-Carlo approximation of the sliced W_1 distance.

4 Contrastive Learning with Sinkhorn

We propose yet another combined method that augments CL with OT. The implementation of the CL part is based on (Chen et al., 2020). Let a mini-batch contain B (≥ 2) sentence pairs, B embeddings $z_i \in \mathbb{S}^{127}$, and binary labels $y_i \in \{0, 1\}$. Positives are all pairs that share the same label, $\mathcal{P} = \{(i, j) \mid i \neq j,\ y_i = y_j\}$. We then fix a temperature of $\tau = 0.08$ and exclude the diagonal term $k = i$:

$$\mathcal{L}_{\mathrm{contr}} = -\frac{1}{|\mathcal{P}|} \sum_{(i,j)\in\mathcal{P}} \log \frac{\exp(\langle z_i, z_j \rangle/\tau)}{\sum_{k\neq i} \exp(\langle z_i, z_k \rangle/\tau)}. \quad (16)$$

We also construct the following bias-free two-layer projection head:

$$z \ = \ \frac{W_2\ \mathrm{ReLU}(W_1 s)}{\|W_2\ \mathrm{ReLU}(W_1 s)\|_2}, \quad (17)$$

$$W_1, W_2 \in \mathbb{R}^{128\times128},\ \text{bias} = 0.$$

We introduce the scalar gate $g = \sigma(\|f\|_2)$ that controls only the classification path. To be more precise, the classifier sees $\tilde{s} = g\, s$, while z is computed from the raw style vector s. From the balanced, debiased Sinkhorn distance $d \geq 0$, we build:

$$u \ = \ \mathrm{LayerNorm}_{\gamma,\beta}\big(W_{\mathrm{OT}} \log(1 + d) + b_{\mathrm{OT}}\big),$$

$$W_{\mathrm{OT}} \in \mathbb{R}^{128\times1},\ b_{\mathrm{OT}}, \gamma, \beta \in \mathbb{R}^{128}. \quad (18)$$

(`LayerNorm` uses $\varepsilon = 10^{-5}$ by default.)
Finally, suppose we have logits $\ell \in \mathbb{R}$, then the binary cross-entropy term is:

$$\mathcal{L}_{\mathrm{BCE}} = -\tfrac{1}{B} \sum_{i=1}^{B} \big[y_i \log \sigma(\ell_i) + (1 - y_i) \log(1 - \sigma(\ell_i)) \big]. \quad (19)$$

During training, our aim is to minimize:
$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}} + 0.1\, \mathcal{L}_{\mathrm{contr}}$.

5 Results

All reported results have been rounded up to the nearest ten-thousandth. We call the model with the best performance on the Hard set the *best model*. Under *best overall model*, we understand the model with the greatest average performance. Unless indicated otherwise, all models were trained for three epochs.

For full transparency and to allow for some degree of comparison, we cite the results obtained by the top teams in a similar shared task (see G). However, note that the task involved a different dataset and that some authors trained their models on additional data, which we have deliberately not done.

5.0.1 Factorized Attention

To assess the generalization capability of our FA model across varying difficulty levels, we report the macro F1 score on the Easy, Medium, and Hard validation subsets. As shown in Figure 1, different weights of orthogonality loss result in significant variation in F1 scores. In addition, we explored different dimensions of style embedding and various feature extraction approaches. These results are also presented in Figure 1. Observe that the best model shares the best average macro F1 score.

Feature Set 1 refers to the original feature extraction set, which includes average word length, number of function words, number of nouns, number of verbs, average sentence length, number of

[5] Given the assumption that each sentence contains at least one lexical token, the denominator never vanishes since m_* is positive.

Model Variant	Features	OT Variant	Macro F1			
			Easy	Medium	Hard	Mean
Balanced OT (blur=0.05)	✓	balanced Sinkhorn	0.9050	0.8114	0.7644	0.8270
Unbalanced OT ($\tau = 0.8$)	✓	unbalanced Sinkhorn	**0.9134**	**0.8224**	0.7853	**0.8404**
CL + OT ($\lambda = 0.1$)	✓	balanced Sinkhorn	0.9024	0.8215	**0.7871**	0.8370
Balanced OT (blur=0.03)	×	balanced Sinkhorn	0.9118	0.8252	0.7698	0.8356
Sliced OT ($n_\pi = 64$, style_dim = 64)	×	sliced W_1	0.9122	0.8260	0.7725	0.8369
Max-proj OT ($n_\pi = 128$, style_dim = 64)	×	max-sliced W_1	0.9133	**0.8273**	**0.7731**	**0.8379**
Multi-scale OT ($n_\pi \in \{8, 32, 128\}$)	×	multi-scale sliced	**0.9140**	0.8262	0.7715	0.8372
Multi-OT (5 epochs) + CL+OT + 2×UB + B (no features)	×/✓	mixed	0.9179	0.8286	0.7933	0.8467
Multi-OT (7 epochs) + 2×CL+OT + 2×UB + B (no features)	×/✓	mixed	0.9186	0.8299	0.7942	0.8476
Multi-OT (5 epochs) + 2×CL+OT + 2×UB + B (no features)	×/✓	mixed	**0.9186**	**0.8301**	**0.7943**	**0.8477**

Table 2: Macro F1 scores and their arithmetic mean for selected configurations in each model family. n_π denotes the number of random directions. Recall that blur = $\sqrt{\varepsilon}$, where ε is the parameter in the entropy-regularized 1-Wasserstein cost. Intuitively, τ is the penalty for creating or destroying mass.
Ensemble results: UB – Unbalanced OT, B – Balanced OT. The number of times represents the voting weight.

punctuation marks, and readability score. *Feature Set 2* is customized for comments on social platforms, including the number of function words, the number of punctuation marks, the type-token ratio, average sentence length, readability score, the uppercase ratio, number of slang terms, number of URLs, whether there is a subject, and the type of the subject. *Feature Set 3* was obtained by modifying *Feature Set 2*, but with the number of slang terms and the number of URLs removed. Instead, part of speech n-gram features were added.

5.0.2 Optimal Transport

We report the macro F1 scores for each OT-based method in Table 2. The best model in the feature-informed group was CL with OT, while its counterpart in the feature-free group was the maximum projection OT variant. The latter reached the highest mean score and the best performance on the Hard set in the feature-free family. It is interesting to note that the maximum projection OT variant did not contribute to the best ensemble models.

5.1 Additional Dataset

As expected, the models underperformed on the additional StackExchange validation dataset. Recall that we did not use any training examples from this dataset. The best model reached an F1 macro score of 0.4168, while an ensemble model achieved 0.4211. We provide the remaining details in Appendix H.

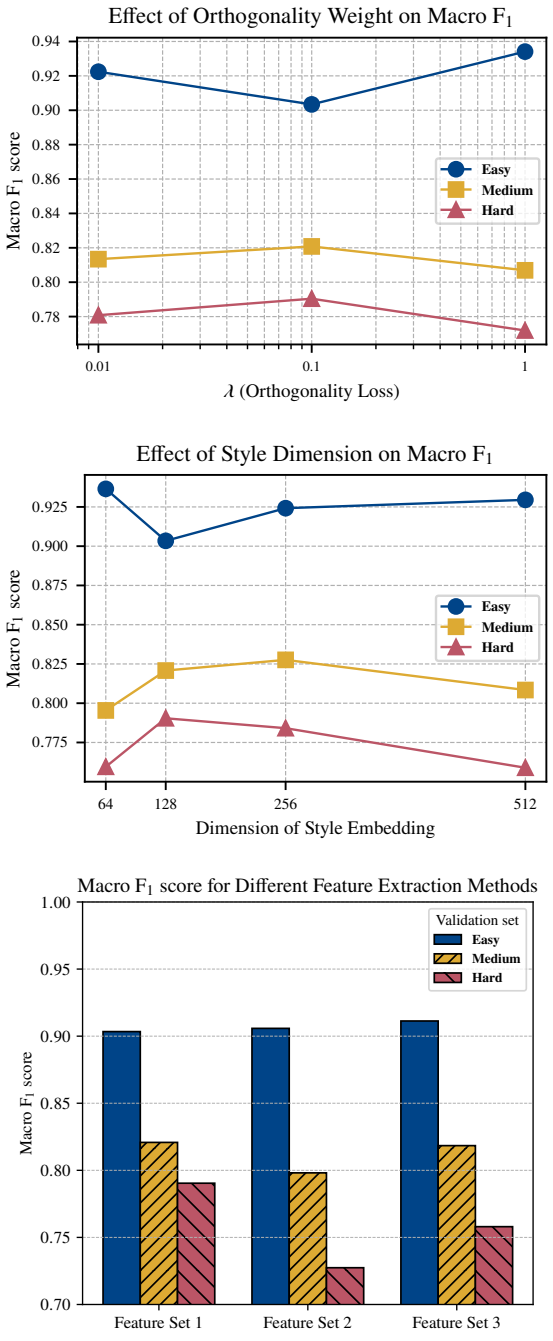


Figure 1: Impact of different hyperparameters on the macro F1 score.

Model Variant	Easy	Medium	Hard	Mean
CLS-only	0.9167	0.8216	0.7669	0.8351
CLS + Orthogonality	0.9025	**0.8258**	0.7852	0.8378
CLS + Style Features + Content	**0.9179**	0.8015	0.7551	0.8248
CLS + Orthogonality + Style Features (Full Model)	0.9034	0.8208	**0.7904**	**0.8382**

Table 3: Macro F1 scores across difficulty levels for different ablations.

5.2 Ablation Experiments

5.2.1 Factorized Attention

To evaluate the contribution of different components in our FA model, we conduct an ablation study using the macro F1 score on three subsets of the validation set: Easy, Medium, and Hard. We compare the full FA model against several ablated variants:

- **CLS-only**: A simple classifier using only the CLS token embedding from the transformer.
- **CLS + Orthogonality**: Applies an orthogonality constraint between style and content projections.
- **CLS + Style Features + Content Embedding**: Uses content projection along with style features, but not style embedding.

5.2.2 Optimal Transport

We relegate the full ablation results to Appendix I since the effects of the respective hyperparameter variations (orthogonality loss, style dimension, projection dimension) were negligible. However, the ablation experiment involving the blur hyperparameter in the balanced OT model with no features yielded statistically significant results. The details are presented in Figure 2.

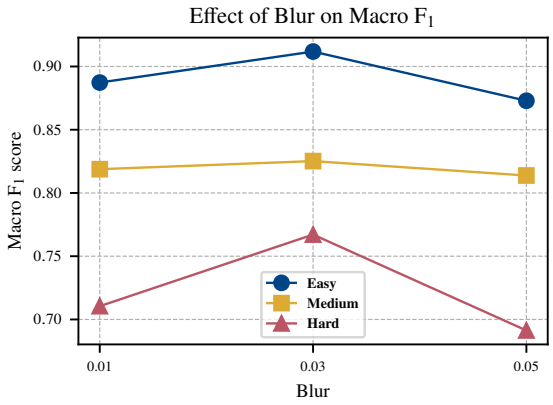


Figure 2: Impact of the blur hyperparameter on the macro F1 score in the balanced OT model with no features.

6 Discussion

6.1 Factorized Attention

Figure 1 implies that the classification of Medium and Hard sets benefits from a moderate orthogonality, while the Easy set uses the strongest penalty. This observation could be explained by recalling that the topical variance is the highest in the Easy dataset, while it gets progressively weaker in the remaining splits. We conclude that the orthogonality term controls content leakage. The practical implication is that λ should be increased in experiments involving clear topical differences, and it ought to stay between 0.01 and 0.1 for inputs with fewer topical cues.

As regards the style dimension, one immediately notices that 64 dimensions suffice for an optimal prediction of the Easy set, whereas the other sets require up to $128-256$ but deteriorate significantly in higher dimensions. This implies a low-rank stylistic signal, *i.e.,* additional capacity simply "memorizes" topical artifacts, which negatively impacts generalization.

Finally, we obtain relatively stable performance on the three feature sets. The most significant difference lies in the score of the Hard dataset, where the greatest score is in the set with the original features, and the lowest is seen in the one augmented with social platform-specific features. A plausible explanation of this trend is that social media markers actually *decrease* performance on topically homogeneous sets but have a negligible positive effect on performance in the case of more heterogeneous sets.

6.2 Optimal Transport

The best model obtained by OT with CL demonstrates that the combination of orthogonal representations works well on sets with fewer stylistic cues. This result is explained by the fact that the two approaches work on complementary signals. Unsurprinsgly, the ablation experiments w.r.t. the orthogonality weight, the number of random projections, and the style dimension produced very simi-

lar results (within 0.003 of each row mean). This outcome is expected since the OT scalar dominates and explains the greatest part of between-class variance. In the feature-free group, every method converges to the same results when rounded up to the hundredth, which confirms that a single OT scalar captures the discriminative signal. Interestingly, the feature-free methods outperform the best FA model on both the Easy and Medium datasets, but fare worse by ≈ 0.017 on the Hard set.

The results imply that feature-free methods could potentially generalize better in similar datasets (for instance, comments from different social media platforms). However, they reach significantly worse scores than FA in more specialized contexts, as shown by the StackExchange evaluation.

As regards blur tuning in the feature-free balanced OT method, the results are fully consistent with Sinkhorn theory. We obtain larger bias with growing $\varepsilon = \texttt{blur}^2$, and higher variance once blur converges to 0, hence the intermediate value of 0.03 performs best in this scenario.

The evidence that the unbalanced approach outperforms its balanced counterpart (both using features) suggests that letting the OT plan change mass preserves stylistic cues, while the balanced constraint diffuses the mass across lower-density tokens. The balanced version might still perform well if the two distributions have similar support. The unbalanced OT model is also the best overall model. The fact that the OT-based ensemble outperforms single models is expected since weighted voting favors the strongest method but the conjunction of models still allows for diversity. It is clear that this research direction ought to be explored in the future, especially with more advanced OT techniques.

7 Limitations

Unarguably, our approach suffers from several limitations. First and foremost, due to limited computational resources, we were unable to perform an exhaustive hyperparameter search or train a larger model. As a result, our models do not reflect optimal performance. We tested only one feature extraction method but more approaches are required for a comprehensive study. Similarly, the OT-based algorithms were fairly standard. More implementations are needed for a full empirical comparison.

Moreover, while our experiments span multiple difficulty levels, the overall volume of training data remains relatively low. Given that the current dataset is entirely derived from Reddit comments, the generalization ability of our model is largely constrained. Finally, since our study focuses on sentence pairs, extending this method to paragraph- or document-level segmentation may require more sophisticated attention mechanisms and memory modeling to capture longer-range dependencies.

8 Future Research

We propose several research directions that seem particularly promising based on our findings. Given the low performance of our methods on a dataset from another stylistic and topical domain (StackExchange *vs* Reddit), it is immediately clear that the approach has to be adapted to other domains, both specialized and general, in order to achieve satisfactory cross-domain performance. One obvious solution would be to create varied and balanced training datasets.

The analyses clearly demonstrate that OT and its combination with orthogonal approaches reaches relatively high scores. Given the promising scores of the ensembles, it would be reasonable to introduce other feature extraction methods and test their impact on the combined OT-driven approach. Further experiments on general domains are required to test the hypothesis whether feature-free OT methods suffice for successful style shift detection. Their relatively short training time ($\approx$1h/epoch on NVidia GTX 1080 Ti) appears promising.

By extension, it would be appealing to apply the OT-augmented methods to tasks from computational diachronic semantics (Kishino et al., 2024). We hypothesize that the OT method could work particularly well in sense shift detection, diachronic lexical substitution, or even the detection of multilingual influence. The only caveat is that one has to identify appropriate corpora and that the source material has to cover sufficiently many years for the diachronic changes to emerge. This issue of the training corpus is closely related to one of the shortcomings of our approach, namely, the method currently handles sentence-level authorship changes but it has to be considerably extended to perform well on corpora that do not offer such fine-grained annotations, and might even lack explicit authorship labels. This challenge needs to be addressed before applying the OT-augmented variants to any diachronic corpora.

References

Rishabh Bhardwaj, Tushar Vaidya, and Soujanya Poria. 2022. Towards solving nlp tasks with optimal transport loss. *Journal of King Saud University-Computer and Information Sciences*, 34(10):10434–10443.

Benedikt Boenninghoff, Steffen Hessler, Dorothea Kolossa, and Robert M. Nickel. 2019. Explainable authorship verification in social media via attention-based similarity learning. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 36–45.

Benedikt Boenninghoff, Henry Hosseini, Robert Nickel, and Dorothea Kolossa. 2024. Who wrote when? author diarization in social media discussions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15721–15734.

Nicolas Bonneel and Julie Digne. 2023. A survey of optimal transport for computer graphics and computer vision. In *Computer Graphics Forum*, volume 42, pages 439–460. Wiley Online Library.

Tulika Bose, Irina Illina, and Dominique Fohr. 2022. Transferring knowledge via neighborhood-aware optimal transport for low-resource hate speech detection. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 453–467, Online only. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR.

Zihao Chen, Chi-Heng Lin, Ran Liu, Jingyun Xiao, and Eva Dyer. 2024. Your contrastive learning problem is secretly a distribution alignment problem. *Advances in Neural Information Processing Systems*, 37:91597–91617.

Sinho Chewi, Jonathan Niles-Weed, and Philippe Rigollet. 2024. Statistical optimal transport. *arXiv preprint arXiv:2407.18163*.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26.

Yongchao Deng, Hongfei Yu, Heng Yu, Xiangyu Duan, and Weihua Luo. 2020. Factorized transformer for multi-domain neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4221–4230, Online. Association for Computational Linguistics.

Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: BERT fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 127–137.

Ahmad Hashemi and Wei Shi. 2025. A survey on writing style change detection: Current literature and future directions. *Machine Intelligence Research*, 22(3):397–416.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Xinyu Hu, Weihan Ou, Sudipta Acharya, Steven HH Ding, Ryan D'Gama, and Hanbo Yu. 2023. Tdrlm: Stylometric learning for authorship verification by topic-debiasing. *Expert Systems with Applications*, 233:120745.

Yingzhou Huang and Leilei Kong. 2024. Team text understanding and analysis at PAN: Utilizing BERT series pre-training model for multi-author writing style analysis. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled representations for style transfer in text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, page 424–434. Association for Computational Linguistics.

Ryo Kishino, Hiroaki Yamagiwa, Ryo Nagata, Sho Yokoi, and Hidetoshi Shimodaira. 2024. Quantifying lexical semantic shift via unbalanced optimal transport. *arXiv preprint arXiv:2412.12569*.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for Information Science and Technology*, 60(1):9–26.

Ewelina Księżniak, Krzysztof Węcel, and Marcin Sawiński. 2024. Team openfact at PAN 2024: Fine-tuning BERT models with stylometric enhancements. In *CEUR Workshop Proceedings*, volume 3740.

Jelena Lazić, Aleksandra Krstić, and Sanja Vujnović. 2023. Sentiment analysis using optimal transport loss function. In *2023 10th International Conference on Electrical, Electronic and Computing Engineering (IcETRAN)*, pages 1–5. IEEE.

Bruce W Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19.

Chen-Yu Lee, Tanmay Batra, Mohammad Haris Baig, and Daniel Ulbricht. 2019. Sliced Wasserstein discrepancy for unsupervised domain adaptation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 10285–10295. Computer Vision Foundation / IEEE.

Tzu-Mi Lin, Yu-Hsin Wu, and Lung-Hao Lee. 2024. Team NYCU-NLP at PAN 2024: Integrating transformers with similarity adjustments for multi-author writing style analysis. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jiajun Lv, Yusheng Yi, and Haoliang Qi. 2024. Team fosu-stu at PAN: Supervised fine-tuning of large language models for multi author writing style analysis. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Quentin Mérigot. 2011. A multiscale approach to optimal transport. In *Computer graphics forum*, volume 30, pages 1583–1592. Wiley Online Library.

Syrielle Montariol, Matej Martinc, and Lidia Pivovarova. 2021. Scalable and interpretable semantic change detection. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4642–4652, Online. Association for Computational Linguistics.

Jonathan Niles-Weed and Philippe Rigollet. 2022. Estimation of Wasserstein distances in the spiked transport model. *Bernoulli*, 28(4):2663–2688.

Nasim Nouri. 2022. Text style transfer via optimal transport. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2532–2541, Seattle, United States. Association for Computational Linguistics.

Sho Otao and Makoto Yamada. 2023. A linear time approximation of Wasserstein distance with word embedding selection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15121–15134, Singapore. Association for Computational Linguistics.

Gabriel Peyré and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Marko Pranjić, Kaja Dobrovoljc, Senja Pollak, and Matej Martinc. 2024. Semantic change detection for Slovene language: a novel dataset and an approach based on optimal transport. *arXiv preprint arXiv:2402.16596*.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6833–6844, Red Hook, NY, USA. Curran Associates Inc.

Xiaorong Shen, Maowei Huang, Zheng Hu, Shimin Cai, and Tao Zhou. 2024. Multimodal fake news detection with contrastive learning and optimal transport. *Frontiers in Computer Science*, Volume 6 - 2024.

Liangliang Shi, Gu Zhang, Haoyu Zhen, Jintao Fan, and Junchi Yan. 2023. Understanding and generalizing contrastive learning from the inverse optimal transport perspective. In *International conference on machine learning*, pages 31408–31421. PMLR.

Justin Solomon. 2018. Optimal transport on discrete domains. *AMS Short Course on Discrete Differential Geometry*, 3.

Justin Solomon, Fernando De Goes, Gabriel Peyré, Marco Cuturi, Adrian Butscher, Andy Nguyen, Tao Du, and Leonidas Guibas. 2015. Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics (ToG)*, 34(4):1–11.

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

Wanbing Tang, Chunhua Wu, Xiaolong Chen, Yudao Sun, and Chen Li. 2019. Weibo authorship identification based on Wasserstein generative adversarial networks. In *2019 IEEE International Conference on Signal, Information and Data Processing (ICSIDP)*, pages 1–5.

Cédric Villani. 2008. *Optimal transport: old and new*, volume 338. Springer.

Jiqing Wu, Zhiwu Huang, Dinesh Acharya, Wen Li, Janine Thoma, Danda Pani Paudel, and Luc Van Gool. 2019. Sliced Wasserstein generative models. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3713–3722. Computer Vision Foundation / IEEE.

Zhanhong Ye, Yutong Zhong, Chen Huang, and Leilei Kong. 2024. Continual transfer learning with progress prompt for multi-author writing style analysis. In *Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum*, volume 3740 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin M Solomon. 2019. Hierarchical optimal transport for document representation. *Advances in neural information processing systems*, 32.

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2022. Overview of the Style Change Detection Task at PAN 2022. In *CLEF 2022 Labs and Workshops, Notebook Papers*, volume 3180 of *CEUR Workshop Proceedings*. CEUR-WS.org.

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

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2024. Overview of the multi-author writing style analysis task at PAN 2024. In *Working Notes Papers of the CLEF 2024 Evaluation Labs*, pages 2513–2522. CEUR-WS.org.

Eva Zangerle, Maximilian Mayerl, Martin Potthast, and Benno Stein. 2025. Multi-author writing style analysis 2025. https://pan.webis.de/clef25/pan25-web/style-change-detection.html. PAN at CLEF 2025 - Multi-Author Writing Style Analysis.

Jingyi Zhang, Ping Ma, Wenxuan Zhong, and Cheng Meng. 2023. Projection-based techniques for high-dimensional optimal transport problems. *Wiley Interdisciplinary Reviews: Computational Statistics*, 15(2):e1587.

A Optimal Transport

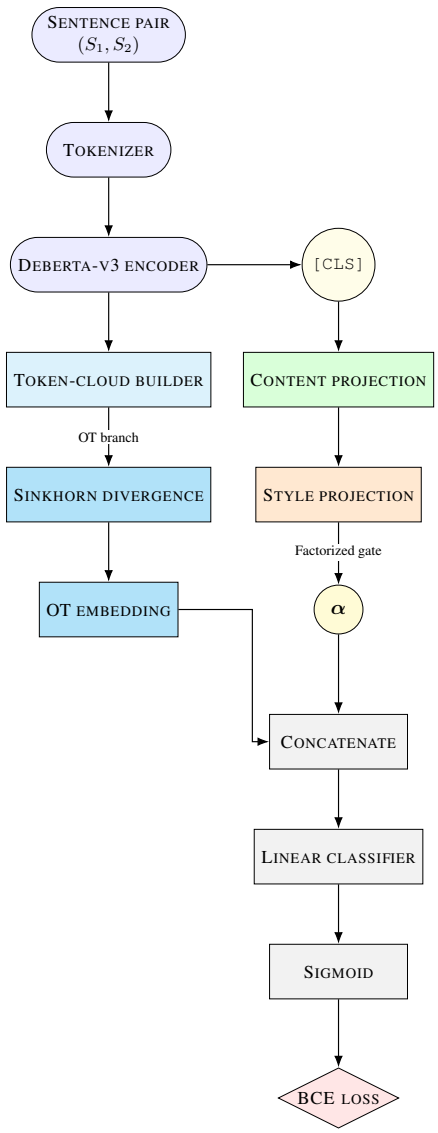


Figure 3: A diagram showing the balanced Optimal Transport method with feature extraction.

B The Sinkhorn Algorithm

The following summary is based on (Cuturi, 2013). Let $X = \{x_i\}_{i=1}^m \subset \mathbb{R}^d$ and $Y = \{y_j\}_{j=1}^n \subset \mathbb{R}^d$ be two finite point clouds with weights $w_X \in \Delta_m$ and $w_Y \in \Delta_n$ ($\Delta_k = \{u \in \mathbb{R}_+^k \mid \sum_i u_i = 1\}$). For the ℓ_2-ground-cost $C_{ij} = \|x_i - y_j\|_2$, we introduce the Gibbs kernel:

$$K_{ij} \;=\; \exp\bigl(-C_{ij}/\varepsilon\bigr), \qquad \varepsilon > 0. \tag{20}$$

Let us now analyze the iterations. We begin with $u^{(0)} = \mathbf{1}_m$, $v^{(0)} = \mathbf{1}_n$, and update for $t = 0, 1, \ldots$:

$$u^{(t+1)} \;=\; \frac{w_X}{K v^{(t)}},\; v^{(t+1)} \;=\; \frac{w_Y}{K^\top u^{(t+1)}}. \tag{21}$$

We halt once the marginal error $\max\bigl\{\|u^{(t)} \odot (K v^{(t)}) - w_X\|_1,\; \|v^{(t)} \odot (K^\top u^{(t)}) - w_Y\|_1\bigr\} < \eta$, with tolerance of $\eta = 10^{-3}$.

Once the algorithm converges, we set: $\pi^\star = \operatorname{diag}(u)\, K\, \operatorname{diag}(v) \in \mathbb{R}_{\geq 0}^{m \times n}$. The cost is precisely the entropic OT cost:

$$\mathrm{OT}_\varepsilon(X, Y) \;=\; \langle C, \pi^\star \rangle. \tag{22}$$

Finally, to remove the entropic bias, we compute:

$$\mathrm{SD}_\varepsilon(X, Y) = \mathrm{OT}_\varepsilon(X, Y) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(X, X) - \frac{1}{2}\,\mathrm{OT}_\varepsilon(Y, Y),$$

which is symmetric, non-negative, and vanishes if and only if the two empirical measures coincide.

Note that in the unbalanced implementation, `GeomLoss` solves the dual fixed-point equations with a scaling parameter of 0.95. This method affects the contraction rate of Sinkhorn iterations but does not alter the objective.

C Balanced Optimal Transport

Note that this is the feature-free variant (see also A). For discrete weighted clouds $(w_X, X) = \{(w_{X,i}, x_i)\}_{i=1}^m$ and $(w_Y, Y) = \{(w_{Y,j}, y_j)\}_{j=1}^n \subset \mathbb{R}^h$ with uniform weights $w_{X,i} = 1/m$, $w_{Y,j} = 1/n$, let $C_{ij} = \|x_i - y_j\|_2$ and fix $\varepsilon = 0.05$. The balanced, entropically regularized OT cost is given by:

$$\mathrm{OT}_\varepsilon(w_X, X;\, w_Y, Y) \;=\; \min_{\pi \geq 0} \bigl\langle C, \pi \bigr\rangle$$

$$-\varepsilon \sum_{i,j} \pi_{ij} \bigl(\log \pi_{ij} - 1\bigr), \tag{23}$$

$$\text{s.t. } \pi \mathbf{1} = w_X,\; \pi^\top \mathbf{1} = w_Y.$$

No KL penalty appears because we set $\tau = 0$. It suffices to replace $\mathrm{OT}_{\varepsilon,\tau}$ by OT_ε ($\tau = 0$). Let us now denote:

$$\mathrm{OT}_\varepsilon(X, Y) \equiv \mathrm{OT}_\varepsilon(w_X, X;\, w_Y, Y) \text{ with } \varepsilon = 0.05$$

In the code, we call `SampleLoss` with the `sinkhorn` and `debias=True` arguments, so `GeomLoss` actually returns the Sinkhorn divergence:

$$d_{\mathrm{OT}} = \mathrm{SD}_\varepsilon(X, Y) := \mathrm{OT}_\varepsilon(X, Y) - \tfrac{1}{2}\,\mathrm{OT}_\varepsilon(X, X) - \tfrac{1}{2}\,\mathrm{OT}_\varepsilon(Y, Y), \quad (24)$$

which is symmetric, non-negative, and vanishes if and only if the two weighted empirical measures coincide. We compute the self-costs $\mathrm{OT}_\varepsilon(X, X)$ and $\mathrm{OT}_\varepsilon(Y, Y)$ with the same ε and uniform weights. In the final steps, we apply the same element-wise map as in the unbalanced version, the learnable embedding, and vector-valued gate with the classifier.

D Multi-scale Optimal Transport

In the multi-scale approach, we reuse the trimmed empirical estimator from the sliced Wasserstein-1 implementation. We compute the estimator for $n_\pi \in \{8, 32, 128\}$ random directions v and concatenate the three scalar distances. Intuitively, the scalar distances could be interpreted as coarse/medium/fine granularity, respectively. The fast low-projection regime emphasizes a large global mismatch between the two clouds, while the high-projection one is more computationally expensive but captures more detailed geometry.

We combine the multi-scale approach with the **fused latent representation** (denoted as $F^{(b)}$). For each mini-batch index b let:

$$C^{(b)} = W_c\, r^{(b)} \in \mathbb{R}^{d_c}, \qquad S^{(b)} = W_s\, r^{(b)} \in \mathbb{R}^{d_s},$$

where $W_c \in \mathbb{R}^{d_c \times 2h}$, $W_s \in \mathbb{R}^{d_s \times 2h}$ and $r^{(b)} \in \mathbb{R}^{2h}$ is the pair feature. We use default latent sizes of $d_c = 192$ (**content**) and $d_s = 128$ (**style**). We also define: $c = W_{\text{content}}\, h_{\text{[CLS]}} \in \mathbb{R}^{d_s}$, $s_{\text{sty}} = W_{\text{style}}\, h_{\text{[CLS]}} \in \mathbb{R}^{d_s}$. We now introduce a single scalar gate γ that controls mixing $g = \sigma(\gamma) \in (0, 1)$ and

$$F = [\, C(1-g) \| \; Sg \,], \quad C = W_c r,\; S = W_s r.$$

In early training, we have $g \approx 0.5$. The network learns whether stylistic or content dimensions better explain the task by adjusting γ. Finally, we define the fused latent representation:

$$F^{(b)} \;=\; \big[\, C^{(b)}(1-g) \,\big\|\, S^{(b)} g \big] \;\in \mathbb{R}^{d_c + d_s}, \quad (25)$$

where the total dimension is $192 + 128 = 320$. Intuitively, one could interpret the values of g as follows. When $g \to 0 \implies F^{(b)} \approx C^{(b)}$, so the style is suppressed. On the other hand, if $g \to 1 \implies F^{(b)} \approx S^{(b)}$, then the content gets suppressed. Once we have formed $F^{(b)}$, we concatenate it with the three-dimensional OT vector $d_{\mathrm{OT}}^{(b)}$ (that is, the scalar distance described above), and obtain a 323-dimensional input to the `LayerNorm` (with the learnable scale γ and a shift parameter β). We use the `LayerNorm` to whiten the vector before applying the 2×323 linear classifier. More precisely, we have: $z^{(b)} = \mathrm{LN}\big[F^{(b)} \| \; d_{\mathrm{OT}}^{(b)}\big]$ and

$$\ell^{(b)} = W_z z^{(b)} + b_z,\; \hat{p}^{(b)} = \mathrm{softmax}(\ell^{(b)}). \quad (26)$$

E Maximum Projection

We also implement a nontrivial variant of the sliced approach. Assume we have two probability measures $\alpha, \beta \in \mathcal{P}(\mathbb{R}^d)$ with finite first moment. The max-sliced Wasserstein-1 distance selects the projection that gives the largest one-dimensional mismatch:

$$W_{1,\max}(\alpha, \beta) := \sup_{v \in \mathbb{S}^{d-1}} W_1\big(\pi_{v\#}\alpha, \pi_{v\#}\beta\big). \quad (27)$$

Our code approximates this equation by first drawing $n_\pi \geq 1$ random directions $v_1, \ldots, v_{n_\pi} \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}(\mathbb{S}^{d-1})$, which are implemented as row-normalized Gaussian vectors.

For each direction, we then compute the trimmed empirical distance $\widehat{W}_{1,\mathrm{tr}}^{(v_k)}(X, Y)$, as in the sliced variant above. The maximum over these n_π values gives the scalar

$$\widehat{W}_{1,\max}^{(n_\pi)}(X, Y) \;=\; \max_{1 \leq k \leq n_\pi} \widehat{W}_{1,\mathrm{tr}}^{(v_k)}(X, Y). \quad (28)$$

Observe that since the trimmed distance is bounded and continuous (except on a measure-zero set), the sequence $\widehat{W}_{1,\max}^{(n_\pi)}(X, Y)$ converges almost surely to the essential supremum as $n_\pi \to \infty$ and, by extension, to (27).

F Optimizer Details

In the OT experiments, we apply the `AdamW` optimizer with cosine decay. We set the base learning rate to $\alpha_0 = 5 \times 10^{-6}$ and use the following schedule:

$$\alpha_t = \begin{cases} \alpha_0\, t/(\rho T), & t < \rho T, \\ \alpha_0 \tfrac{1}{2}\big[1 + \cos\big(\pi \tfrac{t-\rho T}{T-\rho T}\big)\big], & t \geq \rho T, \end{cases} \quad \rho = 0.1. \quad (29)$$

G Previous Works

Team	Easy	Medium	Hard
fosu-stu (Lv et al., 2024)	0.987	**0.887**	0.834
nycu-nlp (Lin et al., 2024)	0.964	0.857	**0.863**
no-999 (Ye et al., 2024)	**0.991**	0.830	0.832
text-understanding-and-analysis (Huang and Kong, 2024)	**0.991**	0.815	0.818

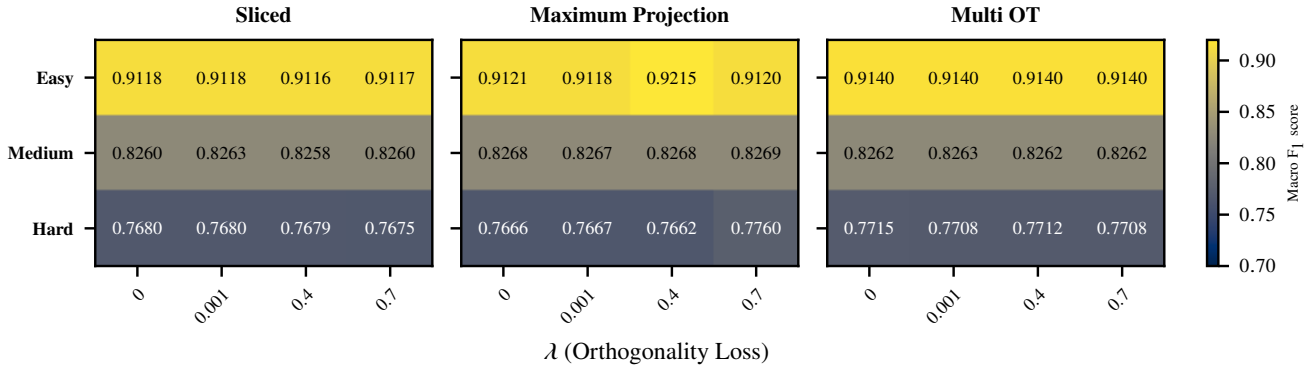
Table 4: F1 scores of different methods on the 2024 PAN Multi-Author Writing Style Analysis task (Zangerle et al., 2024).

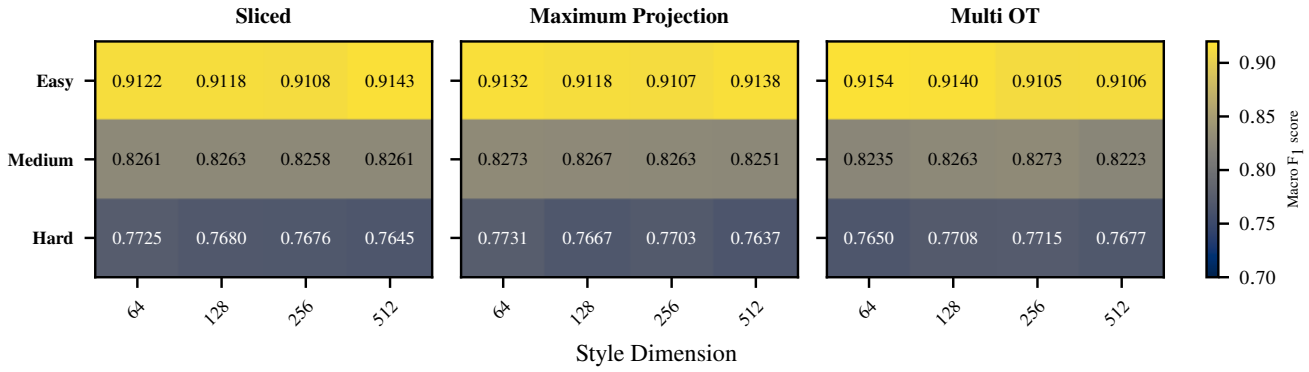
H StackExchange Results

Model	Macro F1
`CLS` + Orthogonality + Style Features	**0.4168**
Multi-OT (7 epochs, $n_\pi \in \{8, 32, 128\}$, no features)	0.4048
Unbalanced OT ($\tau = 0.8$, with features)	0.3947
Multi-OT (5 epochs, $n_\pi \in \{8, 32, 128\}$, no features)	0.3919
Max-proj OT ($n_\pi = 128$, no features)	0.3851
Sliced OT ($n_\pi = 64$, no features)	0.3814
Multi-OT (3 epochs, $n_\pi \in \{8, 32, 128\}$, no features)	0.3806
CL + OT ($\lambda = 0.1$, with features)	0.3769
Balanced OT (`blur` = 0.03, no features)	0.3645
Balanced OT (`blur` = 0.05, with features)	0.3643
2×Unbalanced OT + Multi-OT (5 epochs) + Multi-OT (7 epochs)	**0.4211**
Baseline Random	**0.4975**
Baseline 1	0.3530
Baseline 0	0.3124

Table 5: F1 scores obtained by selected models and all baseline methods on the StackExchange validation set from *dataset 3*, (Zangerle et al., 2022). n_π denotes the number of random directions. Recall that `blur` = $\sqrt{\varepsilon}$, where ε is the parameter in the entropy-regularized 1-Wasserstein cost. Intuitively, τ is the penalty for creating or destroying mass. In the OT variants, `style_dim` was set to 128.

I Full Ablation Results





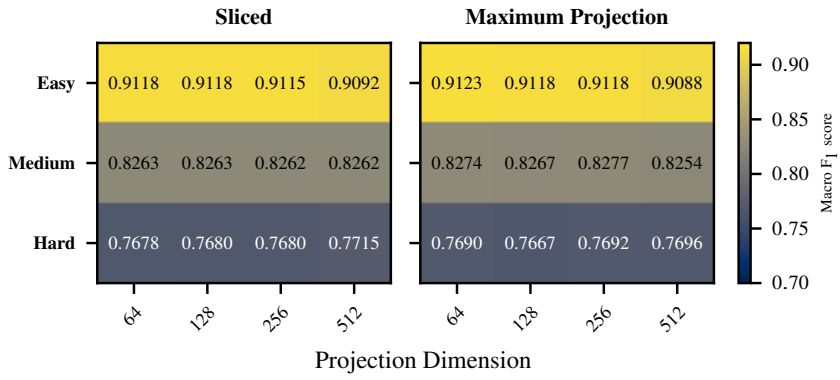


Figure 4: Heatmaps for the following hyperparameters: orthogonality loss, style dimension, and projection dimension. Note that the multi-scale OT approach already computes three projection dimensions (8, 32, 128), so there is no further ablation experiment for this method.

J Factorized Attention

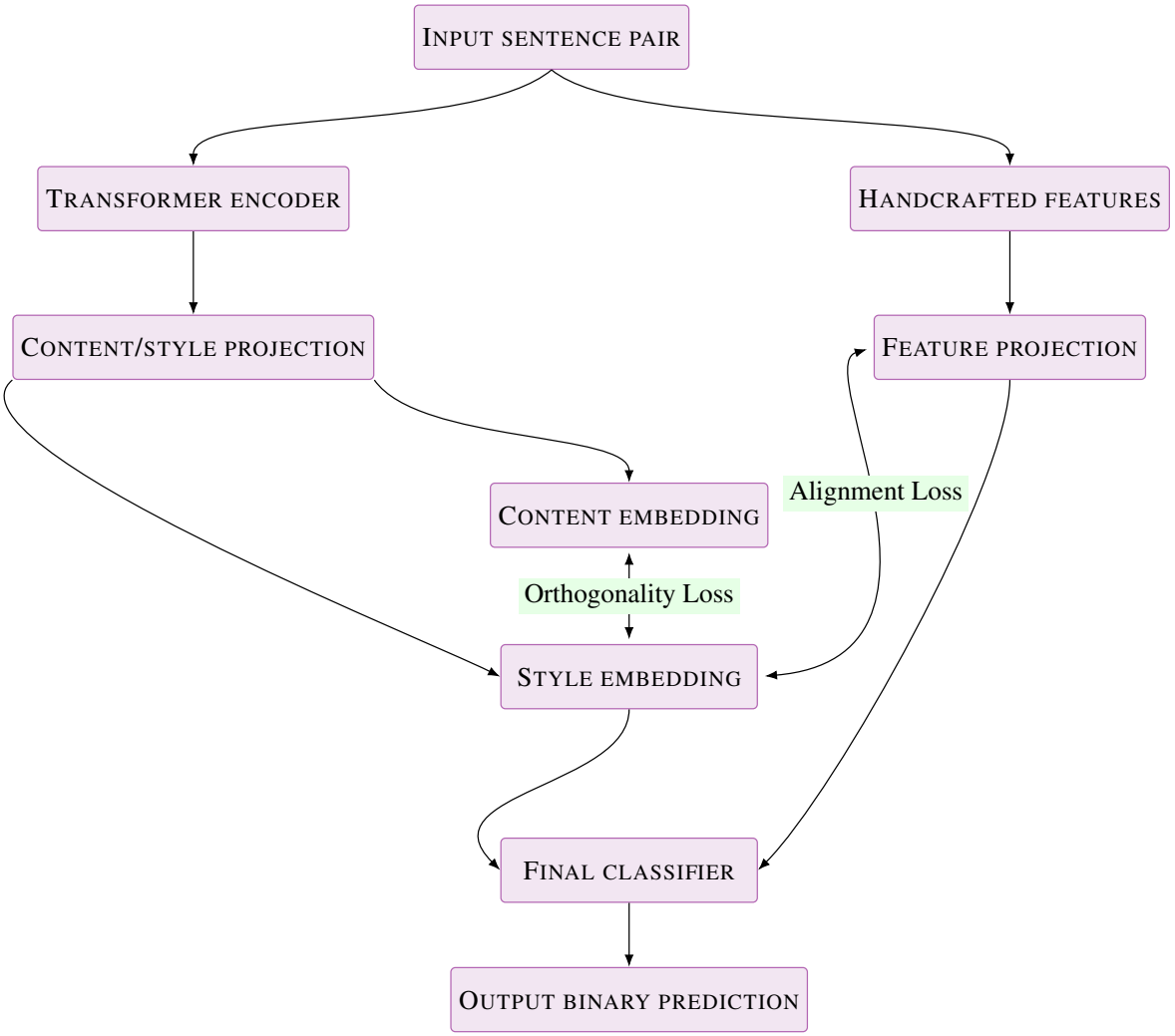


Figure 5: The structure of the Factorized Attention model.