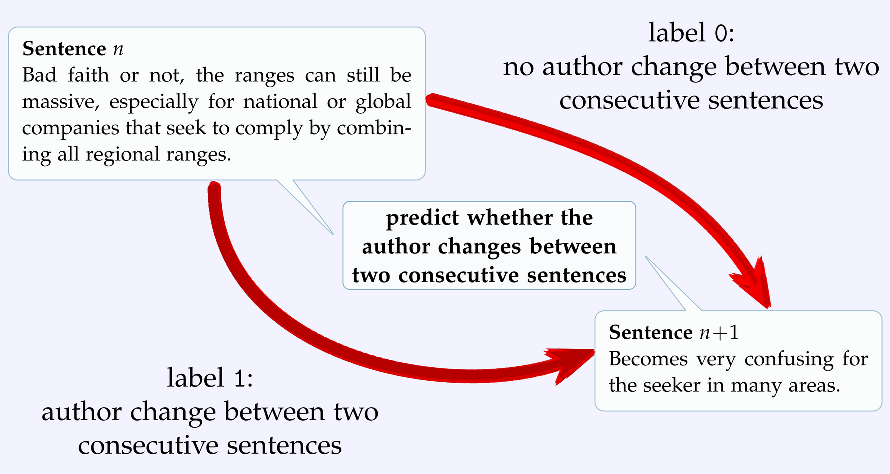
Hybrid Stylistic Shift Detection

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

Course: Computational Semantics for Natural Language Processing (July 11, 2025) ETH zürich

PAPER

Problem



Dataset

We are solving the 2025 PAN shared task on Multi-Author Writing Style Analysis. The dataset consists of Reddit comments. The training set has 158280 labels, and we test on 33654 labels from another set.

There are three difficulty levels:

- **Easy**: The sentences in the document cover various topics, which allows for the use of topic information to detect authorship changes.
- **Medium**: The topic changes in the document are subtle, which forces the method to focus more on style to effectively solve the detection task.
- **Hard**: All sentences in the document involve the same topic.

Results

Model Variant	Features	OT Variant	Macro F1 Easy	Macro F1 Medium	Macro F1 Hard	Macro F1 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
Factorized Attention Model	✓	N/A	0.9034	0.8208	**0.7904**	0.8382
Multi-OT (5 ep) + CL+OT + 2×UB + B (no feats)	✗/✓	mixed	0.9179	0.8286	0.7933	0.8467
Multi-OT (7 ep) + 2×CL+OT + 2×UB + B (no feats)	✗/✓	mixed	0.9186	0.8299	0.7942	0.8476
Multi-OT (5 ep) + 2×CL+OT + 2×UB + B (no feats)	✗/✓	mixed	**0.9186**	**0.8301**	**0.7943**	**0.8477**

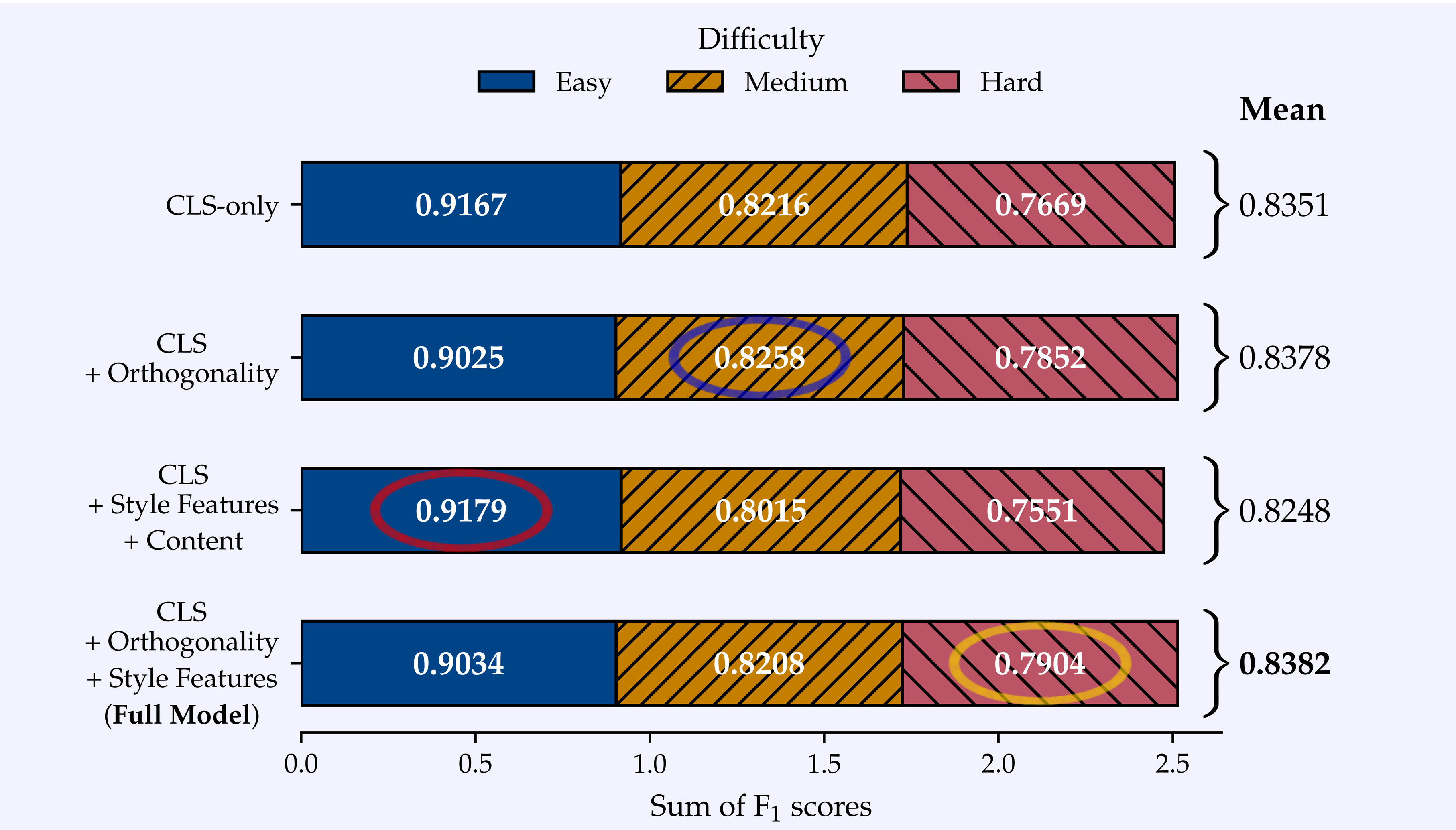
The table presents macro F1 scores and their arithmetic mean for selected configurations in each model family.

- Unless indicated otherwise, all models have been trained for 3 epochs.
- n_π denotes the number of random directions.
- `blur` $= \sqrt{\varepsilon}$, where ε is the parameter in the entropy-regularised 1-Wasserstein cost.
- `style_dim` corresponds to the style dimension embedding.
- Intuitively, τ penalises creating or destroying mass.
- Ensemble results: UB – Unbalanced Optimal Transport, B – Balanced OT, CL – Contrastive Learning, ep. – epoch.
- The multiplier indicates the voting weight in the ensemble.

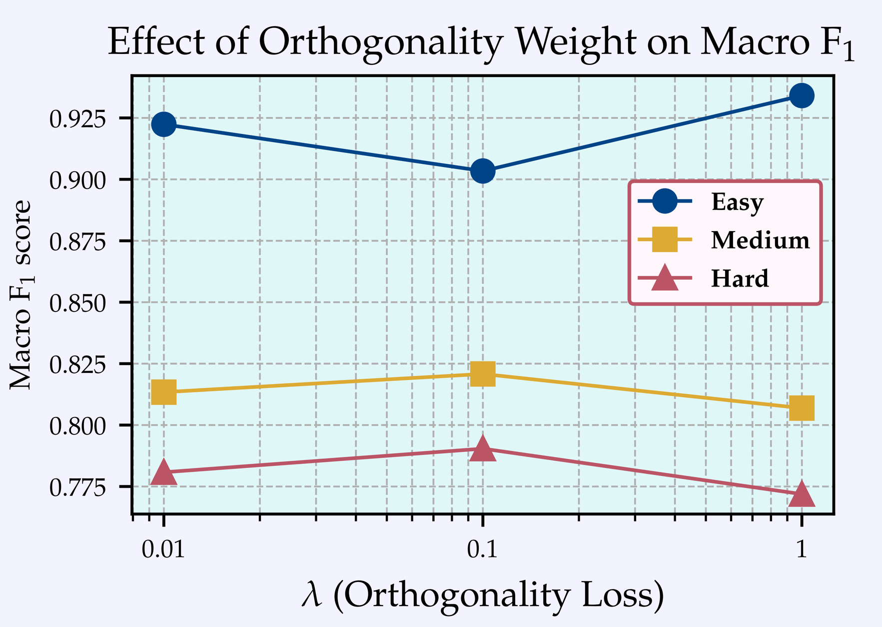
Key Findings

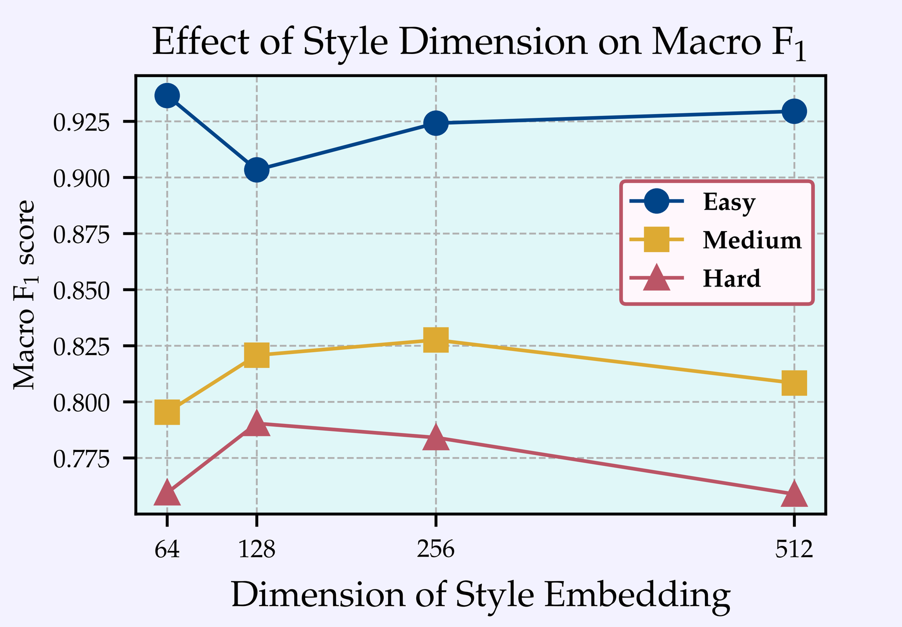
- OT-based ensemble models outperform single methods
- OT-based methods reach better average performance than the Factorized Attention model
- the Factorized Attention model reaches the best score on the Hard set
- **feature-free** OT-based methods are very close to the feature-informed approaches on the Hard set
- poor generalization on a dataset from a specialized domain (StackExchange) (best **ensemble** model: 2×Unbalanced OT + Multi-OT (5 epochs) + Multi-OT (7 epochs) achieved only **0.4211**)
- the methods outperform naïve baselines (≈ 0.45) but do not reach state-of-the-art results (0.863 macro F1 score) from the similar 2024 PAN shared task
- **future research directions:** computational diachronic semantics, extension to paragraph-based authorship detection, feature-free ensemble models

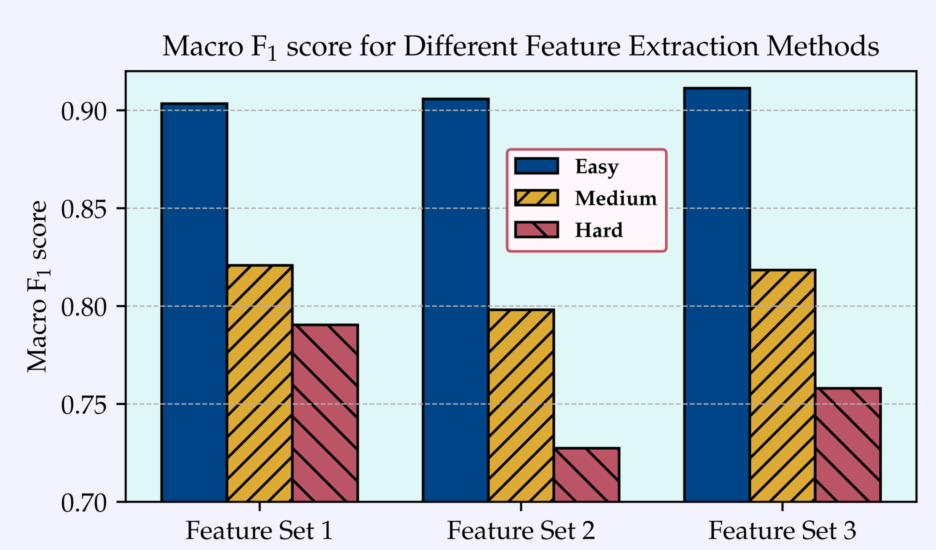
Ablation: Factorized Attention



- `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.





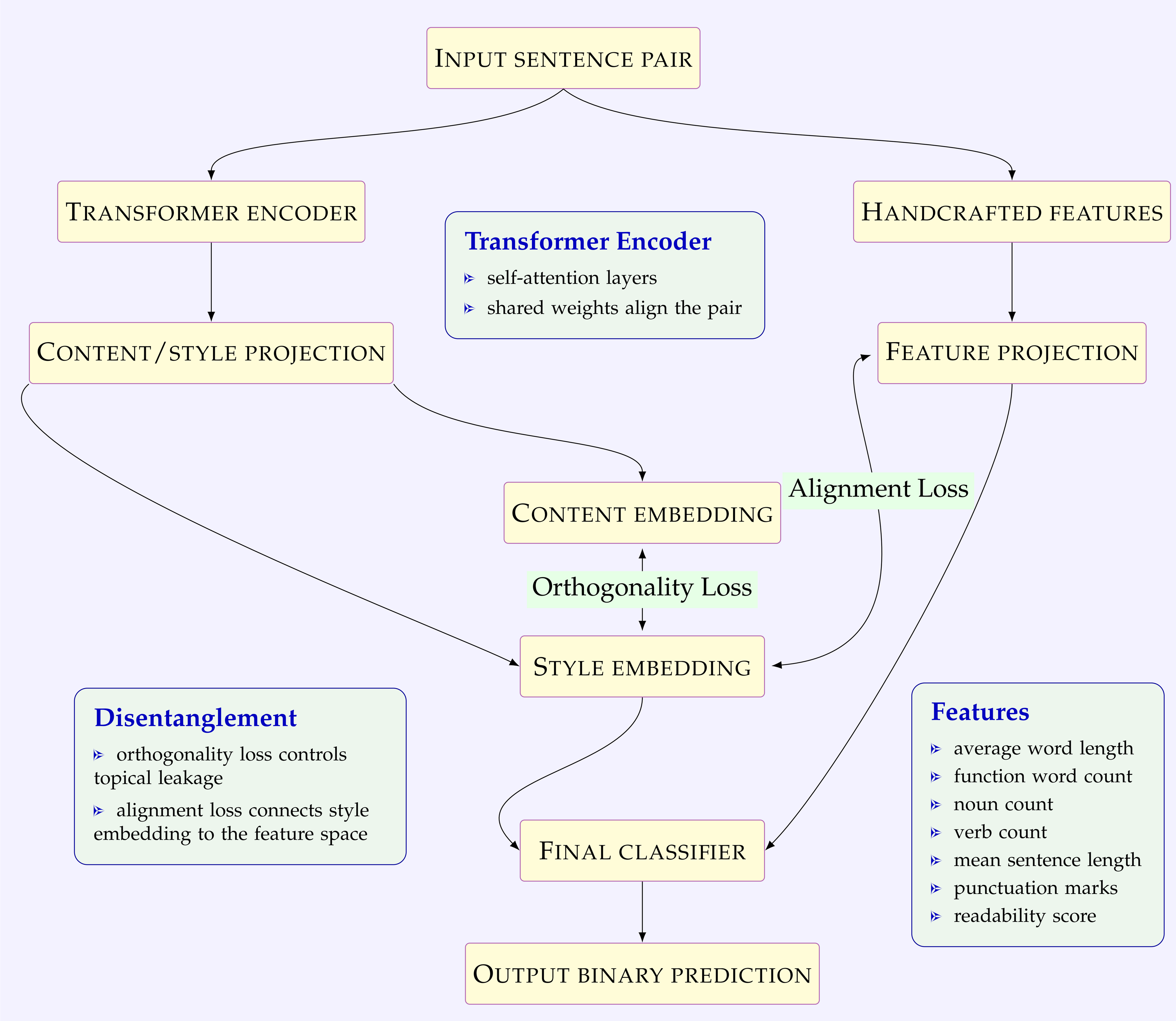


Feature Ablation

- *Feature Set 1* refers to the original feature extraction set.
- *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.

What do the results imply? 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.

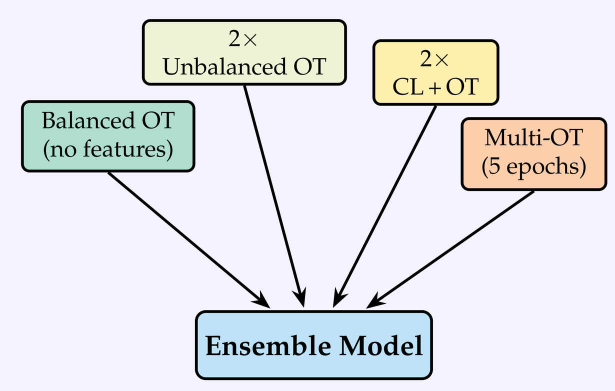
Solution: Factorized Attention

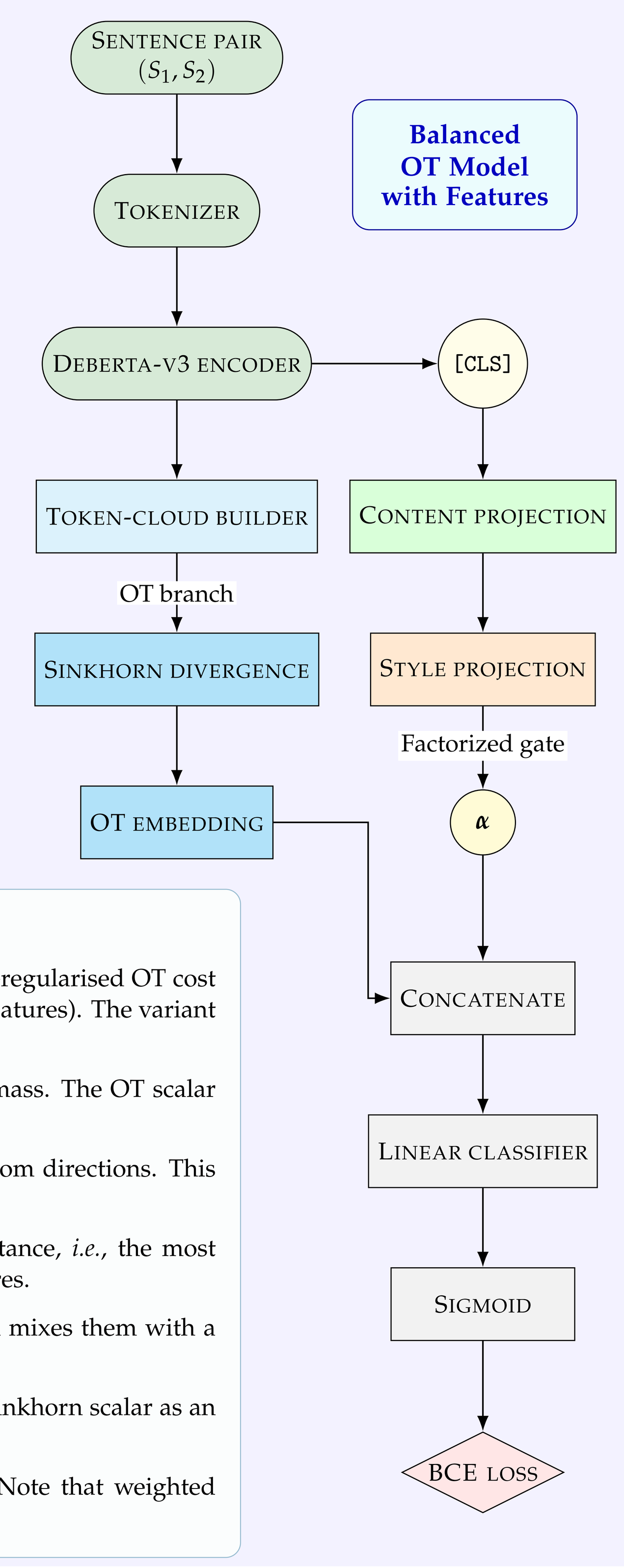


Solution: Optimal Transport

Basics

- **Optimal Transport (OT).** In the context of style shift detection, we treat each sentence as a **cloud** of token embeddings. OT measures the minimal "work" to transform one cloud into the other. This way, we obtain a permutation-invariant style distance.
- **Sinkhorn solver.** This algorithm provides an efficient approximation of the entropy-regularized OT cost by iterative matrix scaling. It makes OT computation feasible by reducing the complexity from $O(n^3)$ to $\approx O(k\,n^2)$, where k is the number of Sinkhorn iterations. Since the updates are basic matrix operations, there exist particularly **fast GPU** implementations.

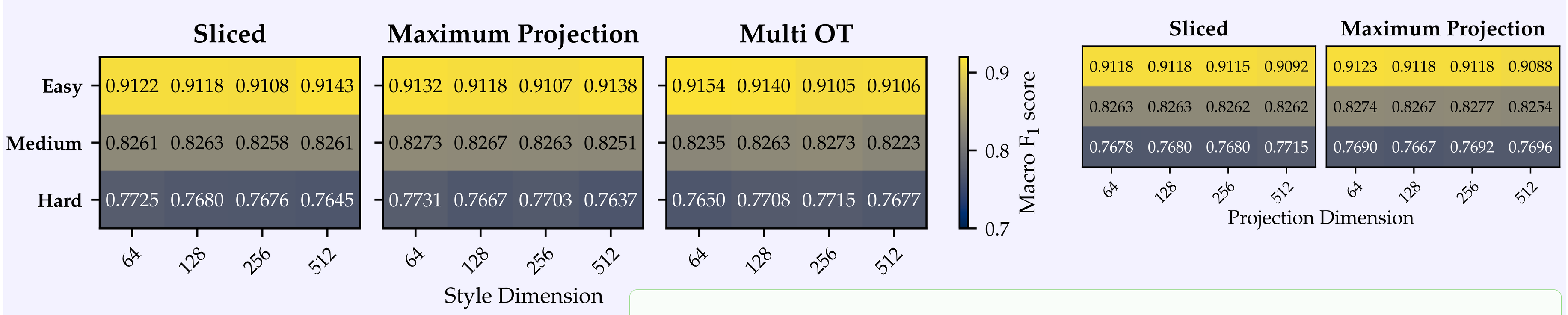


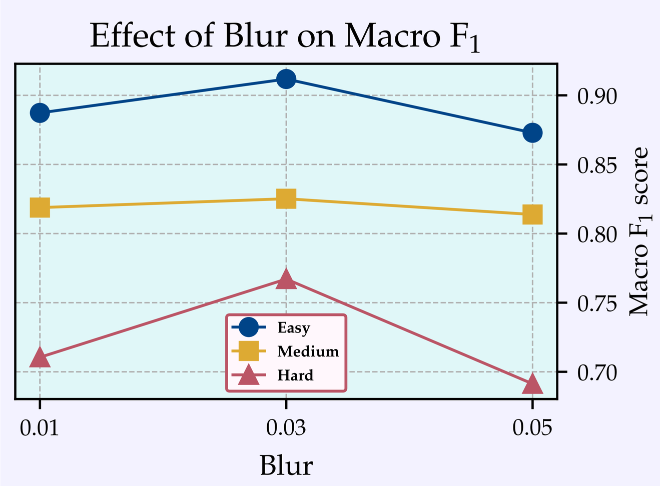


OT-based Models

- **Balanced (Sinkhorn) OT.** Computes the Sinkhorn divergence, a de-biased proxy for the ε-entropy-regularised OT cost on token clouds. The pipeline is: log-scaled OT scalar → `LayerNorm` → classifier (no handcrafted features). The variant with features is schematically described by the figure on the right.
- **Unbalanced (Sinkhorn) OT.** Adds KL penalties, which allows for the creation or destruction of mass. The OT scalar gets fused by a learnable gate with style features and content projection.
- **Sliced OT.** Approximates the 1-Wasserstein distance by averaging 1-D projections across n_π random directions. This variant does not rely on features.
- **Maximum-projection OT.** Same sampling as sliced OT but keeps the maximum directional distance, *i.e.*, the most "discriminative" slice. As a subvariant of the scaled approach, this method does not rely on features.
- **Multi-scale OT.** Concatenates sliced OT scalars at various resolutions ($n_\pi = \{8, 32, 128\}$), and then mixes them with a content-style gate. As a subvariant of the scaled approach, this method does not rely on features.
- **Contrastive Learning + OT.** Combines a contrastive loss on style embeddings with the balanced Sinkhorn scalar as an extra signal. This implementation relies on the balanced OT approach with features.
- **OT Ensembles.** Majority-vote mixes several OT variants to combine complementary signals. Note that weighted voting is possible. The ensembles contain both feature-informed and feature-free variants.

Ablation: Optimal Transport





`blur` tuning on the balanced OT model without features

- **Blur ($\sqrt{\varepsilon}$).** Best results are achieved with `blur` ≈ 0.003. Larger values lead to more bias, smaller values cause more variance, so F1 decreases in either direction.
- **Orthogonality weight.** Varying λ shifts macro F1 by $\approx< 0.003$. This implies that the OT scalar already explains the signal.
- **Style dimension.** Performance is effectively flat between 64 and 256 dimensions, which means that stylistic information is low-rank.
- **Projection dimension.** There is no gain beyond $n_\pi = 32$, and larger values simply lead to slower training.

Apart from `blur` tuning in the feature-free balanced OT method, OT variants are not particularly sensitive to orthogonality weight/projection dimension/style dimension tuning in the context of style shift detection.