OpenFact at PAN 2025: Punctuation-Guided Pretraining for Sentence-Level Style Change Detection

Notebook for the PAN Lab at CLEF 2025

Ewelina Księżniak[1,*], Krzysztof Węcel[1] and Marcin Sawiński[1]

[1]*Department of Information Systems, Poznań University of Economics and Business, Al. Niepodległości 10, 61-875 Poznań, Poland*

Abstract

This paper presents our approach to the PAN 2025 shared task on multi-author style change detection. The task involves identifying sentence-level boundaries where the writing style changes, presumably due to a switch in authorship. Motivated by the stylistic nature of the task, we propose a method based on intermediate-task learning. Specifically, we first perform contrastive pretraining of encoders using auxiliary tasks focused on detecting the presence of stylistic punctuation features—such as question marks and quotation marks—in order to enhance the encoder's sensitivity to fine-grained stylistic variation. These pretrained models are then fine-tuned on the main style change detection task. Additionally, we conducted error analysis and probing experiments to assess the stylistic awareness of the representations.

Keywords

PAN 2025, style change detection, intermediate-task learning

1. Introduction

Multi-author style change detection has been a task organized annually by the PAN lab since 2016. In the 2025 edition, the task focuses on detecting sentence-level style changes, requiring models to identify boundaries where the author switches [1], [2]. Given the stylistic nature of this task, we adopted an approach based on intermediate-task learning. Specifically, we pretrained encoders on contrastive auxiliary tasks designed to sensitize them to the presence or absence of selected punctuation marks—such as question marks or quotation marks—before fine-tuning on the main task. This strategy aims to improve the encoder's awareness of subtle stylistic features that are indicative of authorship changes, especially in contexts where topical signals are less informative.

2. Related Works

Intermediate-task learning is a transfer learning strategy in which a pretrained language model is further fine-tuned on an auxiliary task before being adapted to the target task. Two main paradigms are typically distinguished: *sequential learning*, where the model is trained on the intermediate task and then fine-tuned on the target task, and *multitask learning*, where both tasks are learned jointly, often with task-specific heads or shared representations. Sequential learning aims to transfer task-relevant skills in stages, while multitask learning promotes shared generalization across tasks [3, 4].

Contrastive learning is an approach for learning text representations by encouraging similar samples to be closer in embedding space while pushing dissimilar ones apart. Supervised contrastive learning

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

†These authors contributed equally.

ewelina.ksiezniak@ue.poznan.pl (E. Księżniak); krzysztof.wecel@ue.poznan.pl (K. Węcel); marcin.sawinski@ue.poznan.pl (M. Sawiński)

https://kie.ue.poznan.pl/en/ (E. Księżniak)

0000-0003-1953-8014 (E. Księżniak); 0000-0001-5641-3160 (K. Węcel); 0000-0002-1226-4850 (M. Sawiński)

(SCL) extends this idea by leveraging label information to define positive pairs. It encourages the model to group together representations of samples from the same class [5].

Probing is a widely used technique for analyzing the internal representations of pretrained language models, aiming to assess whether specific linguistic properties are encoded in their hidden states. The basic method involves training lightweight classifiers, known as probes, on top of frozen model layers to predict linguistic features such as part-of-speech tags, syntactic structures, or semantic roles [6]. To validate that the probe truly reflects model knowledge rather than overfitting to surface patterns, control tasks and diagnostic classifiers are often employed [7]. Probing has been extensively applied to examine a range of linguistic phenomena, including morphology, coreference, syntactic depth, and increasingly, stylistic aspects such as punctuation, capitalization, and sentence length.

3. Dataset Analysis and Preparation

To prepare the data for our experiments, we utilized the official datasets provided for the 2025 edition of the style change detection task. Each dataset contains comments sourced from the Reddit platform and corresponds to one of three difficulty levels—**Easy**, **Medium**, and **Hard**—each divided into training and validation subsets. In the Easy subset, documents span a wide range of topics, allowing models to leverage topical cues for authorship change detection. Medium subset includes documents with limited topical diversity, requiring greater reliance on stylistic features. In the Hard subset, all sentences share the same topic, making style the primary discriminative signal [1]. The task is formulated as a binary classification problem, where the label 1 denotes a sentence boundary at which the author changes, and 0 indicates no change in authorship. The label distributions for each subset are summarized in Table 1.

Table 1
Label distribution across training and validation datasets for each difficulty level.

Dataset	Label 0	Label 1
Hard Train	42,246	8,815
Hard Validation	8,742	1,906
Medium Train	46,314	12,503
Medium Validation	10,009	2,750
Easy Train	38,178	10,224
Easy Validation	8,046	2,201

Due to the significant class imbalance—where negative examples (label 0) greatly outnumber positive ones (label 1)—we applied random undersampling to the training datasets to create more balanced class distributions during model training. Additionally, to enable internal model evaluation during development, we randomly sampled 20% of each validation set to construct an internal test set.

To justify our approach, we investigated the extent to which authorship change correlates with shifts in specific punctuation patterns. For each pair of consecutive sentences, we examined whether both sentences contained any of the following features: ellipses (`...`), question marks (`?`), exclamation marks (`!`), quotation marks (`"`), or fully capitalized words (e.g., `IMPORTANT`). We applied the chi-squared test of independence to evaluate whether the distribution of stylistic punctuation changes is associated with authorship transitions. The null hypothesis (H_0) stated that the occurrence of a given punctuation feature change is independent of whether an authorship change occurred. The alternative hypothesis (H_1) posited a dependency between these variables [8].

The results showed statistically significant associations ($p < 0.05$) between authorship change and selectected stylistic features: for the *easy* dataset, significant dependencies were observed for *quotation*, *ellipses*, *question marks*, and *capitalized words*. In the *hard* and *medium* datasets, significant results were found for *quotation*, *ellipses*, *question*, *exclamation marks* and *capitalized words*.

4. Baseline Selection

To establish an internal baseline, we fine-tuned two pretrained language models—`xlm-roberta-base` and `mdeberta-v3-base`—independently for each difficulty level of the dataset. For each model, we experimented with three learning rates: $1e{-}5$, $2e{-}5$, and $3e{-}5$, and for each learning rate, we ran training with three different random seeds to ensure robustness. The data preparation involved concatenating each pair of consecutive sentences using a separator token to form the model input. All models were fine-tuned with a batch size of 16 for `xlm-roberta-base` and a batch size of 4 for `mdeberta-v3-base`, for up to 10,000 steps. We applied early stopping with a patience of 2, based on the F1 score on the validation set. Based on the results of our initial experiments, we selected the following configurations as baselines for each difficulty level: for both the **easy** and **medium** subsets, we used `xlm-roberta-base` with a learning rate of 2×10^{-5}; for the **hard** subset, we used the same model with a learning rate of 3×10^{-5}. All subsequent experiments reported in this paper were conducted using these configurations for the respective subsets.

5. Supervised Contrastive Pretraining

To sensitize the encoder to subtle stylistic cues, we explored pretraining using a supervised contrastive learning objective. First, we extracted sentences from the training subsets (separately for each difficulty level: Easy, Medium, and Hard) that contained at least one of the following stylistic markers: question marks (`?`), exclamation marks (`!`), quotation marks (`"`), ellipses (`...`), and capitalized words (entire tokens in uppercase). Table 2 summarizes the number of sentences per feature and difficulty level.

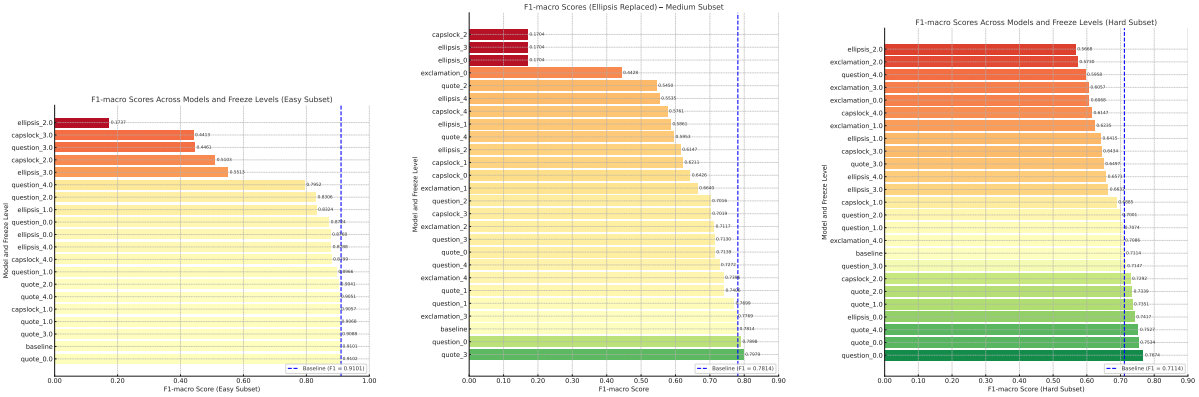
Table 2
Distribution of stylistic features across training datasets by difficulty level.

Level	CAPS	!	None	?	"	Ellipses
Easy	3,671	1,509	30,386	2,847	1,947	536
Medium	4,922	446	39,210	2,425	2,242	767
Hard	4,699	196	25,297	2,183	2,315	570

For each feature, we constructed a binary contrastive dataset consisting of 2,000 sentence pairs: 1,000 labeled as `0` (similar), where both sentences either contain or lack the target feature, and 1,000 labeled as `1` (dissimilar), where the feature appears in only one of the two sentences. These datasets were constructed independently for each difficulty level and each stylistic feature, resulting in 15 datasets in total (5 features $\times$ 3 levels). Each dataset was used to train a dedicated contrastive encoder based on `xlm-roberta-base`, resulting in 15 encoders—each specializing in one stylistic feature and difficulty level. The models were trained using *Supervised Contrastive Loss*. All encoders were trained for 10 epochs with a temperature scaling parameter of 0.2 and a batch size of 16.

6. Final Submission

Subsequently, we fine-tuned each of the contrastively pretrained encoders on the main downstream task of author style change detection. For each difficulty level (*easy*, *medium*, *hard*), we adopted the optimal hyperparameter configurations identified during the baseline model selection phase, including learning rate, batch size, and early stopping criteria. To further explore the generalizability and robustness of the stylistic representations learned during contrastive pretraining, we fine-tuned each encoder multiple times while freezing different numbers of its initial transformer layers: from zero (i.e., full fine-tuning) up to four frozen layers.



(a) Easy (b) Medium (c) Hard

Figure 1: Comparison of F1-macro scores across models and freeze levels for the **easy**, **medium**, and **hard** subsets. Model names follow the convention: *pretraining-task_number-of-frozen-layers*, where the encoder was pretrained on the specified task and fine-tuned to the main task with the given number of frozen layers.

7. Results on Internal Test Dataset

The charts illustrate the *F1-macro* scores achieved by both the baseline and various fine-tuned models, using contrastively pre-trained encoders with different numbers of frozen transformer layers, evaluated on an internal test set. The most substantial performance gain was observed in the **hard** subset, where the baseline *F1-macro* was 71.14%, and the best-performing model—fine-tuned from an encoder pretrained to discriminate between sentence pairs containing question marks, with no frozen layers—achieved 76.74%, indicating a ∼7% relative improvement. A moderate improvement was found for the **medium** subset, where the baseline scored 78.14%, and the top result (79.79%) was obtained by fine-tuning an encoder pretrained on quotation detection with three lower layers frozen. In contrast, the smallest improvement occurred in the **easy** subset, where the best model outperformed the baseline (91.01%) by only 0.01, using a quotation-sensitive encoder with no frozen layers.

This pattern is intuitive given the nature of the data: in the **easy** subset, which likely features more topic diversity and simpler sentence structures, semantic cues dominate over subtle stylistic signals. Meanwhile, the **hard** subset likely benefits more from encoders sensitive to fine-grained stylistic patterns. Nevertheless, despite these promising tendencies—particularly for the **hard** condition—the method exhibits instability. Multiple instances were observed where the same pretrained encoder, when fine-tuned with a different number of frozen layers, led to significant performance degradation, highlighting the need for careful hyperparameter control.

8. Final Submission

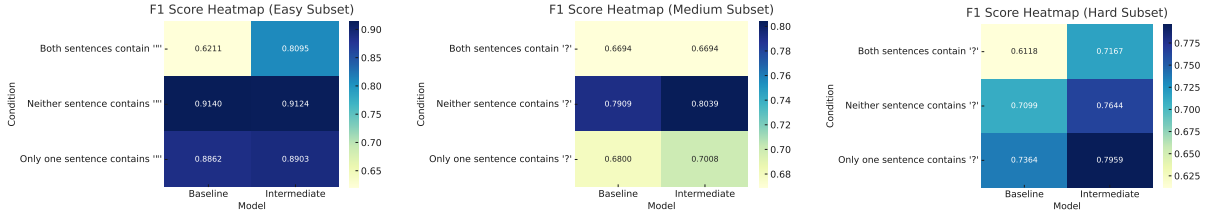
For the final submission, we selected three models based on their performance across the internal validation sets. For the easy subset, we submitted a model fine-tuned on a contrastively pretrained encoder sensitive to quotation marks, with no frozen layers. For the medium subset, we used a model pretrained on question mark discrimination with two lower layers frozen. Notably, although a model pretrained on quotation detection with three frozen layers achieved the best performance on our internal test set for the medium subset, we opted for the question-mark model due to technical issues encountered during final deployment. For the hard subset, we submitted a model fine-tuned on a question-mark-sensitive encoder without any frozen layers. The evaluation results on the official test set provided by the competition organizers are presented in Table 3.

Table 3
Final submission results on the official test set.

Submission ID	Easy (F1)	Medium (F1)	Hard (F1)
unsolvable-adventure	0.919	0.771	0.752

9. Errors Analysis and Model Probing

To assess whether intermediate pretraining improves model sensitivity to specific stylistic cues, we conducted an error analysis using our final submission models. For each subset (easy, medium, hard), the internal test set was partitioned into sentence pairs based on the presence of a target stylistic feature (e.g., question or quotation marks): both sentences containing the feature, neither containing it, or only one. As shown in the heatmaps, for the hard subset we observed improvements across all conditions—including pairs where *neither* sentence contained the relevant marker. This may indicate that the pretraining phase enhanced the encoder's ability to capture not only the targeted punctuation (e.g., question marks), but also broader stylistic or semantic traits. In the easy subset, the most notable gain (nearly 7 percentage points) occurred in pairs where *both* sentences contained quotation marks, while in the medium subset a moderate improvement (about 2 points) was observed specifically in cases where *only one* sentence included the target punctuation. These findings suggest that the benefits of stylistic pretraining may generalize beyond the pretraining signal itself, but manifest differently depending on task difficulty and input structure.



(a) Easy (b) Medium (c) Hard

Figure 2: Heatmaps showing F1-scores for sentence-pair conditions across the **easy**, **medium**, and **hard** subsets. Comparisons are made between models trained with and without intermediate learning.

Building on these findings, we carried out a complementary model probing experiment to further investigate whether the proposed approach enhances the model's sensitivity to specific stylistic features. The task was formulated as binary classification: determining whether a sentence contains a specific punctuation mark. For each stylistic marker, we trained separate logistic regression classifiers using `LogisticRegression(max_iter=1000, random_state=42)`. The results, expressed as F1-scores for the positive class (F1-pos), are presented in Table 4.

We compared three model configurations: (1) XML-RoBERTa base, representing sentence embeddings from the unmodified base model; (2) Target finetuning, where embeddings were taken from a model fine-tuned exclusively on the author style change task; and (3) Pretrained encoder finetuned, where the encoder was first subjected to intermediate contrastive pretraining on a stylistic signal and subsequently fine-tuned on the main task (i.e., the final submission models). For both the easy and hard subsets, the intermediate learning approach led to a notable increase in embedding-level stylistic signal, outperforming standard finetuning not only for the pretraining-related feature, but also for other stylistic tasks. Surprisingly, this effect was not replicated in the medium subset. Here, target finetuning embeddings generally yielded higher or equivalent stylistic separability compared to the intermediate learning setup.

Table 4
F1-pos scores for XML-RoBERTa base, target-task finetuning, and finetuning on pretrained encoder across tasks and difficulty subsets.

Subset	Task	XML-RoBERTa Base	Target Finetuning	Pretrained Encoder
easy	question	0.9352	0.8714	**0.9401**
easy	quote	0.8198	0.8733	**0.8935**
easy	exclamation	0.9532	0.9406	**0.9730**
easy	capslock	0.6261	0.6885	0.3036
easy	ellipsis	0.3284	0.4545	**0.5287**
medium	question	0.9269	0.9641	0.9501
medium	quote	0.8344	0.8452	**0.9599**
medium	exclamation	0.7286	0.8805	0.8481
medium	capslock	0.6748	0.6930	0.3585
medium	ellipsis	0.2793	0.6770	0.6185
hard	question	0.9131	0.9781	0.9437
hard	quote	0.8939	0.8313	**0.9395**
hard	exclamation	0.3750	0.6552	**0.8406**
hard	capslock	0.7144	0.6816	0.4254
hard	ellipsis	0.1575	0.3066	**0.5934**

10. Conclusion and Future Work

Our experiments demonstrate that contrastive intermediate-task pretraining focused on stylistic punctuation features can enhance encoder sensitivity to fine-grained authorial variation, particularly in more challenging subsets of the style change detection task. While the improvements over the baseline are promising, especially for the hard dataset, our results also reveal performance instability across fine-tuning configurations—particularly with varying numbers of frozen layers—which suggests that encoder robustness remains a concern. This variability indicates a need for more systematic regularization or architectural calibration. As a potential direction for future work, we plan to explore ensemble methods that combine multiple stylistically specialized encoders. Such ensembles could help mitigate individual encoder fluctuations while leveraging complementary stylistic representations to further improve detection accuracy.

11. Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT to check grammar, spelling, and style. The tool was applied to selected paragraphs, and all corrections were manually reviewed and approved.

References

[1] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Generative AI Authorship Verification, Multi-Author Writing Style Analysis, Multilingual Text Detoxification, and Generative Plagiarism Detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[2] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[3] J. Phang, T. Févry, S. R. Bowman, Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks, in: arXiv preprint arXiv:1811.01088, 2018.

[4] Y. Pruksachatkun, J. Phang, H. Liu, P. M. Htut, X. Zhang, R. Y. Pang, C. Vania, R. T. McCoy, S. R. Bowman, Intermediate-task transfer learning with pretrained language models: When and why does it work?, in: Proceedings of ACL, 2020.
[5] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems (NeurIPS), 2020.
[6] A. Conneau, D. Kiela, What you can cram into a single vector: Probing sentence embeddings for linguistic properties, in: Proceedings of ACL, 2018.
[7] J. Hewitt, P. Liang, Designing and interpreting probes with control tasks, in: Proceedings of EMNLP, 2019.
[8] R. J. Tallarida, R. B. Murray, R. J. Tallarida, R. B. Murray, Chi-square test, Manual of pharmacologic calculations: with computer programs (1987) 140–142.