Team wqd at Style Change Detection in Multi-Author Writing: A Deep Learning Approach Based on DeBERTa

Notebook for PAN Lab at CLEF 2025

Xiaocan Lin, Chang Liu, Xianbing Duan and Zhongyuan Han*

Foshan University, Foshan, China

Abstract

Style change detection in multi-author writing constitutes a significant research challenge in computational linguistics, with important applications in academic integrity maintenance, forensic investigation, and intelligent writing assistance. This paper proposes a novel DeBERTa-based deep learning approach for sentence-level style change detection. Through systematic comparison with mainstream pre-trained models including RoBERTa across datasets of varying difficulty levels, we conduct comprehensive training and evaluation on three multi-author writing style analysis datasets (Easy, Medium, and Hard) from PAN. As team **wqd**, our proposed method achieves F1-scores of **0.958(ranking 2nd)**, **0.823(ranking 2nd)**, and **0.830(ranking 1st)** on the respective test sets, demonstrating both effectiveness and robustness.

Keywords

PAN 2025, Multi-author writing, Style Change Detection

1. Introduction

With the exponential growth of digital text data, the analysis and processing of multi-author writing have become increasingly important [1][2]. As a key branch of author identification, style change detection focuses on identifying the textual positions where authorial identity shifts within a document. Early studies primarily relied on statistical features (such as word frequency and syntactic features) and clustering algorithms for style change detection[3][4]. Unlike traditional author attribution, style change detection has achieved higher F1 scores and demonstrated broader applicability by combining BERT models with random forest classifiers. Additionally, some studies have proposed ensemble methods that integrate multiple models to improve detection performance[5][6]. In recent years, the PAN series of evaluation tasks have significantly promoted the development of this field, shifting the research focus from paragraph-level analysis to more fine-grained sentence-level studies.

This study presents a systematic experimental investigation into the task of style change detection in multi-author texts. Specifically, we conduct a comprehensive comparative analysis of various pre-trained language models, including advanced architectures such as DeBERTa (e.g., microsoft/deberta-base) and RoBERTa, in combination with multiple data augmentation strategies to evaluate their performance on the task. To further enhance model performance, we incorporate the novel multilingual embedding model BGE-M3 and implement a specialized feature fusion strategy for cross-lingual feature integration, thereby thoroughly exploring effective approaches to improve style change detection.

For evaluation, we employ F1-score as the primary metric and perform multi-dimensional quantitative analyses across three critical subtasks (corresponding to Easy, Medium, and Hard difficulty levels in the dataset). Our evaluation encompasses various aspects including different model configurations and data processing methods. Through empirical validation, we systematically assess the contribution of each technical component (e.g., base models, data augmentation techniques, and feature fusion methods) to

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

*Corresponding author.

linxx6513@gmail.com (X. Lin); lc965024004@gmail.com (C. Liu); 15007473346@163.com (X. Duan); hanzhongyuan@gmail.com (Z. Han*)

0009-0000-0887-9273 (C. Liu); 0000-0001-8960-9872 (Z. Han*)

the overall task performance. The ultimate objective is to identify the optimal model configuration that achieves an optimal balance between computational efficiency and detection accuracy.

2. Method

2.1. Dataset Construction

The experimental dataset consists of text files (`problem-*.txt`) and corresponding annotation files (`truth-problem-*.json`). The text files contain continuous text for detection, while the annotation files use the `changes` field to mark style variations between consecutive sentences (1 indicates variation, 0 indicates no variation).The dataset processing pipeline comprises:

- **Sentence Segmentation**: Utilizing NLTK's `sent_tokenize` tool to split text into sentence sequences;
- **Sample Construction**: Forming (`sent1`, `sent2`, `label`) samples where consecutive sentence pairs serve as input and corresponding `changes[i]` as labels;
- **Anomaly Handling**: Skipping documents where sentence count and annotation count mismatch ($\text{len(sentences)} \neq \text{len(changes)} + 1$) to avoid data noise.

2.2. Model Architecture

The experiment employs a classification framework based on pretrained language models, with the following core structure:

- **Base Models**: Comparative evaluation is conducted on DeBERTa (`microsoft/deberta-base`) and RoBERTa[7][8]. Their semantic understanding capabilities for sentence pairs are leveraged.
- **Classifier**: A two-layer linear network (incorporating ReLU activation and Dropout regularization) processes the hidden state of the `[CLS]` token (for DeBERTa) or pooled features. It outputs binary classification results for style variation detection.
- **Enhanced Model**: For the RoBERTa+BGE-M3 combination, a BGE feature fusion strategy is introduced. Dimensionality reduction is performed via a linear projection layer. Then, the dense features of RoBERTa and BGE are concatenated before being fed into the classifier.

2.3. Training and Evaluation Protocol

- **Training Parameters**: Batch size 16, learning rate 1×10^{-5}, AdamW optimizer, BCEWithLogitsLoss, trained for 5 epochs;
- **Evaluation Metrics**: F1-score as primary metric, with best validation-performing models (highest F1) retained for analysis;
- **Controlled Variables**: Model type (DeBERTa/RoBERTa), BGE-M3 integration, and data augmentation (reversal + transition-focused truncation strategy).

3. Experiments

3.1. Dataset

The experimental datasets comprise three difficulty levels: Easy, Medium, and Hard[9]. The Easy dataset includes documents with diverse themes, the Medium dataset has limited thematic variations, and the Hard dataset is strictly confined to a single theme. This setup systematically evaluates model performance across different scenarios.

3.2. Data Processing

The PAN25 Writing Style Analysis Evaluation Task focuses on sentence-level style change detection in multi-author documents. It requires participating systems to accurately identify all writing style transition boundaries under the conditions of strictly controlling author identity and topic variations. Firstly, the text is automatically split into sentences using the `nltk.sent_tokenize` tool, generating an ordered set of sentences $P = \{p_1, p_2, \ldots, p_n\}$. Subsequently, a series of adjacent sentence pairs $S = \{(p_1, p_2), (p_2, p_3), \ldots, (p_{n-1}, p_n)\}$ are constructed as the basic analysis units[10], where each sentence pair represents a potential style transition point,as illustrated in Table 1.

Table 1
Sample Sentence Pairs from Training Data

sentence_1	sentence_2	label
I learned this about Ukraine a while back and....	It's easy for some to make the mistake as back....	0
It's easy for some to make the mistake as back....	But when ppl see others say it the respectfully....	0

Newline characters in the text are replaced with spaces to avoid sentence segmentation errors. The sentence pairs are encoded using the Tokenizer corresponding to the pretrained model, with padding set to `padding='max_length'` and attention masks generated simultaneously. To address the maximum sequence length constraint of Transformer models, truncation is applied with `max_length=128`. The statistical overview of the sample collection is presented in Table 2.

Table 2
Dataset Statistics

Datasets	Easy		Medium		Hard	
	#documents	#para.	#documents	#para.	#documents	#para.
Training set	4200	46713	4200	54452	4200	48372
Validation set	900	9921	900	11771	900	10283

3.3. Additional Processing for Comparative Experiments

- **Data Augmentation**: The training data is augmented using the "Inversion + Transition Focus Truncation Strategy" to enhance model robustness. The inversion involves adding $(sent2, sent1, label)$ pairs based on the original $(sent1, sent2, label)$ sentence pairs. The attention truncation strategy is implemented by concatenating the last 64 characters of *sentence_1* with the first 64 characters of *sentence_2* before token embedding, followed by encoder processing. This method further improves the model's ability to handle long texts and capture style transition features, thereby enhancing the accuracy and efficiency of style variation detection.
- **Feature Fusion Methodology**: The sentence-level dense embeddings generated by the BGE-M3 model are concatenated with the contextual representations extracted from RoBERTa, forming a hybrid feature vector that serves as input to the classification layer.

3.4. Experiment Results

We carried out four experiments: Using BERT - series pre - trained models, namely DeBERTa - base and RoBERTa, combined with data augmentation and BGE - M3 feature fusion, to train on the training set. The model with the highest accuracy was selected from the validation set to fine - tune the hyperparameters. Subsequently, we deployed the model with the highest F1 score on the test sets corresponding to different difficulty levels submitted by the TIRA platform, and simultaneously compared it with the baseline predictions provided by the competition. The main experimental results are shown in Table 3 and Table 4.

Table 3
F1-scores for author change detection in the multi-author writing style task (test set results)

Method	Task 1	Task 2	Task 3
wqd_felt-bronze	0.958	0.823	0.830
Baseline Predictor	0.439	0.440	0.453

Table 4
Comparison of F1-scores across different model configurations (validation set results)

Configuration	Task1	Task2	Task3
DeBERTa	0.9558	0.8414	0.8331
DeBERTa+DA	0.9131	0.8231	0.8228
RoBERTa	0.9517	0.8352	0.8227
RoBERTa+BGE-M3	0.9390	0.8151	0.8069

4. Conclusions

Through comparative analysis of different models, code versions, and data augmentation strategies for style variation detection, this study yields the following key findings[11]:

- **Model Selection**: DeBERTa demonstrates superior performance over RoBERTa with fewer training epochs, establishing itself as the preferred choice for this task.
- **Data Augmentation**: The implemented "Inversion + Transition Focus Truncation Strategy" failed to improve model performance, potentially due to excessive augmented data volume. Future work should explore reduced augmentation quantities or alternative augmentation approaches.
- **Future Directions**: Promising avenues include optimization of data volume, adjustment of learning rate scheduling strategies, and exploration of more effective feature fusion methods to enhance model generalization capabilities.

Acknowledgments

This work is supported by the National Social Science Foundation of China (24BYY080).

Declaration on Generative AI

During the preparation of this work, the authors utilized the DeepSeek language model for grammar and spelling checks. Following the use of this tool, the authors carefully reviewed and edited the content as necessary and assume full responsibility for the content of this publication.

References

[1] J. Bevendorff, D. Dementieva, M. Fröbe, 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] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[3] M. T. Zamir, M. A. Ayub, A. Gul, N. Ahmad, K. Ahmad, Stylometry analysis of multi-authored documents for authorship and author style change detection, arXiv preprint arXiv:2401.06752 (2024).

[4] R. Singh, J. Weerasinghe, R. Greenstadt, Writing style change detection on multi-author documents., in: CLEF (Working Notes), 2021, pp. 2137–2145.

[5] M. Huang, Z. Huang, L. Kong, Encoded classifier using knowledge distillation for multi-author writing style analysis., in: CLEF (Working Notes), 2023, pp. 2629–2634.

[6] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble pre-trained transformer models for writing style change detection., in: CLEF (Working Notes), 2022, pp. 2565–2573.

[7] A. Karanikola, G. Davrazos, C. M. Liapis, S. Kotsiantis, Financial sentiment analysis: Classic methods vs. deep learning models, Intelligent Decision Technologies 17 (2023) 893–915.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[9] 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.

[10] J. Lv, Y. Yi, H. Qi, Team fosu-stu at pan: Supervised fine-tuning of large language models for multi author writing style analysis, Working Notes of CLEF (2024).

[11] G. Ríos-Toledo, J. P. F. Posadas-Durán, G. Sidorov, N. A. Castro-Sánchez, Detection of changes in literary writing style using n-grams as style markers and supervised machine learning, Plos one 17 (2022) e0267590.