Llama-3 with 4-bit Quantization and IA3 Tuning for Multi-Author Writing Style Analysis

Notebook for PAN at CLEF 2025

Dongjie Chen, Jijie Li and Haoliang Qi*

[1]*Foshan University, Foshan, Guangdong, China*

Abstract

The Multi-Author Writing Style Analysis task aims to detect authorship changes within documents, critical for plagiarism detection and authorship verification. This paper introduces a novel approach combining Llama-3-8B, 4-bit quantization, and IA3 fine-tuning to address this challenge. Our method efficiently adapts large language models to style change detection while minimizing computational costs. Evaluated on the PAN 2025 dataset (Easy/Medium/Hard tasks), our approach achieves F1 scores of 0.461 (Easy), 0.583 (Medium), and 0.484 (Hard), outperforming baselines by +5.0%, +32.5%, and +6.8%, respectively. The results demonstrate IA3's effectiveness in capturing stylistic features, especially under limited topical diversity.

Keywords

Writing Style Analysis, IA3 Tuning, 4-bit Quantization, Llama-3

1. Introduction

Multi-author writing style analysis (MAWSA) remains a pivotal task in natural language processing, focusing on identifying authorial transitions in documents through stylistic variations [1, 2]. This task is crucial for applications like plagiarism detection and authorship verification, but its complexity has increased with diverse text genres and the need to distinguish subtle style differences—especially in topic-consistent contexts where author shifts are harder to detect [2].

Traditional MAWSA methods relying on lexical or syntactic features often fail to capture deep contextual dependencies, while pre-trained transformer models like DeBERTa-v3 have shown promise in integrating contextual understanding for stylistic analysis. However, deploying these models faces challenges in computational efficiency and parameter optimization, particularly across datasets of varying difficulty (easy, medium, hard) [3].

Recent parameter-efficient fine-tuning (PEFT) advances, such as Low-Rank Adaptation (LoRA) [4], have enabled scalable task adaptation. For MAWSA, which requires balancing global context and local stylistic cues, more sophisticated PEFT techniques are needed to optimize model plasticity and stability. This study introduces a framework combining Meta-Llama-3-8B [5] with 4-bit quantization and Improved Attention with Adaptive Instances (IA3) [6] fine-tuning. By adapting attention modules—specifically the *query projection* (`q_proj`) and *value projection* (`v_proj`) layers that transform inputs into query and value vectors in the attention mechanism—IA3 prioritizes stylistic features, while quantization reduces computational overhead. Experiments on the PAN MAWSA dataset show our model achieves F1 scores of 0.461, 0.583, and 0.484 for Tasks 1–3, outperforming baselines (0.439, 0.44, 0.453) and highlighting its efficiency in medium-difficulty stylistic detection.

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

† corresponding author

hellojie1449210489@gmail.com (D. Chen); 2196629893@qq.com (J. Li); haoliang.qi@gmail.com (H. Qi)

0009-0007-1330-8310 (D. Chen); 0009-0009-5566-4166 (J. Li); 0000-0003-1321-5820 (H. Qi)

2. Task and Datasets

2.1. Task Overview

The Multi-Author Writing Style Analysis task in PAN 2025 aims to identify sentence-level authorial changes within multi-author documents. Specifically, for each pair of consecutive sentences, the task requires determining whether a writing style change has occurred. The challenge is designed to evaluate models' ability to distinguish stylistic variations while controlling for topic shifts, with three difficulty levels:

- **Easy**: The sentences of a document cover a variety of topics, allowing approaches to make use of topic information to detect authorship changes.
- **Medium**: The topical variety in a document is small (though still present) forcing the approaches to focus more on style to effectively solve the detection task.
- **Hard**: All sentences in a document are on the same topic.

2.2. Datasets

The datasets are derived from user posts on Reddit, combined into documents with controlled authorial and topic changes. Each dataset is split into training (70%), validation (15%), and test (15%) sets, and provided in English. Key characteristics include:

2.2.1. Data Structure

For each problem instance (document), two files are provided:

- **problem-X.txt**: The text document, formatted as sentences.
- **truth-problem-X.json**: Ground truth in JSON format, containing:
  - `authors`: The number of authors.
  - `changes`: A binary array where each element indicates whether a style change occurs between consecutive sentences (1 for change, 0 for no change).

2.2.2. Data Preprocessing

The input text is preprocessed to remove redundant empty lines and special characters, ensuring consistency. Adjacent sentences are paired to form input samples for model training, with each pair labeled as a style change (1) or no change (0). For sequences exceeding 512 tokens [7], truncation is applied to fit model input constraints.

2.3. Evaluation Metrics

Submissions are evaluated using the macro F1-score, which balances precision and recall across all sentence pairs. The metric is computed independently for each difficulty level (Easy, Medium, Hard) to assess model performance under varying conditions. A provided script facilitates evaluation based on the output JSON files, which must follow the format of the ground truth (i.e., a changes array of binary values for each sentence pair).

3. Methods

Our approach processes sentence pairs through four key stages: (1) Input tokenization and embedding, (2) Quantized transformer processing with IA3-adapted attention, (3) Style feature extraction via modified feedforward networks, and (4) Binary classification. The system first tokenizes sentence pairs with [SEP] markers, then processes them through Llama-3's 4-bit quantized layers where IA3 scaling vectors adapt query/value projections to emphasize stylistic features. Final hidden states are classified using a linear layer trained with cross-entropy loss.

3.1. Task Formulation

We frame the Multi-Author Writing Style Analysis as a binary classification task. Given a document $D = \{p_1, p_2, ..., p_n\}$ segmented into sentences, we construct adjacent sentence pairs (p_i, p_{i+1}). The model predicts a binary label $y \in \{0, 1\}$, where:

- $y = 0$: Consecutive sentences share the same author
- $y = 1$: Author change occurs between sentences

This formulation transforms the style change detection into a sequence classification problem at the sentence-pair level.

3.2. Model Architecture

Our architecture integrates Meta-Llama-3-8B with 4-bit quantization and IA3 tuning. The computation flow for a sentence pair (p_i, p_{i+1}) is defined as:

3.2.1. Input Representation

Given a sentence pair (p_i, p_{i+1}), we concatenate them with a separator token and encode using Llama-3's tokenizer:

$$\mathbf{x}_i = \text{Tokenizer}(p_i \parallel \text{[SEP]} \parallel p_{i+1}, \text{max\_length} = 512, \text{truncation} = \text{True}) \tag{1}$$

where $\parallel$ denotes concatenation and [SEP] is the separation token. The tokenized output includes:

$$\mathbf{x}_i = \{\text{input\_ids}, \text{attention\_mask}\} \in \mathbb{R}^{512} \tag{2}$$

where input_ids are token indices and attention_mask indicates non-padding tokens.

3.2.2. Embedding Layer

The tokenized input $\mathbf{x}_i$ is mapped to dense vector representations through an embedding layer:

$$\mathbf{E}_i = \text{EmbeddingLayer}(\mathbf{x}_i) \tag{3}$$

where $\mathbf{E}_i \in \mathbb{R}^{512 \times d}$ is the embedding matrix, $d = 4096$ is the hidden dimension size, and 512 is the maximum sequence length. This transforms discrete tokens into continuous vectors suitable for transformer processing.

3.2.3. Quantized Transformer Processing

The embeddings are processed through 32 transformer layers with 4-bit quantized weights:

$$\mathbf{H}^{(0)} = \mathbf{E}_i \tag{4}$$

For each layer $l \in [1, 32]$:

$$\mathbf{H}^{(l)} = \text{TransformerLayer}^{\text{quant}}\left(\mathbf{H}^{(l-1)}\right) \tag{5}$$

where weights are quantized using NF4 with double quantization [8]:

$$\mathbf{W}^{\text{quant}} = Q(\mathbf{W}), \quad Q_{\text{NF4}} = \text{BlockwiseQuant}(block\_size = 64) \tag{6}$$

Here Q denotes the quantization function, reducing memory footprint by 68% while preserving model capacity.

3.2.4. IA3 Attention Modification

At each attention layer, IA3 injects trainable scaling vectors (λ) to adapt query and value projections:

$$\mathbf{Q} = (\mathbf{W}_q \odot (\mathbf{1} + \lambda_q))\, \mathbf{H}^{(l-1)} \tag{7}$$

$$\mathbf{V} = (\mathbf{W}_v \odot (\mathbf{1} + \lambda_v))\, \mathbf{H}^{(l-1)} \tag{8}$$

$$\text{Attention} = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{9}$$

where $\mathbf{W}_q$ and $\mathbf{W}_v$ are the original query and value projection matrices, $\odot$ denotes element-wise multiplication, $\lambda_q, \lambda_v \in \mathbb{R}^d$ are task-specific learnable vectors that scale the projections. This adaptation allows the model to dynamically adjust attention patterns for style analysis while keeping most parameters frozen.

3.2.5. Feedforward Network Adaptation

The feedforward network is similarly adapted using scaling vectors:

$$\text{FFN}(\mathbf{x}) = (\mathbf{W}_{\text{down}} \odot (\mathbf{1} + \lambda_d))\, \sigma(\mathbf{W}_{\text{up}}\mathbf{x}) \tag{10}$$

where $\mathbf{W}_{\text{down}}$ and $\mathbf{W}_{\text{up}}$ are the original down-projection and up-projection matrices respectively, σ denotes the activation function (typically GELU), $\lambda_d \in \mathbb{R}^d$ is a learnable scaling vector.This adaptation allows the feedforward network to specialize for style analysis tasks while maintaining parameter efficiency through the lightweight λ_d adjustments.

3.2.6. Final Hidden State Extraction

The contextual representation at the final layer's last token position is extracted:

$$\mathbf{h}_i = \mathbf{H}^{(32)}[\text{last}] \in \mathbb{R}^d \tag{11}$$

This token aggregates information from the entire sequence, capturing pairwise stylistic relationships.

3.2.7. Classification Layer

The hidden state is projected to class probabilities:

$$\mathbf{z} = \mathbf{W}_c\mathbf{h}_i + \mathbf{b}_c, \quad \mathbf{W}_c \in \mathbb{R}^{2\times d} \tag{12}$$

$$P(y_i) = \text{Softmax}(\mathbf{z}) \tag{13}$$

where $\mathbf{z}$ is the logit vector and $P(y_i)$ denotes predicted probabilities for class labels 0 (no change) and 1 (change).

3.2.8. Loss Calculation

Binary cross-entropy loss optimizes model parameters:

$$\mathcal{L}_{ce} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log P(y_i = 1) + (1 - y_i)\log P(y_i = 0)\right] \tag{14}$$

N is batch size, y_i is ground truth label, and $P(y_i)$ is predicted probability.

3.2.9. Implementation Details

Key implementation specifications:

- **Quantization**: NF4 format with double quantization (BitsAndBytesConfig)
- **IA3 Targets**: q_proj, v_proj, down_proj modules
- **Sequence Handling**: Padding/truncation to 512 tokens
- **Optimization**: AdamW ($\eta = 3 \times 10^{-4}$, weight decay $\lambda = 0.01$)

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

We evaluate our method on the PAN 2025 Multi-Author Writing Style Analysis dataset with three difficulty levels:

- **Easy**: 4,200 training documents, 900 validation documents
- **Medium**: 4,200 training documents, 900 validation documents
- **Hard**: 4,200 training documents, 900 validation documents

Data is preprocessed into paragraph pairs with binary labels (change/no-change). Class distribution analysis shows significant imbalance, particularly in Easy task (1:10 ratio).

4.1.2. Model Configuration

- **Base Model**: Meta-Llama-3-8B
- **Quantization**: 4-bit NF4 with double quantization
- **IA3 Targets**: `{q_proj, v_proj, down_proj}`
- **Classification Head**: Single linear layer ($4096 \to 2$)

4.1.3. Training Parameters

The model was trained using the hyperparameters listed in Table 1.

Table 1
Training hyperparameters

Parameter	Value
Batch size	16
Learning rate	3×10^{-4}
Epochs	3
Weight decay	0.01
Warmup ratio	0.1
Max sequence length	512
Gradient checkpointing	Enabled
Mixed precision	FP16

4.1.4. Evaluation Metrics

- Primary metric: Weighted F1-score (handles class imbalance)
- Secondary metrics: Accuracy, Precision, Recall
- Validation: Per-epoch evaluation
- Early stopping: Based on validation F1 improvement

4.1.5. Implementation Environment

- **Hardware**: NVIDIA A800 80GB GPU [9]
- **Frameworks**: PyTorch 2.0, HuggingFace Transformers, PEFT
- **Training Time**: $\approx$ 8 hours per task (3 epochs)

4.2. Results

We finally submitted the model to the TIRA [10]. Table 2 summarizes the performance comparison between our IA3-tuned model (team hellojie) and the naive baseline that always predicts 0 across different difficulty levels on the test set. The proposed approach achieves significant F1-score improvements in all tasks, with the most substantial gain (+32.5%) observed in the Medium-difficulty task.

Table 2
Performance comparison with naive-baseline-predict-0 on test set (F1 scores)

Task	Baseline (predict-0)	Our F1 (hellojie)	Gain
Easy	0.439	0.461	+5.0%
Medium	0.440	0.583	**+32.5%**
Hard	0.453	0.484	+6.8%

Key observations:

- **Medium task dominance**: 32.5% F1 improvement over the naive baseline demonstrates IA3's efficacy in capturing subtle stylistic variations when topic diversity is limited
- **Consistent gains**: Improvements across all difficulty levels validate the robustness of our quantization and tuning approach compared to the trivial baseline

5. Conclusion

This study presents an efficient framework for multi-author writing style analysis by integrating 4-bit quantization and IA3 tuning with the Llama-3-8B model. Our approach demonstrates three key advantages:

1. **Improved performance**: Significant F1 improvements across all difficulty levels (+5.0% Easy, +32.5% Medium, +6.8% Hard), particularly excelling in medium-difficulty tasks where topic consistency demands precise style discrimination.

2. **Computational efficiency**: 4-bit quantization reduces memory requirements by 68% while maintaining competitive accuracy, enabling deployment on resource-constrained systems.

3. **Task-specific adaptation**: IA3's targeted attention modulation (q_proj, v_proj) effectively captures subtle stylistic variations without full parameter updates.

The 32.5% F1 gain in medium-difficulty tasks confirms our hypothesis that IA3 tuning optimizes style representation learning when topic signals are limited. Future work will explore: 1) Dynamic quantization for harder tasks, 2) Multi-task learning across difficulty levels, and 3) Hybrid approaches combining syntactic features with our framework.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No.62276064).

Declaration On Generative AI

During the preparation of this work, the author(s) used DeepSeek in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

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] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korenčić, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of PAN 2024: Multi-author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification Condensed Lab Overview, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. Di Nunzio, L. Soulier, P. Galuščáková, A. García Seco de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction, Springer Nature Switzerland, Cham, 2024, pp. 231–259.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, 2021. URL: https://arxiv.org/abs/2106.09685. arXiv:2106.09685.

[5] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, 2023. URL: https://arxiv.org/abs/2302.13971. arXiv:2302.13971.

[6] H. Liu, D. Tam, M. Muqeeth, J. Mohta, T. Huang, M. Bansal, C. Raffel, Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning, 2022. URL: https://arxiv.org/abs/2205.05638. arXiv:2205.05638.

[7] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, 2020. URL: https://arxiv.org/abs/2004.05150. arXiv:2004.05150.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, Qlora: Efficient finetuning of quantized llms, 2023. URL: https://arxiv.org/abs/2305.14314. arXiv:2305.14314.

[9] V. Korthikanti, J. Casper, S. Lym, L. McAfee, M. Andersch, M. Shoeybi, B. Catanzaro, Reducing activation recomputation in large transformer models, 2022. URL: https://arxiv.org/abs/2205.05198. arXiv:2205.05198.

[10] 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.