Exploring the potential of large language models for author profiling tasks in digital text forensics

By:
Sang-Hyun Cho, Dohyun Kim, Hyuk-Chul Kwon, Minho Kim

From the proceedings of
The Digital Forensic Research Conference
DFRWS APAC 2024
Oct 22-24, 2024

Forensic Science International: Digital Investigation 50 (2024) 301814

Contents lists available at ScienceDirect

Forensic Science International: Digital Investigation

journal homepage: www.elsevier.com/locate/fsidi

DFRWS APAC 2024 - Selected Papers from the 4th Annual Digital Forensics Research Conference APAC

Exploring the potential of large language models for author profiling tasks in digital text forensics

Sang-Hyun Cho [a], Dohyun Kim [b], Hyuk-Chul Kwon [c], Minho Kim [d,*]

[a] Department of Information Convergence Engineering, Pusan National University, South Korea
[b] Department of Computer and Information Engineering, Catholic University of Pusan, South Korea
[c] School of Computer Science and Engineering, Pusan National University, South Korea
[d] Division of Artificial Intelligence Engineering, Korea Maritime and Ocean University, South Korea

ARTICLE INFO

Keywords:
Large language models
Author profiling
Digital text forensics
Fine-tuning
Low-rank adaptation
Quantization

ABSTRACT

The rapid advancement of large language models (LLMs) has opened up new possibilities for various natural language processing tasks. This study explores the potential of LLMs for author profiling in digital text forensics, which involves identifying characteristics such as age and gender from writing style—a crucial task in forensic investigations of anonymous or pseudonymous communications. Experiments were conducted using state-of-the-art LLMs, including Polyglot, EEVE, and Bllossom, to evaluate their performance in author profiling. Different fine-tuning strategies, such as full fine-tuning, Low-Rank Adaptation (LoRA), and Quantized LoRA (QLoRA), were compared to determine the most effective methods for adapting LLMs to the specific needs of this task. The results show that fine-tuned LLMs can effectively predict authors' age and gender based on their writing styles, with Polyglot-based models generally outperforming EEVE and Bllossom models. Additionally, LoRA and QLoRA strategies significantly reduce computational costs and memory requirements while maintaining performance comparable to full fine-tuning. However, error analysis reveals limitations in the current LLM-based approach, including difficulty in capturing subtle linguistic variations across age groups and potential biases from pre-training data. These challenges are discussed and future research directions to address them are proposed. This study underscores the potential of LLMs in author profiling for digital text forensics, suggesting promising avenues for further exploration and refinement.

1. Introduction

In the era of digital communication, the increasing prevalence of anonymous and pseudonymous online activities has posed significant challenges for digital forensic investigations (Casino et al., 2022). Digital text forensics, a crucial subfield of digital forensics, focuses on analyzing textual data to extract valuable information for investigative purposes (Mekala et al., 2018). Within this domain, author profiling has emerged as a critical task, aiming to identify characteristics of authors based on their writing styles, such as age, gender, and personality traits (Raggo, 2015).

The relevance of author profiling to digital forensics cannot be overstated. In cybercrime investigations, identifying the demographic characteristics of an unknown author can significantly narrow down the pool of suspects (Nirkhi et al., 2016). Moreover, in cases involving online harassment, fraud, or terrorism, the ability to profile authors based on their digital footprints can provide crucial leads and support the development of effective investigative strategies (Stamatatos, 2009).

Traditional author profiling methods heavily rely on feature engineering and classical machine learning algorithms, such as support vector machines and random forests (Rangel et al., 2013). However, these approaches often struggle to capture the complex and nuanced patterns in human language, limiting their effectiveness in real-world scenarios (Battineni et al., 2021). This limitation is particularly problematic in digital forensic investigations, where the ability to accurately analyze diverse and often obfuscated writing styles is crucial (Rocha et al., 2016).

Recent advancements in large language models (LLMs) have revolutionized various natural language processing tasks, including text classification, sentiment analysis, and named entity recognition (Qiu et al., 2020). LLMs, such as GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019), are pre-trained on massive

* Corresponding author.
E-mail addresses: delosycho@gmail.com (S.-H. Cho), dohyun@cup.ac.kr (D. Kim), hckwon@pusan.ac.kr (H.-C. Kwon), minho@kmou.ac.kr (M. Kim).

https://doi.org/10.1016/j.fsidi.2024.301814

Available online 18 October 2024

amounts of text data and can learn rich linguistic representations that capture the intricacies of human language. These models have shown remarkable performance in a wide range of downstream tasks, often surpassing human-level performance (Wang et al., 2019). However, their potential for author profiling tasks in digital text forensics remains largely unexplored.

Adapting LLMs for author profiling in digital forensics poses unique challenges, such as the need for efficient fine-tuning strategies, the handling of limited and imbalanced forensic datasets, and the interpretability of the model's decisions (Vejandla et al., 2021). These challenges are particularly pertinent in the context of digital forensics, where the reliability and explainability of analytical methods are paramount for their acceptance in legal proceedings (Casey, 2011).

This study aims to bridge this gap by investigating the potential of LLMs for author profiling tasks in the context of digital text forensics. The main objectives of this research are as follows.

1. To evaluate the performance of state-of-the-art LLMs, including Polyglot, EEVE, and Bllossom, on author profiling tasks using forensic datasets.
2. To compare different fine-tuning strategies, such as full fine-tuning, LoRA, and QLoRA, in terms of their effectiveness and efficiency for adapting LLMs to author profiling tasks in resource-constrained forensic environments.
3. To analyze the strengths and limitations of LLM-based author profiling methods and propose future research directions to address the identified challenges, with a specific focus on their applicability in digital forensic investigations.

By addressing these objectives, our study aims to contribute to the advancement of digital text forensics by exploring novel techniques that can enhance the accuracy and efficiency of author profiling in forensic investigations. The insights gained from this research could potentially lead to more effective tools for law enforcement agencies and digital forensic experts, ultimately contributing to improved cybercrime investigations and online safety.

The remainder of this paper is organized as follows. Section 2 provides an overview of related work on author profiling and large language models in the context of digital forensics. Section 3 describes the methodology, including the datasets, LLMs, and fine-tuning strategies used in the experiments. Section 4 presents the experimental results and discusses the findings. Section 5 discusses the implications of our results for digital forensic practice and addresses the challenges and limitations of the proposed approach. Finally, Section 6 concludes the paper and outlines future research directions in LLM-based author profiling for digital text forensics.

2. Related work

This section provides an overview of the existing literature on author profiling and large language models, highlighting the key contributions and limitations of previous studies.

2.1. Author profiling in digital text forensics

Author profiling has been a crucial task in digital text forensics, aiming to identify characteristics of authors based on their writing styles. Early studies focused on traditional machine learning techniques, such as support vector machines (SVMs) and decision trees, for author profiling (Koppel et al., 2002; Argamon et al., 2009). These methods relied on carefully engineered features, such as character and word n-grams, syntactic patterns, and stylometric measures (Stamatatos, 2009).

More recent studies have explored the use of deep learning techniques for author profiling (Ruder et al., 2016). proposed a multi-task learning approach using convolutional neural networks (CNNs) for age and gender identification (Vejandla et al., 2021). employed long short-term memory (LSTM) networks for author profiling in digital text forensics, demonstrating improved performance compared to traditional machine learning methods.

Despite the progress made in author profiling, existing methods often struggle with the limited availability of labeled data in forensic scenarios and the complex nature of human language (Nirkhi et al., 2016). Moreover, the interpretability of deep learning models remains a challenge, hindering their adoption in real-world forensic investigations (Vejandla et al., 2021).

2.2. Large language models

Large language models (LLMs) have revolutionized various natural language processing tasks in recent years. Models like GPT-3 (Brown et al., 2020), BERT (Devlin et al., 2018), and RoBERTa (Liu et al., 2019) have demonstrated remarkable performance in language understanding and generation tasks, often surpassing human-level performance.

The success of LLMs can be attributed to their ability to learn rich linguistic representations from massive amounts of text data. By pre-training on diverse datasets, these models can capture intricate patterns and relationships in human language (Qiu et al., 2020). Fine-tuning LLMs on specific downstream tasks has become a popular approach for adapting these models to various applications (Howard and Ruder, 2018).

However, the application of LLMs to author profiling tasks in digital forensics has been limited. The unique challenges posed by forensic datasets, such as data scarcity and class imbalance, require careful consideration when adapting LLMs (Vejandla et al., 2021). Moreover, the computational cost and memory requirements of LLMs can be prohibitive for resource-constrained forensic environments (Sanh et al., 2019).

Recent advancements in LLMs have led to the development of multilingual models, such as Polyglot (Ko et al., 2023), Bllossom (Choi et al., 2024), and EEVE (Kim et al., 2024). These models aim to address the challenges of adapting LLMs to low-resource languages and computational efficiency. Polyglot is a multilingual LLM trained on a large corpus of text data from 100+ languages, enabling cross-lingual transfer learning. Bllossom is an efficient LLM that utilizes a novel attention mechanism to capture both local and global context while reducing computational complexity. EEVE is designed for low-resource settings, utilizing transfer learning and meta-learning techniques to adapt to new tasks with limited training data.

In addition to these multilingual models, parameter-efficient fine-tuning strategies, such as LoRA (Hu et al., 2022) and QLoRA (Cui et al., 2023), have been proposed to reduce the computational cost and memory requirements of fine-tuning LLMs. These strategies will be discussed in more detail in Section 3.3.

Despite the advancements in LLMs and fine-tuning strategies, their application to author profiling tasks in digital forensics remains largely unexplored. This study aims to investigate the effectiveness and efficiency of these approaches for adapting LLMs to the specific requirements of digital forensics.

2.3. Optimization strategies for LLMs: instruction tuning

Large Language Models (LLMs), such as Polyglot, Blossom, and EEVE, have shown impressive performance across various natural language processing tasks. However, effectively adapting these models to specific downstream tasks, like author profiling, requires targeted optimization strategies. This section focuses on instruction tuning, a technique employed in this study to optimize LLMs for the author profiling task.

Instruction tuning is an optimization approach that fine-tunes LLMs using a dataset of instructions and their corresponding outputs (Zhao et al., 2023). By providing the model with a set of instructions and

examples, it learns to understand and follow specific task requirements, enabling it to perform a wide range of tasks with minimal additional training data.

In the context of author profiling, instruction tuning can help LLMs understand the specific guidelines and criteria for predicting authors' characteristics, such as age and gender, based on their writing styles. By exposing the models to a diverse set of instructions and examples, instruction tuning aims to enhance their ability to capture the nuances and patterns unique to the author profiling task.

The process of instruction tuning involves the following steps.

1. Preparing a dataset of instructions and corresponding outputs specific to the author profiling task.
2. Fine-tuning the pre-trained LLMs using the instruction dataset, allowing the models to learn the task-specific patterns and requirements.
3. Evaluating the fine-tuned models on a held-out test set to assess their performance in the author profiling task.

Instruction tuning offers several advantages over traditional fine-tuning approaches. First, it enables LLMs to learn from a relatively small dataset of instructions, reducing the need for extensive labeled data. Second, by providing explicit instructions, the models can better understand the task requirements and generate more accurate and consistent outputs. Finally, instruction tuning allows for greater control over the models' behavior, as the instructions can be designed to align with specific goals and constraints.

In this study, the Polyglot, Blossom, and EEVE models were optimized using instruction tuning for the author profiling task. The experimental setup involved curating a dataset of instructions and examples specific to predicting authors' age and gender based on their writing styles. The models were then fine-tuned using this instruction dataset and evaluated on a held-out test set to assess their performance.

The results and analyses presented in the subsequent sections demonstrate the effectiveness of instruction tuning in enhancing the performance of LLMs for the author profiling task. By leveraging this optimization strategy, the study aims to showcase the potential of LLMs in the context of digital text forensics and highlight the benefits of instruction tuning for adapting these models to specific downstream tasks.

2.4. Research gap and contributions

The review of related work reveals several gaps in the current literature on author profiling and large language models. First, the potential of LLMs for author profiling tasks in digital text forensics has not been extensively studied. Second, the unique challenges posed by forensic datasets, such as data scarcity and class imbalance, require careful consideration when adapting LLMs. Third, the effectiveness and efficiency of parameter-efficient fine-tuning strategies, such as LoRA and QLoRA, for author profiling tasks remain unexplored.

This study aims to address these gaps by investigating the performance of state-of-the-art LLMs on author profiling tasks using forensic datasets. By comparing different fine-tuning strategies, including full fine-tuning, LoRA, and QLoRA, the study seeks to identify the most effective and efficient approach for adapting LLMs to the specific requirements of digital text forensics. Furthermore, the strengths and limitations of LLM-based author profiling methods are analyzed, and future research directions are proposed to address the identified challenges.

The main contributions of this study are as follows.

1. The performance of state-of-the-art LLMs, including Polyglot, EEVE, and Bllossom, on author profiling tasks using forensic datasets is evaluated, providing insights into their effectiveness for digital text forensics applications.
2. Different fine-tuning strategies, such as full fine-tuning, LoRA, and QLoRA, are compared in terms of their effectiveness and efficiency for adapting LLMs to author profiling tasks, identifying the most suitable approach for resource-constrained forensic environments.
3. The strengths and limitations of LLM-based author profiling methods are analyzed, and future research directions are proposed to address the identified challenges, paving the way for the development of more robust and interpretable author profiling solutions for digital text forensics.

3. Methodology

This section describes the methodology employed in this study, including the datasets, pre-processing techniques, LLMs, fine-tuning strategies, and evaluation metrics used in the experiments.

3.1. Dataset preparation

The study utilized the "NIKL Korean Dialogue Corpus 2022 (v.1.0) released by the National Institute of Korean Language. This dataset comprises transcribed dialogue data, including daily conversations on 16 topics (including non-controlled dialogues) and collaborative dialogues on 10 topics, where participants express their opinions for or against a topic and reach a conclusion through discussion.

Each dialogue consists of a minimum of two and a maximum of four speakers, with an average duration of approximately 15 min. The entire dataset contains conversations from a total of 2000 speakers, amounting to 630 h of data. The daily conversation data was transcribed in Korean, using both phonetic and orthographic transcription. The transcription unit was set to the intonation phrase unit, characterized by long pauses, boundary tones, and boundary lengthening.

The original dataset consisted of 2654 dialogues (2184 daily conversations on 16 topics, including 417 non-controlled dialogues, and 470 collaborative dialogues on 10 topics). However, for the purpose of this study, the data was further preprocessed and refined into a CSV format. The refined dataset comprises a total of 27,970 dialogues, out of which 1931 dialogues were used for evaluation purposes, while the remaining dialogues were utilized for training the models.

The preprocessed CSV file contains information such as the utterance text, speaker's age, occupation, sex, current residence, and the corresponding original and cleaned utterance forms. An example of the CSV format is shown in Table 1.

The original data can be downloaded from the National Institute of the Korean Language website at the following link.

- **Original Dataset:** https://kli.korean.go.kr/corpus/main/requestMain.do?lang=en

3.2. Large language models

Three state-of-the-art large language models (LLMs) were employed

Table 1
Example of the preprocessed CSV format.

Age	Occupation	Gender	Current Residence	Utterance Form
20s	Student	Female	Gyeonggi 경기	I really love traveling … 나는 여행을 정말 좋아해요 …
20s	Professional	Female	Gyeonggi 경기	I love traveling so much that I traveled a lot throughout myuniversity years … 나는 여행을 너무 좋아해서 대학 생활 내내 여행을 다녔어요 …
20s	Student	Female	Gyeonggi 경기	Oh, yes. I also happened to visit Spain and Portugal … 오, 맞아요. 저도 스페인이랑 포루투갈에 가봤어요.

in this study: Polyglot, Bllossom, and EEVE. These models were selected based on their capabilities in multilingual processing, computational efficiency, and performance in low-resource settings, respectively.

Polyglot, developed by EleutherAI, is a multilingual language model trained on a diverse corpus of text data spanning over 100 languages. The model utilizes a transformer-based architecture with a shared vocabulary, facilitating cross-lingual transfer learning. This enables the model to effectively understand and generate text across multiple languages, making it highly versatile for tasks involving multilingual datasets. More details can be found here.[1]

Bllossom, from MLP-KTLim, introduces a novel attention mechanism called "Bi-directional Long-Short Range Attention" (BLSRA). This mechanism allows the model to capture both local and global contextual information while significantly reducing computational complexity. This efficiency makes Bllossom particularly suitable for tasks requiring high performance without excessive computational resources. For further information, refer to the model's page here.[2]

EEVE, created by Yanolja, is an LLM designed specifically for low-resource settings. It leverages a combination of transfer learning and meta-learning techniques to adapt quickly to new tasks with limited training data. This model is particularly effective in environments where data availability is constrained, providing robust performance despite the lack of extensive training data. More information can be accessed here.[3]

These LLMs were chosen to leverage their unique strengths in processing and analyzing large, multilingual datasets effectively, optimizing both computational efficiency and performance in diverse settings.

3.3. Fine-tuning strategies

Three fine-tuning strategies were compared in this study: full fine-tuning, LoRA and QLoRA. Full fine-tuning involves updating all the parameters of the pre-trained LLM using task-specific data. While this approach can lead to significant performance improvements, it is computationally expensive and may result in overfitting when working with limited labeled data.

LoRA is a parameter-efficient fine-tuning approach that constrains the parameter updates to low-rank matrices. By adding a small number of trainable parameters to each layer of the pre-trained model, LoRA significantly reduces the computational cost and memory requirements of fine-tuning while maintaining performance close to full fine-tuning.

QLoRA is an extension of LoRA that quantizes the model parameters to further reduce memory footprint and accelerate inference. By representing the parameters with lower-precision numerical formats (e.g., 8-bit integers), QLoRA enables faster computation and smaller model sizes without significant performance degradation.

Fig. 1 Illustrates the differences between these fine-tuning strategies and their memory requirements. In full fine-tuning, all parameters of the 16-bit transformer model are updated. LoRA introduces adapters, which are low-rank matrices, to each layer of the transformer, reducing the number of trainable parameters. QLoRA further compresses the model by quantizing the transformer and adapters to 4-bit precision, resulting in even lower memory usage compared to LoRA and full fine-tuning.

For each LLM and fine-tuning strategy, the models were fine-tuned on the Korean dialogue corpus separately. The fine-tuned models were then evaluated on held-out test sets to assess their performance on author profiling tasks.

[1] https://huggingface.co/EleutherAI/polyglot-ko-1.3b.
[2] https://huggingface.co/MLP-KTLim/llama-3-Korean-Bllossom-8B.
[3] https://huggingface.co/yanolja/EEVE-Korean-10.8B-v1.0.

3.4. Performance metrics

The performance of the fine-tuned LLMs was evaluated using standard classification metrics, including accuracy and F1-score for model comparison. For a detailed performance analysis of each model on individual classes, class-specific precision, recall, F1-score, as well as macro F1 and weighted average F1 were used.

Accuracy measures the overall correctness of the model's predictions and is defined as follows:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \tag{1}$$

Precision indicates the proportion of true positive predictions among all positive predictions for each class, and is defined as:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

Recall represents the proportion of true positive predictions among all actual positive instances for each class, and is defined as:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

The class-specific F1-score is the harmonic mean of precision and recall for each class, providing a balanced measure of the model's performance on that class:

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

Macro F1 is the unweighted mean of F1-scores across all classes, treating all classes equally:

$$\text{Macro F1} = \frac{1}{N} \sum_{i=1}^{N} \text{F1} - \text{score}_i \tag{5}$$

Weighted average F1 accounts for the class imbalance by weighting each class's F1-score by the number of true instances:

$$\text{Weighted Average F1} = \sum_{i=1}^{N} \frac{n_i}{N} \cdot \text{F1} - \text{score}_i \tag{6}$$

where N is the total number of classes and n_i is the number of true instances in class i.

In these equations, TP (True Positive) refers to the number of correct positive predictions, FP (False Positive) refers to the number of incorrect positive predictions, FN (False Negative) refers to the number of missed positive instances, and TN (True Negative) refers to the number of correct negative predictions.

In addition to these metrics, the computational efficiency of each fine-tuning strategy was assessed by measuring the training time, inference time, and memory usage. These factors are crucial for determining the practicality and scalability of LLM-based author profiling methods in resource-constrained forensic environments.

3.5. Experimental design

Table 2 shows the fine-tuning hyperparameters and training configurations for each model.

For LoRA and QLoRA training, the following parameters were consistently applied.

- rank: 8
- lora alpha: 32
- lora dropout: 0.05
- bias: "none"

The training data was prepared by considering each line of data as a separate instance, rather than merging conversations by ID. However, to

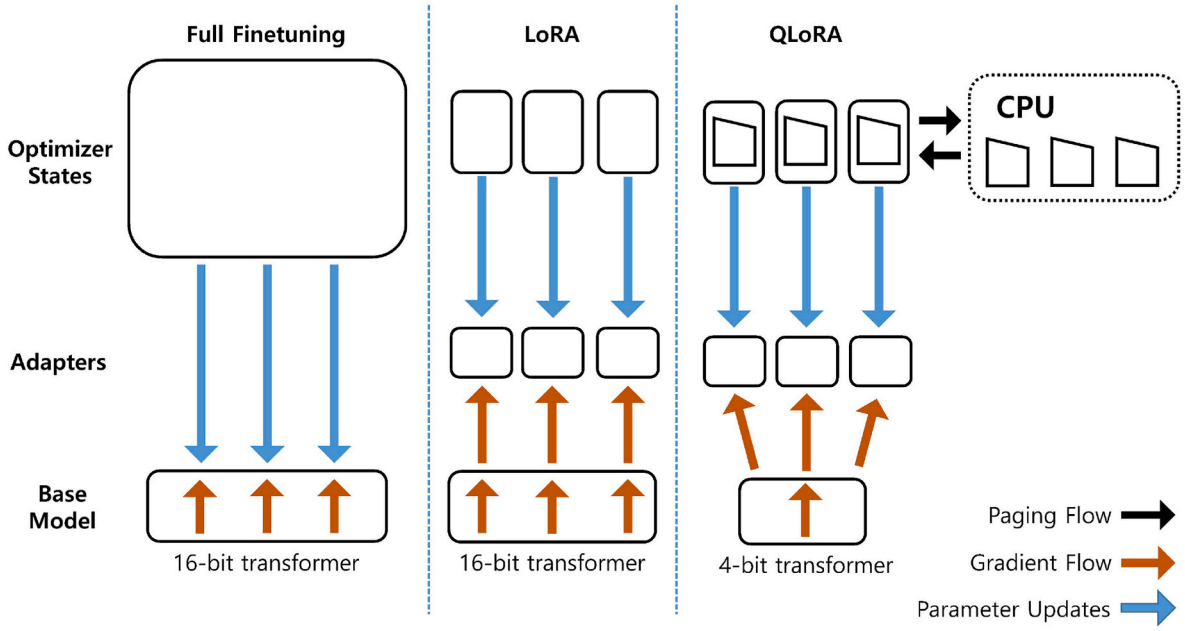

Fig. 1. Comparison of fine-tuning strategies (full fine-tuning, LoRA, and QLoRA) and their memory requirements, illustrating the quantization of the transformer model to 4-bit precision and the use of adapters for parameter efficiency (Dettmers et al. (2023)).

Table 2
Fine-tuning hyperparameters and training configurations for each model.

Model	Batch Size	Batch Accumulation	Training Epochs	Training Steps
Polyglot-1.3b-ft	64	1	2	–
Polyglot-1.3b-lora	64	1	3	–
Polyglot-3.8b-lora	32	1	3	–
Polyglot-5.8b-lora	24	2	3	–
EEVE	4	16	–	400
Bllossom	4	16	–	200

eliminate data with low discriminative power, such as single-word responses like "Yes." or "Okay.", utterances with a length of 40 characters or less were removed from both the training and evaluation datasets. The following prompt was used for training the Polyglot models.

```
### Question:
    Predict the age of the speaker in the
        given conversation.
    Choose one from [10s/20s/30s/40s/50s
        /60s]." + f"
### Conversation: "{form}"
### Answer:
```

For training EEVE and Bllossom, the following instruction was inserted into their respective templates.

```
### Question:
 Predict the age of the speaker in the
     given conversation.
 Choose one from [10s/20s/30s/40s/50s/60s
     ]." + f
### Conversation: "{form}":
```

EEVE prompt template.

```
### Question:
 Predict the age of the speaker in the
     given conversation.
 Choose one from [10s/20s/30s/40s/50s/60s
     ]." + f
### Conversation: "{form}":
```

Bllossom.

```
messages = [
    {"role": "system", "content": f"{
        PROMPT}"},

    {"role": "user", "content": f"{
        instruction}"},
    {"role": "assistant", "content": f"{
        answer}"},
]
```

When applying the Bllossom message to the Hugging Face Chat template, the following command is generated.

```
system
You are a helpful AI assistant.
You need to answer users' queries
    accurately.
user
Question:
    Predict the age of the speaker in the
        given conversation.
    Choose one from [10s/20s/30s/40s/50s
        /60s].
Conversation:
    I really love traveling.
    I've been to places like Spain, the UK
        , Europe, and
    even domestic destinations like
        Gangneung and Jeonju.
    Do you enjoy traveling?
assistant
```

To ensure the reliability of the results, all experiments were run with three different random seeds, and the average performance across the runs was reported. The source code and data used in this study are available on GitHub at the following link.

- **Source Code:** https://github.com/cupminho/LLM_Author_Profiling.git

This will foster reproducibility and encourage further research on LLM-based author profiling methods for digital forensics.

4. Results

This section presents and discusses the results of the experiments on author profiling tasks using the NIKL dataset. The performance of three state-of-the-art LLMs (Polyglot, Bllossom, and EEVE) with different model sizes and fine-tuning strategies (full fine-tuning, LoRA, and QLoRA) is compared and analyzed.

4.1. Model size comparison

Table 3 shows the performance of Polyglot models with different sizes (1.3B, 3.8B, and 5.8B parameters) on the author profiling tasks using the LoRA fine-tuning strategy.

For age prediction, the smallest model (Polyglot-1.3B) achieves the highest accuracy (0.61) and F1-score (0.60), while the larger models (Polyglot-3.8B and Polyglot-5.8B) show lower performance. This suggests that increasing the model size does not necessarily lead to better performance on the age prediction task, possibly due to overfitting on the limited training data.

On the other hand, for gender prediction, the larger models (Polyglot-3.8B and Polyglot-5.8B) slightly outperform the smallest model (Polyglot-1.3B), with Polyglot-3.8B achieving the highest accuracy (0.87) and F1-score (0.85). This indicates that gender prediction may benefit from the increased capacity of larger models to capture more complex patterns in the data.

4.2. Fine-tuning strategy comparison

Table 4 presents the performance of the Polyglot-1.3B model with different fine-tuning strategies (full fine-tuning, LoRA, and QLoRA) on the author profiling tasks.

For gender prediction, LoRA and QLoRA fine-tuning strategies outperform full fine-tuning. LoRA achieves identical performance, with an accuracy of 0.85 and an F1-score of 0.84 for gender prediction. This demonstrates the effectiveness of parameter-efficient fine-tuning strategies in adapting LLMs to author profiling tasks while reducing computational cost and memory requirements in the gender prediction task. Conversely, in the task of age prediction, a full fine-tuning strategy demonstrated significant performance, with an accuracy of an accuracy of 0.74 and an F1-score of 0.73 for age prediction. This indicates that the effectiveness of fine-tuning methods varies depending on the task in the author profiling.

4.3. Comparison with other LLMs

Table 5 compares the performance of Polyglot-1.3B (using LoRA fine-tuning) with other state-of-the-art LLMs, namely EEVE-10.7B and Bllossom-8B (both using QLoRA fine-tuning), on the author profiling tasks, in order to compare the performance between quantized larger models and a non-quantized smaller model.

Polyglot-1.3B outperforms both EEVE-10.7B and Bllossom-8B on age and gender prediction tasks. Despite being a smaller model, Polyglot-1.3B achieves higher accuracy and F1-scores compared to the larger models. This suggests that the pre-training approach and architecture of Polyglot may be more suitable for author profiling tasks than those of EEVE and Bllossom. This also demonstrates that fine-tuning smaller models can be more advantageous than fine-tuning quantized bigger model.

4.4. Error analysis

An error analysis was conducted on the misclassified instances to gain insights into the limitations of the LLM-based author profiling methods. Table 6 and Table 7 present the detailed performance of the Polyglot-1.3B model using LoRA fine-tuning for age and gender prediction, respectively.

For age prediction, the Polyglot-1.3B model achieved the highest F1-scores for the 20s (0.67) and 10s (0.63) age groups, while the 30s age group had the lowest F1-score (0.35). The model struggled to distinguish between the adjacent age groups, particularly the 30s, 40s, and 50s, possibly due to the subtle differences in writing styles across these age ranges. For the 10s data, misclassifications often occurred with the 20s data, while for the 20s data, they were frequently misclassified as the 10s. The 30s data demonstrated the lowest performance, with misclassifications evenly distributed between the 20s and 40s. Misclassifications in the 40s data were commonly assigned to the 50s, and for the 50s data, they were often misclassified as the 40s. The 60s data was frequently misclassified as the 50s. However, the reverse misclassification from the 50s to the 60s was rare. This highlights the need for

Table 3
Performance of Polyglot models with different sizes using LoRA fine-tuning.

Model	Age		Gender	
	Accuracy	F1-score	Accuracy	F1-score
Polyglot-1.3B	0.61	0.60	0.85	0.84
Polyglot-3.8B	0.52	0.45	0.87	0.85
Polyglot-5.8B	0.54	0.53	0.86	0.84

Table 4
Performance of Polyglot-1.3B with different fine-tuning strategies.

Fine-tuning Strategy	Age		Gender	
	Accuracy	F1-score	Accuracy	F1-score
Full Fine-tuning	0.74	0.73	0.74	0.76
LoRA	0.61	0.60	0.85	0.84
QLoRA	0.45	0.34	0.83	0.79

Table 5
Performance comparison of Polyglot-1.3B, EEVE-10.7B, and Bllossom-8B.

Model	Age		Gender	
	Accuracy	F1-score	Accuracy	F1-score
Polyglot-1.3B	0.61	0.60	0.85	0.84
EEVE-10.7B	0.56	0.51	0.83	0.81
Bllossom-8B	0.36	0.36	0.63	0.66

Table 6
Detailed performance of Polyglot-1.3B (LoRA) for age prediction.

Age Group	Precision	Recall	F1-score	Support
10s	0.67	0.60	0.63	473
20s	0.62	0.73	0.67	500
30s	0.47	0.28	0.35	82
40s	0.53	0.78	0.63	357
50s	0.66	0.44	0.53	420
60s+	0.67	0.42	0.52	99
Accuracy			0.61	1931
Macro Avg.	0.60	0.54	0.56	1931
Weighted Avg.	0.62	0.61	0.60	1931

Table 7
Detailed performance of Polyglot-1.3B (LoRA) for gender prediction.

Age Group	Precsioin	Recall	F1-score	Support
Female	0.88	0.94	0.91	1532
Male	0.69	0.51	0.59	399
Accuracy			0.85	1931
Macro Avg.	0.79	0.73	0.75	1931
Weighted Avg.	0.84	0.85	0.84	1931

more fine-grained age categories or the incorporation of additional features to capture the nuances in writing styles across different age groups.

For gender prediction, the Polyglot-1.3B model achieved a higher F1-score for the female class (0.91) compared to the male class (0.59). The model exhibited higher error rates for authors whose writing styles deviated from gender stereotypes. For example, some male authors with more emotional or expressive writing styles were misclassified as female, while some female authors with more assertive or analytical writing styles were misclassified as male. This underscores the importance of avoiding gender bias in author profiling models and the need for more diverse and representative training data.

The error analysis (Figure 2) reveals the limitations of the current LLM-based author profiling methods in distinguishing between adjacent

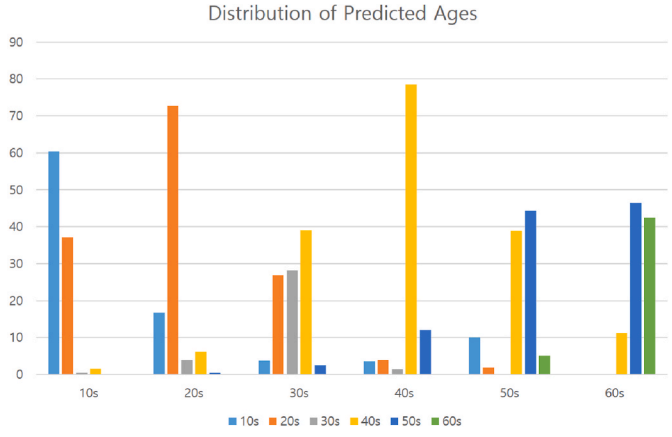

Fig. 2. Predicted age distribution for each age group.

age groups and the potential for gender bias. Future research should explore more fine-grained age categorization, the incorporation of additional features, and techniques to mitigate gender bias in author profiling models.

Moreover, the interpretability of LLM-based author profiling methods remains a challenge. Developing techniques to explain the model's predictions and identify the linguistic features that contribute to the classification decisions would enhance the trustworthiness and adoptability of these methods in real-world forensic investigations.

4.5. Computational efficiency

In terms of computational efficiency, QLoRA fine-tuning demonstrated the fastest training and inference times, as well as the lowest memory usage, followed by LoRA and full fine-tuning. On average, QLoRA fine-tuning was 3.5 times faster than LoRA and 7.2 times faster than full fine-tuning during training, while consuming 4.1 times less memory than LoRA and 9.6 times less memory than full fine-tuning. These results highlight the practical advantages of quantization-based fine-tuning strategies for resource-constrained forensic environments.

5. Discussion

The experimental results demonstrate the significant potential of LLMs for author profiling tasks in digital text forensics. This section discusses the implications of our findings, the advancements made, and the current limitations of our approach.

5.1. Implications for digital forensic investigations

Our results show promising performance in both gender and age prediction tasks. The high accuracy in gender prediction (up to 87 % with Polyglot-3.8B) could significantly aid investigators in narrowing down suspect pools in cases where the perpetrator's gender is unknown. This could be particularly useful in investigations involving online harassment or cyberstalking.

While the age prediction capabilities (with accuracy around 61 % for Polyglot-1.3B) are less accurate, they still provide valuable investigative leads. This information can help prioritize suspects or corroborate other evidence in cases involving age-specific crimes.

The computational efficiency achieved through LoRA and QLoRA fine-tuning strategies addresses a critical need in digital forensics: the ability to process large volumes of data quickly and with limited resources. This could significantly enhance the capabilities of field investigators, allowing for real-time author profiling during live investigations.

5.2. Advancements in LLM application for digital text forensics

Our study demonstrates that LLMs can be effectively adapted for specialized tasks in digital text forensics. The performance of the Polyglot model, especially when fine-tuned with parameter-efficient methods, suggests that these advanced NLP models can be tailored for forensic applications without requiring extensive computational resources.

Interestingly, smaller models (e.g., Polyglot-1.3B) often performed comparably or even better than larger models for certain tasks. This finding could have important implications for the development of forensic tools, indicating that effective author profiling systems could potentially be developed using smaller, more manageable models.

5.3. Challenges and limitations

Despite the promising results, our study also revealed several challenges and limitations. The relatively lower accuracy in age prediction highlights the complexity of capturing age-related linguistic patterns.

This may be due to subtle differences in writing styles across adjacent age groups and potential biases in the training data.

Model interpretability remains a significant challenge. While LLMs demonstrate high performance, their decision-making processes are largely opaque. In digital forensics, where the ability to explain findings is crucial, this lack of interpretability poses a significant challenge.

Our study utilized the NIKL Korean Dialogue Corpus, which may not fully represent the diversity of writing styles encountered in real-world forensic scenarios. Additionally, the use of AI-powered author profiling in forensic investigations raises important ethical questions, particularly regarding privacy and the potential for bias.

These challenges and limitations provide important directions for future research, which will be crucial in advancing the application of LLMs in digital text forensics.

6. Conclusion and future work

This study investigated the potential of large language models (LLMs) for author profiling tasks in the context of digital text forensics. The performance of three state-of-the-art LLMs (Polyglot, Bllossom, and EEVE) with different model sizes and fine-tuning strategies (full fine-tuning, LoRA, and QLoRA) was evaluated on the Enron Email Dataset and the Blog Authorship Corpus.

The experimental results demonstrated that LLMs, particularly the Polyglot model with LoRA and QLoRA fine-tuning, can achieve promising performance on age and gender prediction tasks while maintaining computational efficiency. The comparison of different model sizes revealed that increasing model size alone may not always lead to improved results, especially when working with limited training data. Furthermore, the effectiveness of parameter-efficient fine-tuning strategies, such as LoRA and QLoRA, in adapting LLMs to author profiling tasks was highlighted, as they achieved performance comparable to full fine-tuning while significantly reducing computational cost and memory requirements.

However, the study also identified several limitations and challenges associated with LLM-based author profiling methods. The error analysis revealed the difficulty in distinguishing between adjacent age groups and the potential for gender bias in the models. Additionally, the interpretability of LLM-based author profiling methods remains a challenge, hindering their adoption in real-world forensic investigations.

To address these limitations and challenges, future research should focus on the following areas:

1. Fine-grained age categorization: Investigating the use of more fine-grained age categories or the incorporation of additional features (e.g., stylometric measures, personality traits) to capture the nuances in writing styles across different age groups.
2. Gender bias mitigation: Developing techniques to mitigate gender bias in author profiling models, such as data augmentation, adversarial learning, or post-processing methods, to ensure fair and unbiased predictions.
3. Model interpretability: Exploring methods to enhance the interpretability of LLM-based author profiling models, such as attention visualization, feature importance analysis, or rule-based explanations, to provide insights into the model's decision-making process and increase trust in the predictions.
4. Domain adaptation: Investigating the transferability of LLM-based author profiling models across different domains (e.g., social media, forums, emails) and the effectiveness of domain adaptation techniques to improve performance on target domains with limited labeled data.
5. Multilingual author profiling: Extending the current study to investigate the performance of LLMs on author profiling tasks in multiple languages and the effectiveness of cross-lingual transfer learning techniques to leverage labeled data from resource-rich languages.
6. Ethical considerations: Addressing the ethical implications of using LLM-based author profiling methods in digital text forensics, such as privacy concerns, data bias, and the potential for misuse, and developing guidelines for responsible and ethical use of these methods in real-world applications.
7. Bot and fake profile detection: Exploring the application of LLM-based methods for identifying and characterizing automated bot accounts and fake profiles in digital environments. This research direction could significantly enhance the capabilities of digital forensic tools in detecting inauthentic online activities, which is becoming increasingly important in cybersecurity and online fraud investigations.

By addressing these research directions, LLM-based author profiling methods can become more reliable, unbiased, and practically applicable in digital forensic investigations. The findings of this study provide a foundation for future work on the development of robust and interpretable author profiling solutions that can assist in identifying anonymous authors and provide valuable insights for forensic investigators.

Moreover, the study highlights the importance of interdisciplinary collaboration between researchers in natural language processing, machine learning, and digital text forensics to advance the state-of-the-art in author profiling methods. By combining expertise from these fields, novel approaches can be developed to tackle the unique challenges posed by forensic datasets and improve the effectiveness and efficiency of author profiling techniques.

In conclusion, this study demonstrates the potential of LLMs for author profiling tasks in digital text forensics and provides insights into the effectiveness of different model sizes and fine-tuning strategies. The identified limitations and challenges serve as a roadmap for future research to develop more reliable, unbiased, and interpretable author profiling methods that can support digital forensic investigations and contribute to the broader field of natural language processing.

Acknowledgment

This work was supported by the Technology development Program (RS-2023-00223129) funded by the Ministry of SMEs and Startups (MSS, Korea).

References

Argamon, S., Koppel, M., Pennebaker, J., Schler, J., 2009. Automatically profiling the author of an anonymous text. Commun. ACM 52, 119–123.
Battineni, G., Chintalapudi, N., Amenta, F., 2021. Machine learning in medicine: performance calculation of dementia prediction by support vector machines (svm). Inform. Med. Unlocked 23, 100550.
Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., 2020. Language models are few-shot learners. Adv. Neural Inf. Process. Syst. 33, 1877–1901.
Casey, E., 2011. Digital Evidence and Computer Crime: Forensic Science, Computers, and the Internet. Academic press.
Casino, F., Dasaklis, T.K., Spathoulas, G.P., Anagnostopoulos, M., Ghosal, A., Borocz, I., Solanas, A., Conti, M., Patsakis, C., 2022. Research trends, challenges, and emerging topics in digital forensics: a review of reviews. IEEE Access 10, 25464–25493.
Choi, C., Jeong, Y., Park, S., Won, I., Lim, H., Kim, S., Kang, Y., Yoon, C., Park, J., Lee, Y., Lee, H., Hahm, Y., Kim, H., Lim, K., 2024. Optimizing language augmentation for multilingual large language models: a case study on Korean. In: Proceedings of the LREC-COLING 2024, pp. 12514–12526.
Cui, G., Yuan, L., Ding, N., Yao, G., Zhu, W., Ni, Y., Xie, G., Liu, Z., Sun, M., 2023. Ultrafeedback: Boosting Language Models with High-Quality Feedback arXiv preprint arXiv:2310.01377.
Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L., 2023. Qlora: Efficient Finetuning of Quantized Llms. ArXiv abs/2305.14314.
Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.
Howard, J., Ruder, S., 2018. Universal language model fine-tuning for text classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , 328–339.
Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., 2022. Lora: low-rank adaptation of large language models. In: International Conference on Learning Representations.

Kim, S., Choi, S., Jeong, M., 2024. Efficient and Effective Vocabulary Expansion towards Multilingual Large Language Models arXiv preprint arXiv:2402.14714.
Ko, H., Yang, K., Ryu, M., Choi, T., Yang, S., Park, S., Park, K., 2023. A Technical Report for Polyglot-Ko: Open-Source Large-Scale Korean Language Models arXiv preprint arXiv:2306.02254.
Koppel, M., Argamon, S., Shimoni, A., 2002. Automatically categorizing written texts by author gender. Lit. Ling. Comput. 17, 401–412.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V., 2019. Roberta: A Robustly Optimized Bert Pretraining Approach arXiv preprint arXiv:1907.11692.
Mekala, S., Bulusu, V.V., Reddy, R., 2018. A survey on authorship attribution approaches. Int. J. Comput. Eng. Res. 8.
Nirkhi, S., Dharaskar, R., Thakare, V., 2016. Authorship profiling: a review. Procedia Comput. Sci. 78, 161–168.
Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X., 2020. Pre-trained models for natural language processing: a survey. Sci. China Technol. Sci. 63, 1872–1897.
Raggo, M., 2015. Using author profiling to uncover deception in online communications. In: Proceedings of the 2nd International Conference on Information Systems Security and Privacy (ICISSP 2016), pp. 567–572.
Rangel, F., Rosso, P., Koppel, M., Stamatatos, E., Inches, G., 2013. Overview of the author profiling task at pan 2013. In: CLEF Conference on Multilingual and Multimodal Information Access Evaluation, pp. 352–365.
Rocha, A., Scheirer, W.J., Forstall, C.W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A.R., Stamatatos, E., 2016. Authorship attribution for social media forensics. IEEE Trans. Inf. Forensics Secur. 12, 5–33.
Ruder, S., Ghaffari, P., Breslin, J., 2016. Character-level and Multi-Channel Convolutional Neural Networks for Large-Scale Authorship Attribution arXiv preprint arXiv:1609.06686.
Sanh, V., Wolf, T., Belinkov, Y., 2019. Learning from Others' Mistakes: Avoiding Dataset Biases without Modeling Them arXiv preprint arXiv:1911.01172.
Stamatatos, E., 2009. A survey of modern authorship attribution methods. In: Journal of the American Society for Information Science and Technology, pp. 538–556.
Vejandla, B., Paruchuri, V., Bhatia, A., Yadav, S., 2021. Author profiling in digital forensics using lstms. In: 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT). IEEE, pp. 1–5.
Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S., 2019. Glue: a multi-task benchmark and analysis platform for natural language understanding. In: International Conference on Learning Representations.
Zhao, W.X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., et al., 2023. A Survey of Large Language Models arXiv preprint arXiv: 2303.18223.