Native Language Identification Improves Authorship Attribution

Ahmet Yavuz Uluslu and **Gerold Schneider**
University of Zurich
{ahmet.uluslu, gerold.schneider}@linguistik.uzh.ch

Can Yildizli
PRODAFT
can@prodaft.com

Abstract

This study investigates the integration of native language identification into authorship attribution, a previously unexplored aspect that is particularly important in multilingual contexts. We introduce AA-NLI50, a new dataset containing both native language and authorship information. We propose a novel chain-of-thought approach for native language identification. Our findings demonstrate that our system significantly enhances authorship attribution performance, with results showing a mean accuracy improvement of 9% over baseline methods.

1 Introduction

Authorship attribution—the task of identifying the author of a given document based on a set of possible candidates—is crucial in various forensic applications (Koppel et al., 2009). The complexity of the task increases significantly with the number of potential candidates and the scarcity of training data (Luyckx and Daelemans, 2011; Rocha et al., 2016). Most recent studies integrated author profiles that include attributes such as gender, education level and age to refine the attribution process and narrow down the pool of suspected authors (Deutsch and Paraboni, 2023). Despite these advances, the impact of native language in authorship analysis remains largely unexplored, even though it is frequently mentioned in forensic applications, particularly in the context of cybercrime (Perkins, 2021).

Native Language Identification (NLI) is the task of automatically identifying the native language (L1) of an individual based on their writing or speech in another language (L2). The underlying hypothesis is that L1 affects L2 production due to cross-linguistic influence (Yu and Odlin, 2016). Recent findings in neuroscience suggest that structural differences in the brain can result from the influence of the native language (Wei et al., 2023).

The majority of NLI research relies on learner corpora, such as TOEFL11 (Blanchard et al., 2013) and ICLE (Granger et al., 2009). The training examples consist of formal writings in classroom settings that differ considerably from the context and register of ransomware notes or forum posts on the dark web (Jin et al., 2022). The mismatch can have a negative impact on the analysis and affect the overall performance (Grieve, 2023). Previous studies have demonstrated that state-of-the-art NLI systems often struggle to generalize across different topics and linguistic registers (Malmasi and Dras, 2018). While datasets derived from social media platforms such as Reddit (Goldin et al., 2018; Murauer and Specht, 2021) offer a diverse range of samples within an informal register, which helps to mitigate this issue, they still encounter significant challenges related to topic imbalance.

From an architectural standpoint, NLI followed the trend towards attention-based transformation models such as BERT (Steinbakken and Gambäck, 2020), BigBird (Kramp et al., 2023) and GPT-2 (Lotfi et al., 2020). To address practical problems, more recent work has focused on improving scalability (Uluslu and Schneider, 2022) and extending these models to languages other than English (Malmasi and Dras, 2017; Uluslu and Schneider, 2023). The emergence of more powerful large language models such as GPT-4 (Achiam et al., 2023) and Mixtral 8x7B (Jiang et al., 2024) enabled increasing capabilities in zero-shot learning, achieving state-of-the-art performance on various tasks and benchmarks (Chang et al., 2023). Early results in NLI demonstrate the potential to overcome existing limitations, including scalability to more languages, generation of explanations, identification of direct translations, and cross-domain adaptability (Zhang and Salle, 2023). While LLMs provide new capabilities in this field, they also introduce significant challenges related to robustness, as well as increased risks of hallucinations and biases.

The main contributions and findings of this study are threefold: (1) We create a new dataset called AA-NLI50 which includes both the author's native language and authorship information; (2) We propose a zero-shot chain-of-thought (CoT) approach to mitigate hallucinations found in previous NLI studies; (3) We show that predicting the author's native language significantly improves attribution performance in multilingual contexts.

2 Related Work

In the NLI shared tasks of 2013 and 2017, the best performing approaches primarily used linguistic features together with traditional machine learning algorithms (Tetreault et al., 2013; Malmasi et al., 2017). Various feature types were investigated, including spelling errors, word and lemma n-grams, character n-grams, dependency parsing and morphosyntax (Malmasi and Dras, 2018). The combination of these diverse features was shown to be highly effective in achieving the best results for NLI (Markov et al., 2022). More recently, the deep generative approach was introduced, involving the fine-tuning of a GPT-2 model to identify each L1, achieving state-of-the-art performance (Lotfi et al., 2020). However, this approach was found to be resource-intensive with considerable performance bottlenecks (Uluslu and Schneider, 2022). The replication attempts highlighted generalization issues across various domains, suggesting potential overfitting to the benchmark dataset (Vian, 2023). Most recent studies found that newer LLMs, such as GPT-4 (Achiam et al., 2023), achieve state-of-the-art performance in zero-shot settings using prompt-based approaches, which represents a significant advancement over previous methods (Zhang and Salle, 2023; Goswami et al., 2024).

Authorship profiling has been explored as a valuable tool for authorship attribution where it assists in narrowing down the pool of potential candidates by filtering based on characteristics such as gender, age, and educational background (Yang and Chow, 2014; Deutsch and Paraboni, 2023). Psychological profiling was also shown to be effective in differentiating between authors in various contexts (Boyd, 2018). The impact of the author's native language has not yet been explored due to the scarcity of data and specific use cases. The significance of native language in cybercrime investigations cannot be overstated, as evidenced by its repeated utility in forensic analyses (Perkins, 2021).

3 Data

Due to the absence of available authorship attribution datasets that include the native language of the author, we scraped a new dataset from the social media platform Reddit, following the methodologies of Murauer and Specht (2019, 2021); Goldin et al. (2018). The dataset features English posts from authors in five different L1: French (FR), Dutch (NL), Turkish (TR), Russian (RU), and German (DE). We included posts that were assigned the topic *politics*, most of which discuss recent migration and economic issues in Europe. We ensured a minimum of 10 authors for each L1, with each author contributing at least 20 documents. A document is defined as a concatenation of individual posts until the minimum document length is reached. Following the pre-processing steps of Murauer and Specht (2019), we required each document to have a minimum length of 4,000 characters and replaced URLs and user tags with special tokens. The final corpus consists of 1,000 documents in total.

4 Methodology

We build upon the work presented by Deutsch and Paraboni (2023) by incorporating profiling systems to enhance closed-set authorship attribution. This approach utilizes an ensemble architecture comprising word and character-level n-gram models (Custódio and Paraboni, 2019). The output probabilities from the word and character-level n-gram models, combined with the one-hot encoded native language prediction from the LLM, are fed into a second-level logistic regression classifier to identify the author of the input documents. The entire pipeline is illustrated in Figure 1. To select the most suitable language model, we conducted preliminary experiments using the TOEFL11 (Blanchard et al., 2013) dataset, the de facto benchmark for NLI. The results presented in Appendix A.1 show that Llama3 (AI@Meta, 2024) performs better than Mixtral 7Bx8 (Jiang et al., 2024) but is slightly outperformed by GPT-4 (Achiam et al., 2023). Due to the confidential nature of forensic work, we only consider open-source models and utilize *llama3-70b-8192* for our experiments. The discrepancies between GPT-4 and Llama3 were primarily observed in the problematic pair within the benchmark (Hindi-Telugu). Early results in NLI revealed various types of hallucinations (Zhang and Salle, 2023), likely due to the cultural and contextual cues

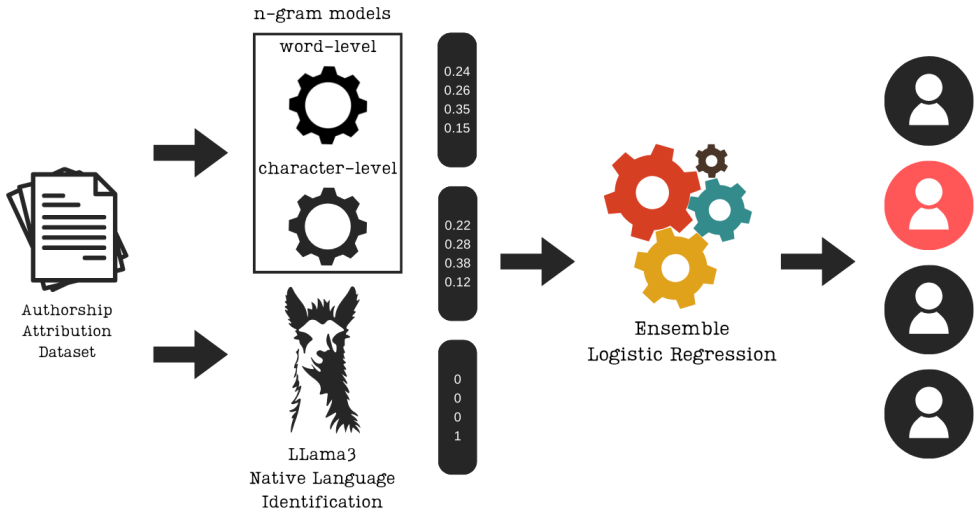

Figure 1: System architecture demonstrating the integration of native language profiling with word and character-level n-grams for authorship attribution.

present in the benchmark. Learner corpora often contain biographical information about the author, which the LLM leverages when trying to identify the native language. Using self-reported information raises the question of whether this constitutes cheating the task (Balloccu et al., 2024), as it can result in cultural analysis rather than understanding the linguistic patterns. This concern aligns with previous findings that model-generated explanations are unreliable indicators of the model's reasoning process (Madsen et al., 2024). The model can hallucinate in order to maintain self-consistency, even in the absence of linguistic cues. To address this issue, we propose a CoT approach, which does not eliminate hallucinations per se, but was shown to restrict the model behavior to the instructions (Dhuliawala et al., 2023). We redacted various name entity information from the text, including geopolitical entities (GPE), locations (LOC), and nationalities or religious or political groups (NORP), which can reveal the author's origins. To enforce a structured analysis and delay the prediction until all instructions are followed, we introduce XML formatting and prefill the model's response. The full system prompt used in our study can be found in Appendix B.1.

5 Results

We first present the results of NLI as an independent task. Following this, we integrate the most realistic setup into the authorship attribution pipeline to derive the final results. This two-step approach ensures that our evaluation captures both the isolated and integrated performance of the models.

5.1 Native Language Identification

We conducted four experiments to evaluate the zero-shot performance of LLMs on NLI. Using the entire corpus, we reported the results in terms of precision, recall, and F1 score. For comparison, we implement the open-set classification prompt from Zhang and Salle (2023) as well as our CoT approach on the dataset. We report results after redacting entity information in both experiments to assess the model dependency. Previous research has reported near-perfect accuracy on benchmarks for most language pairs under zero-shot settings. We argue that these results should be interpreted cautiously, as state-of-the-art approaches typically require approximately 10,000 examples to achieve similar performance levels and often encounter generalization issues across different datasets. While LLMs may exhibit an emergent ability for NLI, it is also possible that they have merely adapted to the datasets, finding shortcuts due to the task's inherent complexity (Schaeffer et al., 2024). An example of such hallucinations can be observed in B.2.

Approach	P	R	F1
Baseline	0.68	0.68	0.69
- Redacted ↓	0.49	0.46	0.48
CoT Prompt (Ours)	0.54	0.53	0.54
- Redacted ↓	0.46	0.46	0.47

Table 1: Performance comparison of open-set classification and CoT approaches before and after redaction.

As shown in Table 1, the open-set classification prompt achieved a high performance of 69% un-

der zero-shot settings on a more complex dataset compared to the benchmark. However, redacting entity information resulted in a substantial performance decrease of 20%. In comparison, the CoT approach requires the model to document its findings before making a final prediction, relying more on the model's ability to identify linguistic features. While the CoT prompt also experienced a performance drop due to the removal of entity information, the decrease was not as significant as with the original prompt. Both approaches converged to a similar level of performance in the follow-up experiments.

# Authors	Accuracy	
	Baseline	**+ Native Language**
5	0.65	0.66
10	0.46	**0.55**
15	0.37	**0.49**
20	0.33	**0.43**
25	0.27	**0.42**
30	0.25	**0.34**
35	0.22	**0.33**
40	0.20	**0.29**
45	0.18	**0.28**
50	0.17	**0.27**
Mean	0.32	**0.41**

Table 2: Authorship attribution mean accuracy and SD results for the standalone and integrated pipeline.

5.2 Authorship Attribution

Following the evaluation methodology of Deutsch and Paraboni (2023), we completed multiple evaluation experiments to assess the system's performance under varying conditions. We employed a zero-shot classification system, eliminating the need to split the dataset between attribution and profiling tasks. We created a balanced testing set comprising 20% of the entire dataset (200 documents), including 50 authors and five different L1s. We conducted the experiments using the CoT approach, as it offers a more realistic performance given the absence of self-reported information in forensic contexts. To evaluate the system, we sampled random sets of candidate authors from the 50-author test set, varying the number of candidate authors from 5 to 50. To minimize the effects of random selection, each evaluation was repeated 20 times. For each iteration, we varied the candidate authors randomly and selected different training and testing documents. This repetition aimed to provide more reliable and robust results by averaging out the variability introduced by random selection. The results of the authorship attribution experiments are reported in terms of accuracy, as shown in Table 2. The table presents the mean accuracy scores obtained by the open-set classification baseline and the integration of NLI into the stack ensemble. The best results for each candidate set are highlighted in bold. Based on McNemar's test, the differences in performance between the baseline model and the proposed model were found to be statistically significant ($p < 0.05$) after 5 authors. The results indicate that incorporating native language outperforms the baseline as the number of candidate authors increase. Overall, we achieved a 9% increase in mean accuracy, with the baseline at 33% and the enhanced ensemble model at 41%.

6 Conclusion

Our study demonstrates that integrating native language into authorship attribution systems significantly enhances attribution accuracy, which is particularly important in multilingual contexts such as cybercrime investigations (Perkins, 2021). This improvement aligns with the gains observed from other profiling attributes like age, gender, and education (Deutsch and Paraboni, 2023). Our study highlights the shortcuts taken by LLMs in profiling tasks, with a particular focus on how certain background information in the text (e.g. ethnicity) can lead to superficial analysis and hallucinations. Therefore, we argue that model generations should not be considered true explanations of the reasoning process. We found that employing CoT prompts can mitigate this tendency by encouraging systematic documentation of relevant linguistic features. While these findings offer promising advancements, they also underscore the need for cautious interpretation of LLM outputs in forensic sciences. Future research should focus on developing more robust profiling techniques that account for diverse linguistic factors, including the effects of register, genre, and topic. As LLMs continue to play increasingly important roles in authorship analysis, our work emphasizes the ongoing need to investigate their behaviors and limitations.

Acknowledgments

This work was supported by the collaboration between the University of Zurich and PRODAFT as part of the Innosuisse innovation project AUCH 103.188 IP-ICT (Author profiling to automatize attribution in cybercrime investigations).

References

OpenAI Josh Achiam, Steven Adler, and Sandhini Agarwal. 2023. GPT-4 Technical Report.

AI@Meta. 2024. Llama 3 model card.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondřej Dušek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source llms. *arXiv preprint arXiv:2402.03927*.

Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. Toefl11: A corpus of non-native english. *ETS Research Report Series*, 2013(2):i–15.

Ryan L Boyd. 2018. Mental profile mapping: A psychological single-candidate authorship attribution method. *PloS one*, 13(7):e0200588.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2023. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*.

José Eleandro Custódio and Ivandré Paraboni. 2019. An ensemble approach to cross-domain authorship attribution. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 201–212. Springer.

Caio Deutsch and Ivandré Paraboni. 2023. Authorship attribution using author profiling classifiers. *Natural Language Engineering*, 29(1):110–137.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2023. Chain-of-verification reduces hallucination in large language models. *arXiv preprint arXiv:2309.11495*.

Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 3591–3601.

Dhiman Goswami, Sharanya Thilagan, Kai North, Shervin Malmasi, and Marcos Zampieri. 2024. Native language identification in texts: A survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3149–3160.

Sylviane Granger, Estelle Dagneaux, Fanny Meunier, Magali Paquot, et al. 2009. *International corpus of learner English*, volume 2. Presses universitaires de Louvain Louvain-la-Neuve.

Jack Grieve. 2023. Register variation explains stylometric authorship analysis. *Corpus Linguistics and Linguistic Theory*, 19(1):47–77.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Youngjin Jin, Eugene Jang, Yongjae Lee, Seungwon Shin, and Jin-Woo Chung. 2022. Shedding new light on the language of the dark web. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5621–5637, Seattle, United States. Association for Computational Linguistics.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *Journal of the American Society for information Science and Technology*, 60(1):9–26.

Sergey Kramp, Giovanni Cassani, and Chris Emmery. 2023. Native language identification with big bird embeddings. *arXiv preprint arXiv:2309.06923*.

Ehsan Lotfi, Ilia Markov, and Walter Daelemans. 2020. A deep generative approach to native language identification. In *Proceedings of the 28th international conference on computational linguistics*, pages 1778–1783.

Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Literary and linguistic Computing*, 26(1):35–55.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful?

Shervin Malmasi and Mark Dras. 2017. Multilingual native language identification. *Natural Language Engineering*, 23(2):163–215.

Shervin Malmasi and Mark Dras. 2018. Native language identification with classifier stacking and ensembles. *Computational Linguistics*, 44(3):403–446.

Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel Tetreault, Robert Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 62–75.

Ilia Markov, Vivi Nastase, and Carlo Strapparava. 2022. Exploiting native language interference for native language identification. *Natural Language Engineering*, 28(2):167–197.

Benjamin Murauer and Günther Specht. 2019. Generating cross-domain text classification corpora from social media comments. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10*, pages 114–125. Springer.

Benjamin Murauer and Günther Specht. 2021. Dtgrams: Structured dependency grammar stylometry for cross-language authorship attribution. *arXiv preprint arXiv:2106.05677*.

Ria C Perkins. 2021. The application of forensic linguistics in cybercrime investigations. *Policing: A Journal of Policy and Practice*, 15(1):68–78.

Anderson Rocha, Walter J Scheirer, Christopher W Forstall, Thiago Cavalcante, Antonio Theophilo, Bingyu Shen, Ariadne RB Carvalho, and Efstathios Stamatatos. 2016. Authorship attribution for social media forensics. *IEEE transactions on information forensics and security*, 12(1):5–33.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2024. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Stian Steinbakken and Björn Gambäck. 2020. Native-language identification with attention. In *Proceedings of the 17th international conference on natural language processing (icon)*, pages 261–271.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the eighth workshop on innovative use of NLP for building educational applications*, pages 48–57.

Ahmet Yavuz Uluslu and Gerold Schneider. 2022. Scaling native language identification with transformer adapters. In *Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022)*, pages 298–302.

Ahmet Yavuz Uluslu and Gerold Schneider. 2023. Turkish native language identification. *Natural Language and Speech Processing (ICNLSP-2023)*, page 303.

Matias Johansen Vian. 2023. A study of transformers for cross-corpus native language identification. Master's thesis, NTNU.

Xuehu Wei, Helyne Adamson, Matthias Schwendemann, Tomás Goucha, Angela D Friederici, and Alfred Anwander. 2023. Native language differences in the structural connectome of the human brain. *Neuroimage*, 270:119955.

Min Yang and Kam-Pui Chow. 2014. Authorship attribution for forensic investigation with thousands of authors. In *ICT Systems Security and Privacy Protection: 29th IFIP TC 11 International Conference, SEC 2014, Marrakech, Morocco, June 2-4, 2014. Proceedings 29*, pages 339–350. Springer.

Liming Yu and Terence Odlin. 2016. *New perspectives on transfer in second language learning*, volume 92. Multilingual Matters.

Wei Zhang and Alexandre Salle. 2023. Native language identification with large language models. *arXiv preprint arXiv:2312.07819*.

A Appendix – Preliminary Experiments

Model	TOEFL11 Test Set (%)
Random Baseline	9.1
GPT-2 (Lotfi et al., 2020)	89.0
GPT-3.5 (Zero-shot)	74.0
Mixtral 8x7B (Zero-shot)	74.0
LLama3 70B (Zero-shot)	85.4
GPT-4 (Zero-shot)	91.5

Table A.1: Performance comparison of various models on the TOEFL11 test set.

B Appendix – Supplementary Material

B.1 System Prompt

You are a forensic linguistics expert responsible for analyzing texts written by non-native speakers. Identify linguistic cues such as direct translations, spelling errors, syntactic patterns, and grammatical errors to identify the native language of the author. It is important to note that the self-reported information or cultural references provided in the text can be misleading.

<transcript> {input_text} </transcript>

Think step by step on how to analyze the <transcript> within the provided <sketchpad>.

In the <sketchpad>, return a list of <findings> and their corresponding <types>.

Then, check that <sketchpad> items are factually consistent with the <transcript>.

Finally, identify the native language of the author based on the <sketchpad>.

Figure B.1: System Prompt

B.2 LLM-generated Authorship Profiling Outputs

Text	Prediction
If the state wants to implement something bad, they protest like mad men, until the state listens to them. The Gezi protests failed, because we didn't go hard enough. <text continues>	Turkish <hallucination analysis>
If the state wants to implement something bad, they protest like mad men, until the state listens to them. The Moscow protests failed, because we didn't go hard enough. <text continues>	Russian <hallucination analysis>
If the state wants to implement something bad, they protest like mad men, until the state listens to them. The <redacted> protests failed, because we didn't go hard enough. <text continues>	Random or No prediction <complication message> <random message>

Figure B.2: LLM-generated outputs for the NLI task based on Reddit posts under different conditions.