Causal Language Model Perplexity for Human Authorship Attribution

Anonymous ACL submission

Abstract

In this paper, we introduce an authorship attribution method that identifies the most likely author of a questioned document based on the perplexity of the questioned document calculated for a set of GPT-2 models fine-tuned on the writings of each candidate author. We evaluate our method on corpora representing the writings of 50 fiction authors. We find that the perplexity of causal large language models is able to distinguish among these 50 authors with an overall f-score of 0.99 and a macro average accuracy of 0.99, considerably outperforming other state-of-the-art methods applied to other datasets with similar numbers of authors. We also test how the performance of our method depends on the length of the questioned document and the amount of training data for each author. We find that to reach a 0.90 f-score with 50 possible authors via our method, the minimum training data required is 28,000 tokens, while the minimum testing data required is 70 tokens.

1 Introduction

For centuries, researchers have developed methods for *authorship attribution* to resolve cases of disputed authorship by comparing the style of a questioned document to writing samples from a set of candidate authors (Juola, 2006; Stamatatos, 2009). The goal of authorship attribution is to identify the candidate whose style of writing is most similar to a questioned document. Stylometry is the quantitative analysis of style and is a common approach to authorship attribution (Juola, 2006; Stamatatos, 2009). A wide range of different measurements and methods for authorship attribution have been developed in stylometry (Grieve, 2007; Stamatatos, 2009). The most popular techniques include Principal Component Analys of function word frequencies (Binongo, 2003; Grieve, 2023) and distance-based comparisons of the frequencies of common words (Argamon, 2007; Burrows, 2002). Although stylometric approaches are very useful for resolving certain types of authorship attribution tasks, there are clear limitations with these techniques: the overall performance of these methods drastically declines when the number of candidate authors increases (Grieve, 2007; Luyckx and Daelemans, 2011), when the length of the question document decreases (Eder, 2015; Grieve et al., 2018), and when the amount of training data from the candidate authors decreases (Luyckx and Daelemans, 2011).

The power of modern large language models (LLMs), however, has the potential to address these issues. Examples of such approaches include building universal authorial embedding via Siamese BERT (Rivera-Soto et al., 2021) or Character BERT (El Boukkouri et al., 2020), and using BERT for classification (Fabien et al., 2020; Tyo et al., 2022). In addition, in response to increasing concerns about the misuse of LLMs (Bommasani et al., 2022; Gehrmann et al., 2019; Tian et al., 2023; Wu et al., 2023; Gehrmann et al., 2019; Wu et al., 2023), the task of LLM identification has gained prominence, with causal language model perplexity being proposed as a potential indicator, encompassing applications in fully automated detection and computer-assisted methods such as GLTR (Gehrmann et al., 2019) and GPTZero (Chakraborty et al., 2023). Although researchers have attempted authorship attribution via LLM perplexity for PoS-tags (Fourkioti et al., 2019) and pre-trained BERT perplexity (i.e. pALM; Barlas and Stamatatos, 2020), the performance of these systems is poor (Tyo et al., 2022).

In this paper, we address human authorship attribution via LLM perplexity by introducing the concept of *authorial causal language models*. Our approach involves fine-tuning a series of authorial GPT-2 models based on the known writing of a series of candidate authors, creating one model for each author. We then calculate the perplexity of the questioned document given each of these *authorial LLMs*. Finally, we attribute the questioned document to the author whose authorial LLM has the lowest perplexity for the questioned document. We show that the method achieves remarkably high performance, while addressing various limitations with stylometric methods for authorship attribution.

2 Methodology

2.1 Data

To evaluate our method, we require a large number of texts written in a relatively consistent genre/register by a large number of authors. Furthermore, as our method relies on fine-tuning of authorial GPT-2, we prefer a dataset containing as many texts as possible. Considering existing datasets, we found that CCAT50 (Lewis et al., 2004), Blogs50 (Schler et al., 2006), and GutenbergAA (Tyo et al., 2022) contain a sufficient number of authors, but they are either too small or imbalanced in genre/register, text count, or token count by author. We therefore compiled our own Gutenberg English Fiction Authorship corpus (GEFA) to evaluate our method.

To compile GEFA, we first extracted texts from Project Gutenberg[1], controlling for genre/register and the time of publication. We collected the content of books tagged as fiction from 1830 to 1920 from 50 authors. We chose to work with 50 authors because distinguishing between such a large number of authors is generally considered to be challenging (Grieve, 2007) and because this is comparable with existing datasets used to evaluate methods for authorship attribution in NLP (Lewis et al., 2004; Schler et al., 2006; Tyo et al., 2022). We then cleaned the texts (e.g., removing Project Gutenberg specific labels) and split each document into texts of at least 512 tokens. This decision was guided by the general difficulty in attributing shorter texts in stylometry, where a minimum text length of 500 words is commonly used (Grieve, 2007) and even stricter criteria are often recommended (Grieve, 2007; Eder, 2015). This procedure led to texts with similar token counts, which we labelled as *GEFA Unbalanced*, because the number of texts per author has not been controlled. Next, we randomly sampled an equal number of texts from each author, equal to the smallest number of texts in any author's corpus. We then split the data into training (80%) and test (20%) sets, which we labeled as *GEFA full*. Based on this full dataset, we also produced a series of downsampled datasets to test our method as the amount of training data for each author decreases. Downsampled GEFA versions are labelled GEFA-X where X is the percentage of texts in GEFA full. To ensure the reproducibility of our results, all GEFA versions are made accessible here[2].

2.2 Authorial GPT Fine-Tuning

In our study, we chose GPT-2 small as the base model to minimize training costs. Furhermore, preliminary research has suggested that the link between perplexity accuracy and model size is limited (Radford et al., 2019). With fine-tuning epoch counts set to 100, we fine-tuned 50 authorial GPT-2 based on the texts of 50 authors in the training set, and we labeled fine-tuned GPT-2s with the corresponding author names. We conducted fine-tuning on Graphcore IPU Pod 4 Machine at PaperSpace. We make all scripts accessible online.[2].

2.3 Perplexity

Perplexity of a fixed-length causal language model M over a token sequence $T = \{x_1, x_2, ..., x_t\}$ is

$$ppl\left(M, T\right) = exp\left\{-\frac{1}{t}\sum_{i}^{t} p_M\left(x_i|x_{<i}\right)\right\}$$

In practice, we calculate perplexity as the cross entropy between the true token and predicted logits, namely $exp\left\{CrossEntropy\left(Logits, T\right)\right\}$. Given a questioned test text Q, and a pre-trained authorial GPT-2 model M, we first pass Q to GPT-2 BPE Tokenizer to form a true token sequence T. This token sequence is then passed to M for language modeling, where $Logits$ is in the model outputs. Then we calculated cross entropy of T and $Logits$ in $torch.nn.CrossEntropyLoss$ of PyTorch. Finally, we obtain the perplexity of Q under M as exponentiated cross entropy with base e.

2.4 Authorship Prediction

As perplexity measures how well an LLM predicts a text, we apply this concept to human authorship attribution as follows: given a text Q from author i, and a set of authorial GPT-2 models $\{M_1, M_2, ...M_n\}$ fine-tuned on texts from a set of candidate authors $\{1, 2, ..., n\}$,

[1] https://www.gutenberg.org

[2] https://anonymous.4open.science/r/CLMPPL-ACLARR2024

Dataset	A	T	TK	T/A	TTL
GEFA(full)	50	43k	28M	856	665
GEFA-5	50	2k	1.5M	42	665
CCAT50	50	5k	2.5M	100	506
Blogs50	50	66k	8.1M	1.3k	122
GAA	4.5k	29k	1.9B	6	66k

A: author count; T: text count; TK: token count; TTL: test text length, in token count

Table 1: Facts on Datasets

Method	**Dataset**	**Acc.**	**Dataset**	**Acc.**
CLMppl	**GEFA**	**99.8**	**GEFA5**	**92.6**
Ngram	CCAT50	76.7	Blogs50	72.3
BERT	CCAT50	65.7	Blogs50	75.0
	GAA	59.1		
PPM	CCAT50	69.4	Blogs50	72.2
	GAA	57.7		
pALM	CCAT50	63.4		

Table 2: Comparison Between Our Method (CLMppl) And Recent SOTA Studies (Tyo et al., 2022)

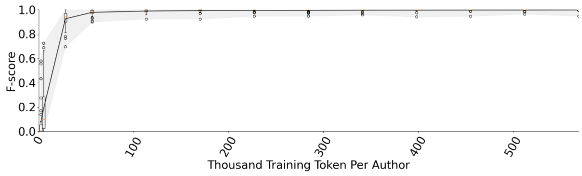

Figure 1: Training Data Token Count and Our Method

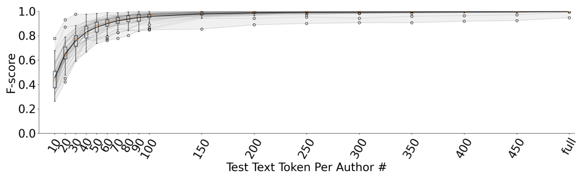

Figure 2: Test Text Length and Our Method

we expect $ppl\left(M_i, Q\right)$ to be lowest among $\{ppl\left(M_1, Q\right), ppl\left(M_2, Q\right), ..., ppl\left(M_n, Q\right)\}$. This is because we expect the text Q to be most predictable for the model that was fine-tuned on the training corpus for author i. To test these assumptions, we evaluated our method on the full GEFA dataset as well as a range of GEFA downsampled datasets.

3 Results

We evaluated the performance of our method on GEFA using f-score and accuracy. On GEFA full, our method achieves a near-perfect 0.998 on both criteria, and our method still achieves an f-score of 0.926 on the downsampled GEFA-5 training set, which consists of only 5% of the data in GEFA full. Both results are excellent compared to both stylometric approaches and recent SOTA studies in NLP, as presented in Table 1 and Table 2. The table shows that when compared against recent SOTA studies with similar author count, and on the same benchmark of macro-average accuracy, our method outperforms other studies. This is even true when we evaluate on GEFA-5: with training data as small as 42 texts per author, or approximately 28k tokens per authors, our method still achieves the best macro-average accuracy by a considerable amount.

Furthermore, to evaluate the robustness of our method across authors, we calculated single author f-scores. Table 3 in the Appendix shows that we achieve perfect performance for most authors: 82% of authors have a perfect f-score of 1, while 98% have f-scores over 0.99. Only one author is below this threshold (Alexander, Mrs.); however, an f-score of 0.95 was still obtained.

In addition to evaluating our method on *GEFA-full* and *GEFA-5*, we evaluated our method on various GEFA downsampled datasets to test how its performance is affected when the training token count is varied. We plot f-scores across GEFA downsampled corpora in Figure 1, which shows how performance changes as training data token count increases. We also evaluated our method using different test text length by calculating perplexity for truncated test texts. We plot f-scores for different text lengths in Figure 2, which shows how performance changes as test text length increases. As expected, we found that increasing the amount of training data or the length of the test texts resulted in higher f-scores and smaller inter-quartile ranges. However, both f-scores and inter-quartile ranges stabilize as the median f-score hits 0.90, which requires at least 28,000 tokens in training data or 70 tokens in test texts. This result demonstrates that our method is still highly accurate on 50 authors with limited training data or short test texts.

4 Discussion

We attribute the excellent performances of our method to the fact that perplexity provides access to the token-level authorial features that are inaccessible in type-based methods. Compared with standard type-based methods in stylometry, which are based on the relative frequencies of common words, n-grams, and other forms, our method is capable of capturing authorial information for each word token rather than each word type, thereby offering greater flexibility and granularity.

For instance, the use of the word *baseball* is not generally a good feature for stylometric authorship attribution for two reasons. First, it is relatively infrequent, making it difficult to obtain a meaningful measurement of relative frequency of this word type. This is why stylometric methods tend to focus on high frequency forms. Second, even if sufficient data were available, the relative frequency of this word type in any text or corpus would primarily reflect the topic of that text or corpus, as opposed to the style of the author. For example, a text with the frequent use of the word *baseball* will tend to be about this sport. This is problematic in the context of authorship attribution because the goal is to attribute texts to the correct authors regardless of the topical content of the text. This is why stylometric methods tend to focus not only on common features, but on grammatical features, like function word frequency.

Our token-based approach, however, avoids these issues as it assigns a perplexity score to every token in a text. In general, we can assume that if a questioned document is about baseball, occurrences of the word *baseball* will generally carry very little authorial information, and that the perplexity of the tokens of the word *baseball* in that text will consistently be low for all authors. However, given a questioned document on some other topic, a token of the term *baseball* (e.g., as an example or as a metaphor) would potentially be highly discriminatory – extremely unexpected for most authors, unless, for example, an author often uses baseball metaphors out of context.

In this sense, our method is similar to the type of qualitative stylistic authorship analysis often conducted manually in a forensic context, where forensic linguists examine a questioned document word by word (Coulthard et al., 2016; Grant, 2008). Like a forensic stylistic analysis, a great advantage of our approach compared to a standard stylometric analysis is that we can extract considerably more information from each text: every token is now a valid feature, whereas for traditional methods only frequent types can potentially be features. Our method can therefore produce highly accurate results, even when presented with large numbers of candidate authors and short questioned documents.

5 Conclusion

We developed an authorship attribution method based on the perplexity of fine-tuned causal language models, achieving a near perfect f-score of 0.99, outperforming existing authorship attribution methods tested on datasets of similar dimensions. In addition, we found that to reach an f-score of 0.90 on our evaluation corpus, our method requires at least 28,000 tokens of training data or test texts consisting of at least 70 tokens. Future research may focus on few-shot authorship attribution via perplexity of in-context-learning-capable LLM like Llama-2 (Touvron et al., 2023) and authorship profiling with filtered perplexity. In addition, we believe that our perplexity-based approach constitutes a general method for comparative research in corpus linguistics, allowing for automated token-level comparative linguistic analysis.

6 Limitation

Currently, our method has only been tested on different versions of GEFA. Though GEFA is made openly accessible to ensure replicability of this research, it is still important to evaluate our method on existing datasets to test the robustness and comparability of our current results. It is also important to evaluate our method when the amount of training data and test text length are varied simultaneously, especially to assess performance when both of these factors are limited. Finally, additional examination of subtle biases in GEFA is important, despite our best efforts to balance the corpus.

7 Ethics and Impact

Our research is based on publicly available base model and, and we are not aware of specific risks except biases inherited from data or base model, which needs to be examined before any implementations in a large scale. Moreover, when put in practice, the predicted author from this method should be treated as reference to form the final decision of authorship together with other clues and evidences.

References

Shlomo Argamon. 2007. Interpreting burrows's delta: Geometric and probabilistic foundations. *Literary and Linguistic Computing*, 23(2):131–147.

Georgios Barlas and Efstathios Stamatatos. 2020. *Cross-Domain Authorship Attribution Using Pre-trained*

Language Models, volume 583 of *IFIP Advances in Information and Communication Technology*, page 255–266. Springer International Publishing, Cham.

José Nilo G. Binongo. 2003. Who wrote the 15th book of oz? an application of multivariate analysis to authorship attribution. *Chance*, 16(2):9–17.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, et al. 2022. On the opportunities and risks of foundation models. (arXiv:2108.07258). ArXiv:2108.07258 [cs].

John Burrows. 2002. "delta": a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

Souradip Chakraborty, Amrit Singh Bedi, Sicheng Zhu, Bang An, Dinesh Manocha, and Furong Huang. 2023. On the Possibilities of AI-Generated Text Detection. ArXiv:2304.04736 [cs].

Malcolm Coulthard, Alison Johnson, and David Wright. 2016. *An introduction to forensic linguistics: Language in evidence*. Routledge.

Maciej Eder. 2015. Does size matter? authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2):167–182.

Hicham El Boukkouri, Olivier Ferret, Thomas Lavergne, Hiroshi Noji, Pierre Zweigenbaum, and Jun'ichi Tsujii. 2020. Characterbert: Reconciling elmo and bert for word-level open-vocabulary representations from characters. In *Proceedings of the 28th International Conference on Computational Linguistics*, page 6903–6915, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. 2020. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, page 127–137.

Olga Fourkioti, Symeon Symeonidis, and Avi Arampatzis. 2019. Language models and fusion for authorship attribution. *Information Processing & Management*, 56(6):102061.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, page 111–116, Florence, Italy. Association for Computational Linguistics.

Tim Grant. 2008. Approaching questions in forensic authorship analysis. *Dimensions of forensic linguistics*, 5:215–229.

Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3):251–270.

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

Jack Grieve, Isobelle Clarke, Emily Chiang, Hannah Gideon, Annina Heini, Andrea Nini, and Emily Waibel. 2018. Attributing the Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities*, 34(3):493–512.

Patrick Juola. 2006. *Authorship attribution*. Foundations and trends in information retrieval. Now Publ, Boston, Mass.

David D Lewis, Yiming Yang, Tony G Rose, and Fan Li. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of machine learning research*, 5:361–397.

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.

Alec Radford, Jong Wook Kim, and Jeff Wu. 2019. Gpt2 output dataset.

Rafael A. Rivera-Soto, Olivia Elizabeth Miano, Juanita Ordonez, Barry Y. Chen, Aleem Khan, Marcus Bishop, and Nicholas Andrews. 2021. Learning universal authorship representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, page 913–919, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, page 199–205.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.

Yuchuan Tian, Hanting Chen, Xutao Wang, Zheyuan Bai, Qinghua Zhang, Ruifeng Li, Chao Xu, and Yunhe Wang. 2023. Multiscale positive-unlabeled detection of ai-generated texts. (arXiv:2305.18149). ArXiv:2305.18149 [cs].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Jacob Tyo, Bhuwan Dhingra, and Zachary C. Lipton. 2022. On the state of the art in authorship attribution and authorship verification. (arXiv:2209.06869). ArXiv:2209.06869 [cs].

Kangxi Wu, Liang Pang, Huawei Shen, Xueqi Cheng, and Tat-Seng Chua. 2023. Llmdet: A large language models detection tool. (arXiv:2305.15004). ArXiv:2305.15004 [cs].

Author Name	Text #	Token #	Mean Token # per Text	F1
Abbott, Jacob	856	556134	649.69	1.00
Allen, James Lane	856	551245	643.98	1.00
Anderson, Sherwood	856	524301	612.5	1.00
Brame, Charlotte M.	856	560221	654.46	1.00
Bridges, Victor	856	553755	646.91	1.00
Bullen, Frank Thomas	856	561322	655.75	1.00
Coleridge, Christabel R. (Christabel Rose)	856	568590	664.24	1.00
Crane, Stephen	856	568984	664.7	1.00
Cummings, Ray	856	565409	660.52	1.00
Disraeli, Benjamin, Earl of Beaconsfield	856	580202	677.81	1.00
Farrar, F. W. (Frederic William)	856	575684	672.53	1.00
Ferber, Edna	856	590245	689.54	1.00
Forster, E. M. (Edward Morgan)	856	571506	667.65	1.00
Frey, Hildegard G.	856	558387	652.32	1.00
Hains, T. Jenkins (Thornton Jenkins)	856	554547	647.84	1.00
Harrison, Henry Sydnor	856	571045	667.11	1.00
Hendryx, James B. (James Beardsley)	856	579162	676.59	1.00
Hudson, W. H. (William Henry)	856	553653	646.79	1.00
Jefferies, Richard	856	566789	662.14	1.00
Kingsley, Charles	856	589668	688.86	1.00
Knox, Thomas Wallace	856	558877	652.89	1.00
Major, Charles	856	557868	651.71	1.00
Marchant, Bessie	856	546549	638.49	1.00
May, Sophie	856	587586	686.43	1.00
McElroy, John	856	571328	667.44	1.00
McKenna, Stephen	856	562750	657.42	1.00
Moodie, Susanna	856	565542	660.68	1.00
Mühlbach, Luise	856	576502	673.48	1.00
Ray, Anna Chapin	856	573470	669.94	1.00
Saintsbury, George	856	593408	693.23	1.00
Sayler, H. L. (Harry Lincoln)	856	563583	658.39	1.00
Scott, John Reed	856	585912	684.48	1.00
Senarens, Luis	856	578822	676.19	1.00
Seton, Ernest Thompson	856	567763	663.27	1.00
Smedley, Frank E. (Frank Edward)	856	581661	679.51	1.00
Stephens, Ann S. (Ann Sophia)	856	571313	667.42	1.00
Taylor, Meadows	856	581298	679.09	1.00
Tuttle, W. C. (Wilbur C.)	856	585312	683.78	1.00
Wallace, Lew	856	582231	680.18	1.00
Warner, Anna Bartlett	856	570210	666.13	1.00
Yates, Dornford	856	585614	684.13	1.00
Arthur, T. S. (Timothy Shay)	856	559509	653.63	0.99
Broughton, Rhoda	856	577235	674.34	0.99
Eliot, George	856	568850	664.54	0.99
Ewing, Juliana Horatia	856	570841	666.87	0.99
Freeman, R. Austin (Richard Austin)	856	564832	659.85	0.99
Morris, William	856	574831	671.53	0.99
Morrison, Arthur	856	575161	671.92	0.99
Sinclair, Bertrand W.	856	558971	653	0.99
Alexander, Mrs.	856	570113	666.02	0.95

Table 3: Performance of Our Method on Each of the 50 Authors