

AUTHORSHIP VERIFICATION OF ENGLISH TEXT: TRADITIONAL STYLOMETRY, TRANSFORMERS, AND LARGE LANGUAGE MODELS COMPARED AND DEPLOYED

Jiří Pokorný¹, Michaela Veselá¹, Oldřich Trenz¹

¹Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/ecos2026_027

ABSTRACT

This article compares three families of methods for authorship verification, the binary decision whether two texts share one author. We frame the task as pairwise classification and test traditional stylometric pipelines (bag-of-words and TF-IDF with scikit-learn classifiers), fine-tuned transformers (Longformer, BigBird-RoBERTa, ModernBERT), and large language models used through prompting. Training used the PAN 2020 corpus, testing the disjoint PAN 2021 corpus, together 295,800 English text pairs. A fine-tuned Longformer classification head reached the highest F1 of 92.38%, but a much cheaper TF-IDF pipeline with a linear classifier trailed it by less than one percentage point at 91.50%. The way two text vectors are combined into one feature vector mattered more than the classifier itself: absolute difference beat element-wise multiplication everywhere. The three open large language models we tried produced unusable output. Because hosting was constrained, the deployed web application runs TF-IDF with logistic regression (F1 91.28%), which returns the probability the interface needs.

Keywords: authorship verification, stylometry, TF-IDF, transformers, Longformer, large language models

JEL Code: C45

1 INTRODUCTION

Authorship verification asks a narrow question with wide consequences: were these two texts written by the same person? In this article we treat the task as an instance-based pairwise binary classification problem, without constructing author profiles. The setting is relevant to forensics, law, cybersecurity, and identity-theft detection (Stamatatos, 2008; Koppel *et al.*, 2004).

The practical question behind this work is which method a developer should actually pick. Traditional stylometry counts words and syntactic patterns and feeds them to a classifier; it is cheap and well understood (Savoy, 2020). Transformer language models promise stronger text



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license. <https://creativecommons.org/licenses/by-sa/4.0/deed.en>

representations but cost far more to train and serve (Vaswani *et al.*, 2017; Devlin *et al.*, 2018). Large language models can be prompted to answer directly, with no training at all (Huang *et al.*, 2024). This work therefore compares these three families on the same PAN-based corpus and then takes the selected model one step further, into a running web application.

1.1 Related work

Stylometric verification has a long record. Koppel *et al.* (2004) cast it as one-class classification; later surveys traced the field as it moved from hand-crafted features toward learned representations (Stamatatos *et al.*, 2016; Tyo *et al.*, 2022). The PAN shared tasks have set the common benchmark for cross-domain verification, including the 2020 and 2021 editions on fanfiction text whose corpora we use here (Kestemont *et al.*, 2020; Kestemont *et al.*, 2021). Ensembles of traditional classifiers have stayed competitive: Abbasi *et al.* (2022) report strong results from ensemble learning, and the best PAN 2021 entry was itself an ensemble.

Neural approaches have grown alongside. Siamese networks learn a similarity space for author identification at scale (Saedi *et al.*, 2021), and graph-based features encode syntax explicitly through integrated syntactic graphs (Gómez-Adorno *et al.*, 2016). The most recent thread asks whether large language models can judge authorship by prompting alone; Huang *et al.* (2024) report over 88 % F1 on an attribution variant using a proprietary model. Whether smaller open models reach that bar, and whether a fine-tuned encoder beats classical stylometry by enough to justify its cost, is what this work tests.

2 METHODOLOGY

We frame verification as pairwise binary classification. Each text becomes a feature vector; the two vectors of a pair are merged into one, and a classifier labels the pair as same-author or different-author. We compared two merge operators throughout: the absolute difference of the two vectors and their element-wise multiplication. The intuition is that multiplication highlights features the texts share, while absolute difference captures where they diverge.

Four experimental directions were run on identical data. The first used bag-of-words and TF-IDF representations (unigrams) with classifiers from scikit-learn: Multinomial Naive Bayes, Logistic Regression, LinearSVC, k-Nearest Neighbors, Random Forest, and XGBoost. The second replaced those representations with integrated syntactic graphs built per text by the text2graphapi library, turning each text into a frequency vector over a node dictionary and reusing the same classifiers. The third fine-tuned a classification head on three long-context transformers: Longformer-base-4096, BigBird-RoBERTa-large, and ModernBERT-large. The fourth prompted three open large language models (Llama-2-70B-chat-GPTQ, Gemma-3-4b-it, and Qwen3-4B-Instruct-2507) with no fine-tuning, using conservative decoding (low temperature, low top-p).

Hyperparameters were tuned with Optuna, 30 trials per representation-and-classifier combination, optimizing F1. Transformer trials trained on a reduced set of 19,999 pairs and validated on 2,500, with input capped at 4096 tokens; BigBird-RoBERTa and ModernBERT can take 8192 tokens, but we kept 4096 to hold training time down. We report accuracy, precision, recall, and F1, with F1 as the deciding metric, plus normalized confusion matrices.

The corpus is the PAN fanfiction dataset: the larger PAN 2020 variant for training and the disjoint PAN 2021 set for testing (Kestemont *et al.*, 2020; Kestemont *et al.*, 2021), 295,800 English pairs in

total. The test set holds 19,999 pairs (10,000 same-author, 9,999 different-author), and its authors do not appear in training. Texts are long, with mean and median length around 21,000 characters, which is why long-context transformers were a natural fit. Experiments ran on e-INFRA CZ / MetaCentrum (CESNET): NVIDIA H100 GPUs for transformers and language models, A10 and A40 cards for the stylometric and graph runs.

The verification model also had to run, not just score well offline. The best practical model was wrapped in a two-tier web application: a React 18.3.1 frontend in TypeScript and a FastAPI 0.128.0 backend exposing a `/predict` endpoint with Pydantic validation, deployed on the Railway cloud. The architecture is shown in Figure 1.

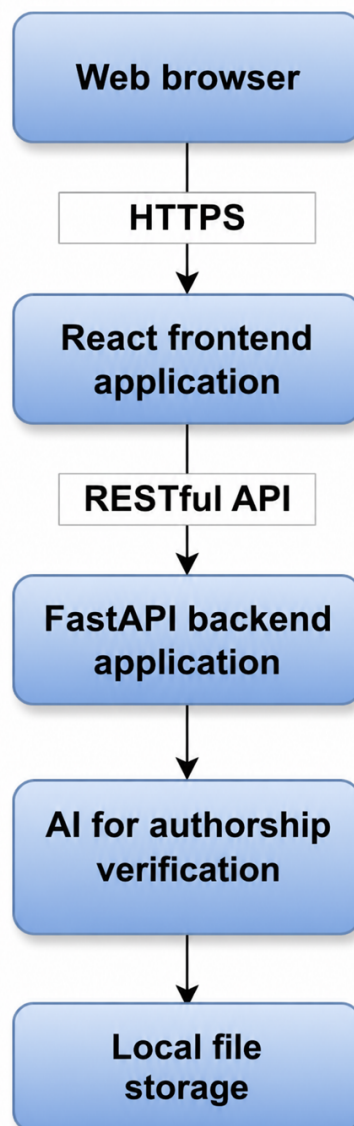


Fig. 1: System architecture of the authorship-verification web application, showing the layered flow from the web browser through the React frontend (HTTPS) and the FastAPI backend (RESTful API) to the AI authorship-verification module and the local file storage (adopted from Pokorny, 2026).

3 RESULTS

Table 1 lists the strongest run in each of the four directions, ranked by F1; the best overall row is in bold. The winner is a fine-tuned Longformer classification head at F1 92.38%. Second is the classical TF-IDF pipeline with LinearSVC at 91.50%, less than a percentage point behind. Two TF-IDF variants even posted higher accuracy (93.33% and 93.25%) than the Longformer (92.48%), though lower F1 (Pokorny, 2026).

Inside the stylometric direction (Table 2) the merge operator decided more than the classifier. With absolute difference, TF-IDF and LinearSVC reached F1 91.50%; the best multiplication run (BoW with LinearSVC) fell to 86.68%, and k-Nearest Neighbors collapsed to 63.75% with BoW and multiplication. TF-IDF beat plain bag-of-words across the board, since it weights author-specific words more heavily.

Among transformers (Table 3) Longformer led, BigBird-RoBERTa followed closely at 91.48%, and ModernBERT came last at 90.56% despite more than three times the training time of the others. The graph direction was weakest: its best result was Logistic Regression with absolute difference at F1 86.69%, and LinearSVC with multiplication dropped to 50.00% accuracy, no discriminative power at all. The three large language models all failed; their outputs were inconsistent or merely echoed the prompt, so they carry no numeric table.

Tab. 1 Best experiment per direction, ranked by F1

Representation/model	Combination	Classifier	Accuracy	Precision	Recall	F1
Longformer-base-4096	—	Classification head	92.48%	93.58%	91.22%	92.38%
TF-IDF	Absolute difference	LinearSVC	93.33%	89.39%	91.32%	91.50%
BigBird-RoBERTa-large	—	Classification head	91.64%	93.28%	89.76%	91.48%
TF-IDF	Absolute difference	Logistic Regression	93.25%	89.00%	91.08%	91.28%
ModernBERT-large	—	Classification head	90.72%	92.18%	88.99%	90.56%

Tab. 2 Best BoW / TF-IDF results, absolute-difference combination

Representation	Combination	Classifier	Accuracy	Precision	Recall	F1
TF-IDF	Absolute difference	LinearSVC	93.33%	89.39%	91.32%	91.50%
TF-IDF	Absolute difference	Logistic Regression	93.25%	89.00%	91.08%	91.28%
TF-IDF	Absolute difference	XGBoost	91.18%	88.96%	90.05%	90.17%
BoW	Absolute difference	Logistic Regression	90.21%	87.29%	88.73%	88.91%
BoW	Absolute difference	LinearSVC	90.72%	83.88%	87.17%	87.65%

Tab. 3 Best transformer results (classification head)

Representation	Classifier	Accuracy	Precision	Recall	F1
Longformer-base-4096	Classification head	92.48 %	93.58 %	91.22 %	92.38 %
BigBird-RoBERTa-large	Classification head	91.64 %	93.28 %	89.76 %	91.48 %
ModernBERT-large	Classification head	90.72 %	92.18 %	88.99 %	90.56 %
BoW	Logistic Regression	90.21%	87.29%	88.73%	88.91%
BoW	LinearSVC	90.72%	83.88%	87.17%	87.65%

		Prediction	
		Positive	Negative
Actual value	Positive	93.6 %	6.4 %
	Negative	10.6 %	89.4 %

Fig. 2: Normalised confusion matrix of the best TF-IDF model (TF-IDF representation with LinearSVC and absolute-difference vector combination), reaching 93.6 % true positives and 89.4% true negatives (adopted from Pokorny, 2026)

4 DISCUSSION

The headline number favours the transformer, but the gap is too small to be the whole story. Less than one percentage point of F1 separates a fine-tuned Longformer from a TF-IDF pipeline that needs a fraction of the compute and no GPU to serve. If the only criterion is maximum F1 and sufficient GPU resources are available, the Longformer is the strongest tested option. For a service that must fit a modest cloud plan, the classical TF-IDF pipeline offers a better cost-performance trade-off without giving up much accuracy. That is why the deployed model is neither of the two top scorers but TF-IDF with logistic regression at F1 91.28 %: LinearSVC scores marginally higher yet returns only a hard label, while the interface needs a probability to show how confident the verdict is.

Why did the merge operator outweigh the classifier? Absolute difference keeps the dimensions where two texts disagree, which is exactly the signal verification needs: two different authors should differ somewhere. Multiplication rewards shared features, and shared vocabulary says little about identity when both texts come from the same genre. The same logic explains why TF-IDF beat bag-of-words: down-weighting common words sharpens the author-specific tail that carries the style.

The two disappointments are worth naming. The integrated syntactic graph never matched the stylometric baselines; we attribute this to not reproducing the exact construction of Gómez-Adorno *et al.* (2016), and the representation was left untuned. The large language models failed outright. Huang *et al.* (2024) got past 88 % F1, but they used a strong proprietary model on an attribution task; our open models, chosen partly to avoid sending text to a third-party service, were not up to pairwise verification by prompting alone. We are not sure how much of that gap is model size and how much is the prompting format itself.

Set against the field, the best result here would rank near fifth among PAN 2021 entries, whose top score was 95.45 % from a five-classifier ensemble (Kestemont *et al.*, 2021). The comparison is loose, since that contest let systems leave hard cases undecided and we did not, but it places the work in a believable range. One detail stands out: no PAN 2021 team used a fine-tuned transformer head, so the Longformer result adds a data point the contest did not have.

The limits are real. Everything was measured on long English fanfiction (~21,000 characters), so short messages and other languages remain open, and there was no systematic out-of-distribution test. The transformer head also offers little interpretability, which matters in forensic and legal use where a verdict may need to be explained.

5 CONCLUSION

Across one corpus of 295,800 English pairs, a fine-tuned Longformer gave the best authorship-verification F1 at 92.38%, but classical TF-IDF stylometry came within a percentage point at a small fraction of the cost. For most deployments that trade-off points to the classical pipeline, which is what we shipped: TF-IDF with logistic regression, serving a probability through a React and FastAPI application on the Railway cloud. The choice of how to combine the two text vectors, absolute difference over multiplication, turned out to matter more than the choice of classifier. Open large language models prompted without training were not usable for this task, and the syntactic-graph route underperformed. Worthwhile next steps include pairing transformer embeddings with traditional classifiers, testing short-text and multilingual robustness, building a Czech corpus, and adding retrieval-grounded explanations so a verdict can be justified, not just stated.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-financed from the Operational Programme Research, Development and Education.

Anthropic Claude was used to summarize the text of the final thesis (Pokorný, 2026) and transform it into the form of this compact paper. The paper was then reviewed and fine-tuned by the author. The author is responsible for the correctness of the final text.

REFERENCES

- ABBASI, A. *et al.* 2022. Authorship identification using ensemble learning. *Scientific Reports*. 12, 9537. <https://www.nature.com/articles/s41598-022-13690-4>
- DEVLIN, J. *et al.* 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. *arXiv:1810.04805* <http://arxiv.org/abs/1810.04805>
- GÓMEZ-ADORNO, H. *et al.* 2016. Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. *Sensors*. 16(9), 1374. <https://doi.org/10.3390/s16091374>
- HUANG, B. *et al.* 2024. Can Large Language Models Identify Authorship? *EMNLP*, 445–460. <https://aclanthology.org/2024.findings-emnlp.26.pdf>
- KESTEMONT, M. *et al.* 2020. Overview of the Cross-Domain Authorship Verification Task at PAN 2020. CLEF. PAN. <https://pan.webis.de/clef20/pan20-web/author-identification.html>

- KESTEMONT, M. *et al.* 2021. Overview of the Authorship Verification Task at PAN 2021. CLEF. *PAN*. <https://pan.webis.de/clef21/pan21-web/author-identification.html>
- KOPPEL, M. *et al.* 2004. Authorship Verification as a One-Class Classification Problem. In: *ICML 04*, pp. 62. <https://doi.org/10.1145/1015330.1015448>
- POKORNY, J. 2026. *Authorship verification of text using selected artificial intelligence methods* [in Czech]. Master's thesis, head: Trenz, O. Brno: Mendel University in Brno.
- SAEDI, C. *et al.* 2021. Siamese networks for large-scale author identification. *Computer Speech & Language*. 70, 101241. <https://doi.org/10.1016/j.csl.2021.101241>
- SAVOY, J. 2020. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*. Cham: Springer Nature. ISBN 978-3-030-53359-5.
- STAMATATOS, E. 2008. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*. 538–556. <https://doi.org/10.1002/asi.21001>
- STAMATATOS, E. *et al.* 2016. Authorship Verification: A Review of Recent Advances. *Research in Computing Science*. 9–25. <https://doi.org/10.13053/racs-123-1-1>
- TYO, J. *et al.* 2022. On the State of the Art in Authorship Attribution and Authorship Verification. *arXiv:2209.06869*. <http://dx.doi.org/10.48550/arXiv.2209.06869>
- VASWANI, A. *et al.* 2017. Attention Is All You Need. *arXiv:1706.03762* <https://arxiv.org/abs/1706.03762>

Contact information

Jiri Pokorny: e-mail: xpokor17@node.mendelu.cz

Michaela Veselá: e-mail: michaela.vesela.ui@mendelu.cz

Oldrich Trenz: e-mail: oldrich.trenz@mendelu.cz