

Recognizing a Person Using Stylometry

Everyone writes with a fingerprint. This thesis asks a machine one question — **were these two texts written by the same person?** — and teaches it to answer from writing style alone, not from what the texts happen to be about.

AUTHOR
Ing. Samuel Šimún
SUPERVISOR
Ing. Tomáš Goldmann, Ph.D.
DEFENDED AT
Brno University of Technology
Faculty of Information Technology · 2026

Why it matters

Stylometry is a **cognitive biometric**: it identifies people not by the body but by how they think in language — the function words they lean on, how they punctuate. These habits stay stable over years, are hard to fake, and survive in plain text.

- 01 **Forensics** — tie an anonymous threat or extortion note to a known writing sample.
- 02 **Fake accounts & fraud** — reveal one person behind many profiles.
- 03 **Academic integrity** — does this essay read like the student's own earlier work?
- 04 **Open-set identity** — two texts are enough; no list of candidates needed.

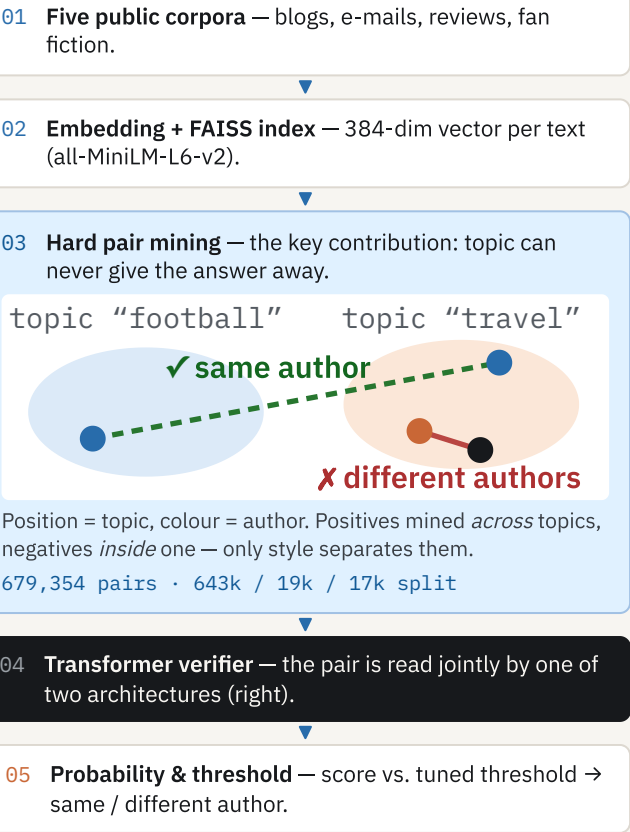
WHAT THE MODEL ACTUALLY READS

- function words
- punctuation
- sentence rhythm
- syntax
- word choice

THE PROBLEM THIS WORK ATTACKS

Deep models cheat: two texts about football get "same author" because the *topic* matches, not the person. Real cases are the opposite: one author writes about anything, and different authors constantly share topics.

The pipeline



Two architectures, compared

ENCODER · THE EFFICIENT ONE

BERT / RoBERTa + co-attention

Shared encoder weights plus a co-attention layer that lets both representations exchange information; a **style-regularisation branch** penalises reliance on meaning.

Best ROC-AUC **0.98** in-domain · 829M params · 19 ms/pair · 6.4 GB

DECODER · THE WINNER

Qwen3-4B fine-tuned with LoRA

Both texts sit in one prompt, so the model is a cross-encoder: every token of one attends to all of the other. Only lightweight **LoRA adapters** are trained.

Best ROC-AUC **0.9903** in-domain · 4.0B params · 28 ms/pair · 7.8 GB

What the experiments showed

- **Size helps once fine-tuned** — 4B beats 1.7B on every metric; zero-shot, it was worse.
- **Encoders stay relevant** — 5× smaller, within 0.01 AUC, cheaper to deploy.
- **Domain shift hits negatives hardest**: false positives rise 5% → 17%; calibration is next.

RESULTS

0.9903
ROC-AUC on the custom test set
(17,062 hard pairs)

0.9462
ROC-AUC on the PAN21 benchmark —
never trained on it

95.2%
F1 score · 4.8% equal error rate · MCC
0.90

679k
topic-controlled pairs mined for
training and evaluation

Competitive with the winners of the PAN21 verification competition — which trained *on that very data*, while this model met it for the first time at test time.