Team David Villate at PAN 2026: RoBERTa-LC, Length-Controlled Supervised Detection of AI-Generated Text

Notebook for the PAN Lab at CLEF 2026

David Villate[1,*]

[1]*Independent researcher*

Abstract

Detecting AI-generated text is becoming increasingly difficult because modern large language models are explicitly instructed to mimic human style and because benchmarks can contain length shortcuts that inflate validation performance. This paper presents RoBERTa-LC, a supervised RoBERTa-base system for the Voight-Kampff task at PAN @ CLEF 2026. The system combines architectural length control, random-crop augmentation, and class-balanced cross-entropy to reduce length confounding and prevent majority-class collapse. The study also reports negative results: Persistent Homology Dimension and Fast-DetectGPT, two zero-shot approaches that performed well in earlier benchmarks, do not provide robust signal on modern generators such as gpt-4o, o3-mini, and gpt-4.5-preview. RoBERTa-LC obtains 0.9854 mean on internal validation and 0.966 mean on the official TIRA smoke-test. Ablations show that removing length control improves validation by 0.0041 mean, confirming that part of the uncontrolled model's performance comes from exploiting length rather than style.

Keywords

AI-generated text detection, supervised classification, length control, RoBERTa, adversarial benchmarks

1. Introduction

The Voight-Kampff task at PAN @ CLEF 2026 is a binary classification problem: given a text, a system must output a score in $[0, 1]$ estimating the probability that the text was generated by a language model. This formulation is simple, but the task has two practical adversarial properties. First, the generators were instructed to mimic specific human styles, reducing obvious lexical signatures such as repetitive connectives, template-like paragraph structures, and stock phrases. Second, the organizers state that the test set contains surprises: generators unseen during training and unknown obfuscations. A system that memorizes signatures of the known generators can therefore degrade substantially out of distribution [1].

A central difficulty is the length confound. In the PAN training data, human-written texts tend to be long, often exceeding the standard 512-token limit of transformer encoders, whereas AI-generated texts have more variable lengths. A naive supervised classifier can learn the shortcut "maximally truncated text implies human" and achieve high validation scores without learning robust stylistic evidence. This shortcut is risky because a test-time obfuscation can change length without changing authorship.

The working hypothesis is that a simple supervised system can remain competitive if it explicitly addresses three risks: length confounding, class imbalance, and silent failure modes during transformer fine-tuning. RoBERTa-LC implements a RoBERTa-base classifier [2] with architectural length control, class-weighted loss, and preflight checks before training (lightweight sanity tests on model outputs and gradients, described in Section 4.3, designed to catch silent training failures before committing compute). The system uses no ensembles, no heuristic post-processing rules, and no statistics computed from the test batch.

CLEF 2026 Working Notes, September 21-24 2026, Jena, Germany

*Corresponding author: David Villate.

jdvillate@hotmail.com (D. Villate)

The contributions are: (1) a tokenization scheme that physically prevents the model from observing length during training; (2) an ablation study showing that validation performance without length control is optimistic; (3) negative results for Persistent Homology Dimension (PHD) [3], a topological estimator of the intrinsic dimension of a text's contextual embedding cloud, and Fast-DetectGPT [4]; and (4) a reproducible implementation compatible with TIRA.

2. Related Work

Zero-shot detection of machine-generated text has proposed both probabilistic and geometric signals. DetectGPT argues that generated text lies in negative-curvature regions of a language model's log-probability landscape [5]. Fast-DetectGPT accelerates this idea with a closed-form estimate of conditional probability curvature [4]. Binoculars separates human and LLM text by comparing perplexities from an observer model and a performer model [6]. PHD estimates the intrinsic dimension of contextual embedding point clouds and assumes that human text has higher dimension than generated text [3].

Supervised transformer classifiers remain strong for sequence classification. BERT [7], RoBERTa [2], and DeBERTaV3 [8] provide robust text encoders, and continued pretraining can adapt encoders to domain and task distributions [9]. In this setting, however, the critical choice is not simply architectural capacity. The main challenge is preventing the model from using spurious differences in length, encoding, or formatting; related AI-versus-human studies also emphasize stratified analysis across properties such as length and domain [10].

Calibration is also relevant because the official metric combines ranking quality, binary decisions, and probabilistic quality. Guo et al. show that modern neural networks are often over-confident and that temperature scaling can correct softmax probabilities without changing ranking [11]. This calibration is implemented, but it is not deployed in the final run to avoid overfitting a single validation split.

3. Task and Metric

Given a text x, the system outputs a score $s(x) \in [0, 1]$. Scores below 0.5 indicate human authorship, scores above 0.5 indicate AI generation, and exactly 0.5 indicates abstention. The official evaluation averages five metrics: ROC-AUC, Brier complement, C@1, F1, and F0.5u. The final metric is

$$\text{mean} = \frac{1}{5}(\text{AUC} + \text{Brier} + \text{C@1} + \text{F1} + \text{F0.5u}). \tag{1}$$

C@1 follows Peñas and Rodrigo [12]:

$$\text{C@1} = \frac{1}{n}\left(n_c + \frac{n_c}{n - n_u} n_u\right), \tag{2}$$

where n is the number of examples, n_c the number of correct decisions, and n_u the number of abstentions.

TIRA, the shared-task evaluation platform used by PAN, imposes three operational restrictions: no network access during inference, independent processing of each test case, and a command-line interface that writes JSONL predictions. These constraints favor a self-contained model with no batch-dependent normalization and no external calls. The shared-task context is described in the current PAN overview and the PAN notebook materials, and submissions are executed through TIRA [1, 13, 14, 15, 16].

4. System

Figure 1 summarizes RoBERTa-LC. The system has three blocks: representation with length control, supervised training, and TIRA-compatible inference. Inference reads JSONL, tokenizes each record independently, performs batched forward passes for throughput, and writes one $P(\text{AI})$ score per identifier—the model's predicted probability, in $[0, 1]$, that the corresponding text was generated by an AI system.

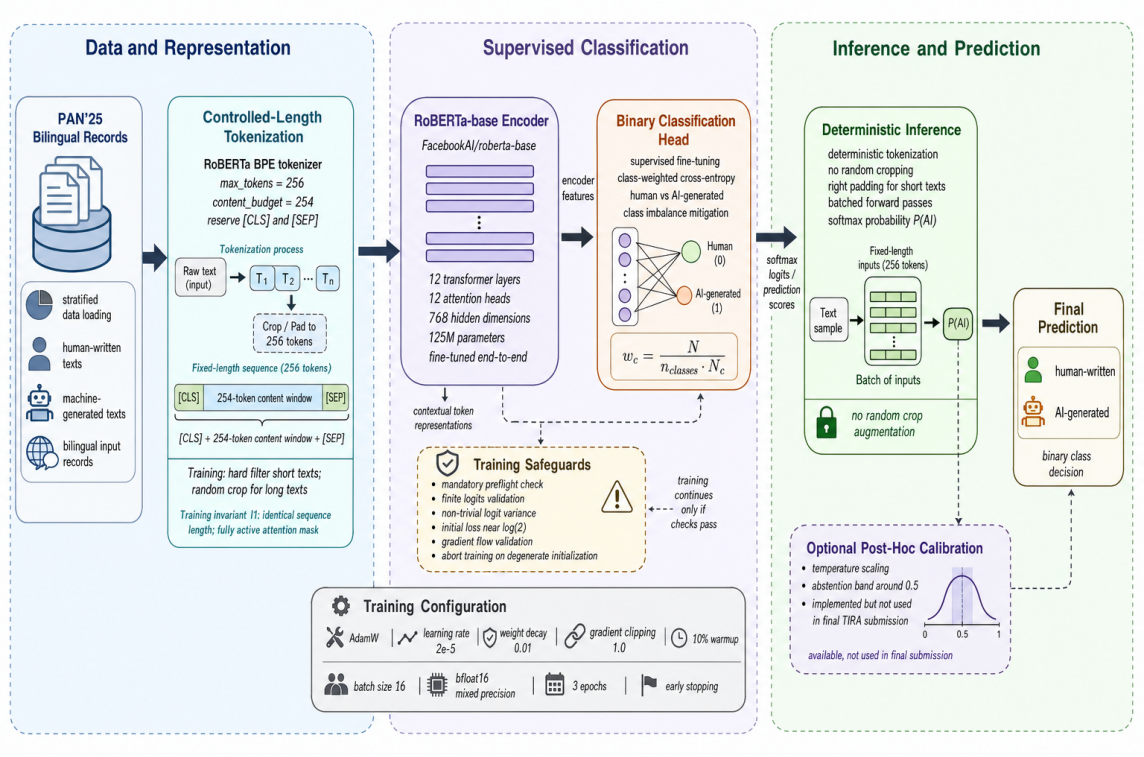

Figure 1: RoBERTa-LC architecture. The pipeline separates data reading, length-controlled tokenization, supervised training, optional calibration, and self-contained TIRA inference.

4.1. Encoder

The encoder is `FacebookAI/roberta-base`: 12 layers, 12 attention heads, and 768-dimensional hidden states. An initial evaluation considered DeBERTaV3-base, but identified a practical failure mode associated with LayerNorm parameter names in some checkpoints: `gamma/beta` versus `weight/bias`. The model could load without an error and converge toward degenerate predictions. Although a robust loader that inspects and remaps LayerNorm parameters was implemented, RoBERTa-base was selected for simplicity and auditability.

4.2. Architectural Length Control

The configuration sets `max_tokens=256` and `content_budget=254`, reserving two positions for special tokens. During training, each text is tokenized without truncation and without special tokens. If the sequence contains fewer than 254 subwords, it is removed from the effective dataset. If it contains at least 254 subwords, a random window of exactly 254 tokens is sampled and `[CLS] + window + [SEP]` is built. The `attention_mask` is 1 in all 256 positions.

This design guarantees that every training example has identical sequence length and a fully active attention mask. Therefore, the length channel cannot carry discriminative signal. During evaluation and inference, texts are not discarded: they are deterministically truncated to the first 254 tokens and right-padded if needed. The 256-token budget reduces the quadratic cost of self-attention while preserving enough stylistic signal for the task.

4.3. Weighted Loss and Preflight Checks

The training data are imbalanced: 9,101 human texts and 14,606 AI texts. To avoid the trivial minimum of always predicting AI, weighted cross-entropy is used with

$$w_c = \frac{N}{n_{\text{classes}} N_c}. \tag{3}$$

Table 1
Dataset summary.

Split	Total	Human	AI	% AI
Train	23,707	9,101	14,606	62
Validation	3,589	1,277	2,312	64

where N is the total number of training examples after the length filter, $n_{\text{classes}} = 2$ is the number of classes, and N_c is the number of examples belonging to class c. After applying the length filter, the approximate weights are $w_{\text{human}} = 1.302$ and $w_{\text{AI}} = 0.812$.

Before training, the system runs a preflight check. First, a small batch must produce finite logits with variance greater than 10^{-4}. Second, a cross-entropy backward pass must produce an initial loss near $\log 2$ and a gradient norm greater than 10^{-6} in the classification head. If any condition fails, training aborts.

4.4. Optimization and Calibration

Training uses AdamW [17], learning rate 2×10^{-5}, weight decay 0.01, gradient clipping at norm 1.0, 10% linear warmup, batch size 16, and bfloat16 mixed precision. Training runs for three epochs with early stopping on PAN mean. The implementation uses PyTorch [18] and Transformers [19]; random seeds are fixed to 42.

Temperature scaling and an abstention band $\delta \in [0, 0.20]$ are implemented as post-hoc calibration. On validation, the temperature scaling parameter $T = 1.86$ and the abstention-band parameter $\delta = 0.01$ yield a marginal 0.001 mean gain. These parameters are not included in the final TIRA run because of the risk of validation overfitting.

4.5. Data and Experimental Setup

The experiments use the official labeled training and internal validation splits. The test set is evaluated only through TIRA. Table 1 summarizes the available data.

The dataset contains 22 generators across three genres: essays, news, and fiction. Families include OpenAI, Google, Meta, Mistral, and open-weight models such as Falcon, DeepSeek, and Qwen. Two variants contain paraphrase-based obfuscation. The splits are disjoint by identifier and stratified by generator.

Preprocessing is deliberately minimal: UTF-8 text is passed directly to the RoBERTa BPE tokenizer. The pipeline does not lowercase, normalize, or clean punctuation because these elements can carry stylistic signal.

5. Results

5.1. Overall Effectiveness

Table 2 reports results on internal validation and the official TIRA smoke-test. RoBERTa-LC obtains 0.9854 mean on validation and 0.966 on the smoke test set.

On validation, the confusion matrix is $\text{TN} = 1223$, $\text{FP} = 54$, $\text{FN} = 16$, $\text{TP} = 2296$. The false-positive rate on human texts is 4.2%, while the false-negative rate on AI texts is 0.7%. The validation-to-smoke-test drop is 0.019 mean: AUC drops only 0.008, but F1 drops 0.052, indicating recall loss on some AI texts from the new distribution.

Table 2
Overall effectiveness. Baselines are taken from the task documentation and are evaluated on the validation split; the RoBERTa-LC TIRA smoke-test row reports the official smoke-test result.

System	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
PPMd baseline [20]	0.786	0.799	0.757	0.812	0.778	0.786
Binoculars baseline	0.918	0.867	0.844	0.872	0.882	0.877
TF-IDF + SVM baseline	0.996	0.951	0.984	0.980	0.981	0.978
RoBERTa-LC validation	0.9982	0.9830	0.9805	0.9850	0.9802	0.9854
RoBERTa-LC TIRA smoke-test	0.990	0.965	0.969	0.933	0.972	0.966

Table 3
Impact of length control on internal validation.

Metric	Controlled	Uncontrolled	Δ
ROC-AUC	0.9982	0.9985	+0.0003
Brier	0.9830	0.9854	+0.0024
C@1	0.9805	0.9861	+0.0056
F1	0.9850	0.9892	+0.0042
F0.5u	0.9802	0.9846	+0.0044
Mean	0.9854	0.9895	+0.0041

5.2. Ablations

Table 3 shows that removing length control increases validation mean from 0.9854 to 0.9895. This increase does not imply a more robust model: a permutation-importance analysis on the uncontrolled model shows that `text_len_words` contributes 26 times more signal than the best content feature.

5.3. Leave-One-Model-Out and Breakdowns

In two leave-one-model-out experiments, all examples from one generator were removed during training and that generator's validation examples were evaluated. For `o3-mini`, recall was 132/132; for `gpt-4-turbo-paraphrase`, recall was 59/59. This perfection must be interpreted cautiously: it can reflect real transfer across generators, but also persistent dataset confounds.

The main validation failure mode is `gemini-1.5-pro`: 10 of the 16 false negatives come from this generator. By genre, mean is 0.990 on essays, 0.989 on fiction, and 0.978 on news. By length, mean decreases from 0.990 in the shortest quartile to 0.977 in the longest, consistent with 256-token truncation discarding much of the content in long texts.

6. Negative Results

PHD was tested as a standalone detector. The original hypothesis predicts lower intrinsic dimension for generated text; in these experiments this relationship is inverted for modern generators such as gpt-4o, gpt-4-turbo, o3-mini, and gpt-4.5-preview. It holds only for older models. A plausible explanation is that modern post-training, including optimization against repetition, increases the geometric diversity of LLM outputs.

Fast-DetectGPT was also not orthogonal to RoBERTa-LC. Using gpt-neo-1.3B as scorer performed worse than gpt2 on frontier generators. In a weighted ensemble with RoBERTa-LC, Fast-DetectGPT corrected few unique errors and introduced many errors that RoBERTa did not make. This behavior is interpreted as saturation: modern LLMs produce text close to high-likelihood regions for strong scorers, reducing discriminative conditional curvature.

7. Limitations and Future Work

The main limitation is separating genuine stylistic signal from dataset-specific confounds. Length control addresses one confound, but differences in encoding, normalization, or whitespace may remain. In addition, LOO evaluation covers only two of 22 generators, active attacks are not tested, calibration is not deployed, and the system uses a single model with 256-token truncation.

Future work should prioritize paraphrase augmentation for OOD robustness, sliding-window inference for long texts, cross-validated calibration, cross-dataset evaluation of the negative results, and ensembles with genuinely orthogonal signals such as classical stylometry or syntactic features.

8. Conclusion

RoBERTa-LC shows that a simple supervised classifier can be competitive in Voight-Kampff when designed against dataset shortcuts. The system obtains 0.9854 mean on internal validation and 0.966 on the smoke test set. The ablation confirms that removing length control improves validation, but by exploiting a confound. The negative results for PHD and Fast-DetectGPT suggest that some zero-shot detectors require reevaluation under modern LLMs. The main remaining error is detecting generators that produce especially human-like text, such as gemini-1.5-pro; therefore, the most promising direction is not only increasing capacity, but auditing and mitigating benchmark confounds.

Declaration on Generative AI

During the preparation of this work, the author used generative AI tools for language editing, translation assistance, and LaTeX drafting. The author reviewed and edited the content and takes full responsibility for the publication's content.

References

[1] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Ta, Y. Wang, E. Zangerle, Overview of PAN 2026: Voight-Kampff Generative AI Detection, Text Watermarking, Multi-author Writing Style Analysis, Generative Plagiarism Detection, and Reasoning Trajectory Detection – Extended Abstract, in: Advances in Information Retrieval. 48th European Conference on IR Research (ECIR 2026), volume 16485 of *Lecture Notes in Computer Science*, Springer Nature, Cham, Switzerland, 2026, pp. 225–232. doi:10.1007/978-3-032-21321-1_32.

[2] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A Robustly Optimized BERT Pretraining Approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692. doi:10.48550/arXiv.1907.11692. arXiv:1907.11692.

[3] E. Tulchinskii, K. Kuznetsov, L. Kushnareva, D. Cherniavskii, S. Nikolenko, E. Burnaev, S. Barannikov, I. Piontkovskaya, Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts, in: Advances in Neural Information Processing Systems, volume 36, Curran Associates, Inc., 2023, pp. 39257–39276. URL: https://proceedings.neurips.cc/paper_files/paper/2023/hash/7baa48bc166aa2013d78cbdc15010530-Abstract-Conference.html.

[4] G. Bao, Y. Zhao, Z. Teng, L. Yang, Y. Zhang, Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature, Proceedings of the Twelfth International Conference on Learning Representations (ICLR), 2024. URL: https://openreview.net/forum?id=Bpcgcr8E8Z.

[5] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature, in: Proceedings of the 40th International Conference on Machine Learning, volume 202 of *Proceedings of Machine Learning Research*, PMLR, 2023, pp. 24950–24962. URL: https://proceedings.mlr.press/v202/mitchell23a.html.

[6] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text, in: Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, PMLR, 2024, pp. 17519–17537. URL: https://proceedings.mlr.press/v235/hans24a.html.

[7] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[8] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-Style Pre-Training with Gradient-Disentangled Embedding Sharing, Proceedings of the Eleventh International Conference on Learning Representations (ICLR), 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

[9] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't Stop Pretraining: Adapt Language Models to Domains and Tasks, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740/. doi:10.18653/v1/2020.acl-main.740.

[10] Y. Ma, J. Liu, F. Yi, Q. Cheng, Y. Huang, W. Lu, X. Liu, AI vs. Human–Differentiation Analysis of Scientific Content Generation, arXiv preprint arXiv:2301.10416 (2023). URL: https://arxiv.org/abs/2301.10416. doi:10.48550/arXiv.2301.10416. arXiv:2301.10416.

[11] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On Calibration of Modern Neural Networks, in: Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 1321–1330. URL: https://proceedings.mlr.press/v70/guo17a.html.

[12] A. Peñas, Á. Rodrigo, A Simple Measure to Assess Non-response, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Portland, Oregon, USA, 2011, pp. 1415–1424. URL: https://aclanthology.org/P11-1142/.

[13] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-author Writing Style Analysis, and Generative Plagiarism Detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), volume 16089 of *Lecture Notes in Computer Science*, Springer Nature, Cham, Switzerland, 2025, pp. 388–411. doi:10.1007/978-3-032-04354-2_21.

[14] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[15] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, Overview of PAN 2026: Voight-Kampff Generative AI Detection, Text Watermarking, Multi-Author Writing Style Analysis, Generative Plagiarism Detection, and Reasoning Trajectory Detection, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[16] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fröbe, M. Potthast, B. Stein, Overview of the third "Voight-

Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[17] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, Proceedings of the Seventh International Conference on Learning Representations (ICLR), 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[18] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. K"opf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An Imperative Style, High-Performance Deep Learning Library, in: Advances in Neural Information Processing Systems, volume 32, Curran Associates, Inc., 2019, pp. 8024–8035. URL: https://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning.

[19] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. Rush, Transformers: State-of-the-Art Natural Language Processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://aclanthology.org/2020.emnlp-demos.6/. doi:10.18653/v1/2020.emnlp-demos.6.

[20] O. Halvani, C. Winter, L. Graner, On the Usefulness of Compression Models for Authorship Verification, in: Proceedings of the 12th International Conference on Availability, Reliability and Security, Association for Computing Machinery, New York, NY, USA, 2017, pp. 54:1–54:10. URL: https://doi.org/10.1145/3098954.3104050. doi:10.1145/3098954.3104050.