Team bedratiuk-lab at PAN 2026: From Embedding Similarity to Semantic Canonization for Generated Plagiarism Detection

Notebook for the PAN Lab at CLEF 2026

Leonid Bedratyuk[1], Anna Bedratiuk[1]

[1]*Khmelnytskyi National University, Khmelnytskyi, Ukraine*

Abstract

This notebook paper describes the participation of team bedratiuk-lab in the PAN 2026 Generated Plagiarism Detection task. Our approach combines dense document retrieval with several local reranking signals designed to capture both semantic similarity and local structural correspondence between suspicious and source documents. In particular, we investigate fragment-level matching, soft semantic attention, and a canonical-configuration module inspired by the idea of semantic canonization in embedding spaces. The final submitted system uses a weighted combination of dense retrieval, fragment matching, semantic attention, and canonical configuration similarity. We report the development process of the method, including local experiments on a custom dataset and the final evaluation on the PAN benchmark. The results show that dense retrieval remains the strongest practical component of the pipeline, while the canonization-inspired module is more valuable as an exploratory structural signal than as a dominant ranking factor. The notebook concludes with a discussion of the limitations of the current proof-of-concept and outlines directions for future work on semantically invariant plagiarism detection.

Keywords

generated plagiarism detection, paraphrase retrieval, dense retrieval, semantic canonization, embedding geometry, PAN

1. Introduction

The PAN generated plagiarism detection task addresses the problem of retrieving source documents for suspicious texts that may contain paraphrased, rewritten, compressed, or otherwise transformed reused content. The task is organized within the PAN lab framework [**?**], described in the task-specific overview of the plagiarism detection task at PAN 2026 [**?**], and executed using the TIRA experimentation infrastructure [**?**]. In contrast to traditional lexical plagiarism detection, this setting requires methods that remain effective even when substantial surface-level rewriting obscures the original source.

Our participation in this task was motivated by a broader research question: whether plagiarism detection can benefit not only from semantic similarity in embedding spaces, but also from a more structured view of meaning-preserving transformations. In the underlying research program behind this submission, meaning is treated as an invariant of linguistic transformation, and paraphrasing is viewed as movement inside a semantic neighborhood rather than as a purely lexical operation. This perspective motivated us to move from global document similarity to local structural and canonization-inspired analysis.

The submitted system therefore follows a two-stage design. First, a dense retriever is used to identify promising candidate source documents. Second, the top-ranked candidates are reranked using several local signals intended to capture finer-grained structural correspondence between suspicious and source documents. These signals include fragment matching, semantic attention, and a canonical-configuration score derived from normalized local embedding configurations.

This notebook has three main goals. First, it documents the final submitted system. Second, it summarizes the development process that led to the final weighting of the model components. Third, it

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

*Corresponding author.

discusses what worked, what did not, and what this implies for future work on semantically informed plagiarism detection.

2. Motivation and Background

Generated plagiarism detection is particularly challenging because source reuse may be obscured by strong paraphrasing, sentence restructuring, lexical substitution, condensation, or partial recombination of multiple sources. This motivates approaches that go beyond surface lexical overlap and rely instead on semantic retrieval.

Our method starts from the observation that document and sentence embeddings often preserve semantic relatedness even under substantial rewriting. However, global semantic similarity alone does not explain why two documents are related, nor does it sufficiently distinguish genuine reuse from broader topical similarity. We therefore augment dense retrieval by local sentence-level comparison mechanisms.

A more exploratory part of our approach is inspired by the idea of semantic canonization. If a set of local paraphrastic realizations of the same meaning forms a structured neighborhood in embedding space, then it may be useful to compare not only raw similarity scores but also normalized local configurations. In the present notebook, this idea is explored in a preliminary way through canonical-configuration reranking.

3. Method

3.1. Overview

Our final submission follows a two-stage architecture:

1. dense retrieval over the full candidate corpus;
2. reranking of the top-k candidates using a weighted combination of local scores.

Let d_q denote a suspicious document and d_c a candidate source document. Each component score is computed for the ordered pair (d_q, d_c). The final score is defined as

$$S_{\text{final}}(d_q, d_c) = 0.80\, S_{\text{dense}}(d_q, d_c) + 0.15\, S_{\text{fragment}}(d_q, d_c) + 0.01\, S_{\text{attention}}(d_q, d_c) + 0.04\, S_{\text{canon}}(d_q, d_c).$$

The component weights were selected empirically on a small custom development dataset built in the same JSONL/TREC-style format as the PAN competition data.

3.2. Dense Retrieval

At the first stage, both suspicious and candidate documents are embedded using a Sentence-BERT style sentence-transformer model [**?**]. In our implementation, we use a compact MiniLM-based encoder [**?**]. Dense cosine similarity between document embeddings is then used to retrieve the most promising candidate source documents.

In all development experiments, dense retrieval proved to be the strongest practical component of the system. For this reason, it remained the dominant term in the final scoring function.

3.3. Fragment Matching

The fragment matching module compares suspicious and candidate documents at the sentence level. Each document is segmented into sentences, sentence embeddings are computed, and for each suspicious sentence the best matching candidate sentences are identified. The fragment score aggregates these local correspondences and includes a simple notion of coverage.

This component is intended to distinguish genuine source reuse from broad topical similarity. Even when two documents are globally close in embedding space, genuine reuse should yield stronger and more systematic local alignment.

3.4. Semantic Attention

The semantic attention component is a softer variant of local alignment. Instead of using only the best local matches, it computes a weighted aggregation over sentence-level similarities, where the weights are obtained through a temperature-controlled softmax.

In our development experiments, this component was informative but did not provide a strong independent gain over fragment matching. Consequently, it received only a very small weight in the final submission.

3.5. Canonical Configuration Reranking

The most exploratory component of our system is the canonical-configuration module. For each suspicious sentence, we construct a local cloud in embedding space consisting of the sentence itself and its best-matching sentences from the candidate document. This cloud is normalized by centering, scale normalization, and PCA-based alignment. From the normalized local configuration we derive a compact descriptor.

The candidate document is thus represented not only by individual similarity scores, but also by the coherence of its local normalized configurations. The purpose of this module is to move one step closer to semantic canonization: instead of comparing only raw similarities, we compare local forms after normalization.

In practical terms, this component behaved as a weak but conceptually important auxiliary signal. It did not outperform dense retrieval, but it provided the clearest connection between the final submission and our broader hypothesis on meaning-preserving transformations.

4. Experimental Setup

4.1. Development Experiments

Before the final submission, we evaluated several variants of the pipeline on:

- the PAN spot-check dataset;
- a small custom development dataset with manually designed source, paraphrase, partial-reuse, mixed-source, and hard-negative examples.

The custom dataset was structured analogously to the PAN competition format, with `queries.jsonl`, `corpus.jsonl.gz`, and `qrels.txt`. This made it possible to test the retrieval pipeline in a controlled environment while preserving interface compatibility with the official task.

We evaluated the following systems:

- dense retrieval only;
- dense + fragment reranking;
- dense + semantic attention reranking;
- dense + canonical configuration reranking;
- the final weighted combination.

On the small development datasets, recall quickly saturated, so nDCG became the more informative ranking metric.

4.2. Model Selection

The development experiments showed that dense retrieval was the strongest standalone component. Fragment matching provided a useful local verification signal, while semantic attention largely duplicated the same information in a softer form. Canonical configuration reranking was conceptually promising, but weaker as a standalone practical ranker.

A simple grid search over component weights suggested that dense retrieval should dominate the final score, with fragment matching as the strongest secondary signal. The final submission therefore used the weighting described in the method overview.

4.3. Submission Environment

The system was packaged as a Docker-based TIRA code submission and executed through the official PAN/TIRA infrastructure [**?**]. After resolving an authentication issue during setup, the final submission was uploaded successfully and executed through the TIRA interface.

5. Results

5.1. Development Findings

In internal development experiments on a small custom dataset, dense retrieval was consistently the strongest standalone component. Fragment matching and semantic attention preserved a reasonable ordering of relevant documents and provided useful local verification signals, while the canonical-configuration module was weaker as an independent reranking score. These observations directly influenced the final weighting scheme, in which dense retrieval received the dominant weight and the other components were used as auxiliary signals.

The experiments also revealed a structural limitation of the pipeline: local reranking components can help only when the correct source document is already present among the top candidates produced by the dense retriever. If the first-stage retriever misses the correct candidate, the downstream fragment-level, attention-based, and geometric signals cannot recover it.

Thus, the development stage suggested that the main practical bottleneck is the quality of the first-stage retrieval. The local modules are useful for refining and interpreting candidate rankings, but they cannot compensate for insufficient recall in the initial candidate generation step.

5.2. Official Main Test Set

On the official main test set, our submitted system achieved the following results:

- nDCG@10: 0.3862
- nDCG@100: 0.3862
- Recall@10: 0.3854
- Recall@100: 0.3854
- Reciprocal Rank: 0.6191

These results were below the official BM25 baseline. This indicates that the current proof-of-concept did not yet achieve competitive retrieval quality on the full task benchmark.

5.3. Interpretation

The most important lesson from the official evaluation is that the main bottleneck lies in candidate retrieval rather than in reranking. The equality of nDCG@10 and nDCG@100, as well as Recall@10 and Recall@100, suggests that the system often either retrieves the relevant source early or fails to retrieve it altogether. This points to insufficient robustness in the first-stage retriever under the actual task distribution.

The results also show that an interesting reranking strategy does not compensate for an insufficiently strong retrieval backbone. In other words, canonization-inspired local analysis becomes useful only after a strong enough retrieval stage has already identified the correct candidate set.

6. Discussion

Our participation produced a mixed outcome. On the one hand, the final competition score was not competitive. On the other hand, the experiments provided a clear diagnosis of where the current method fails and what should be improved next.

Three conclusions are especially important.

First, dense retrieval remains the key practical component. Any future version of the system should begin with a stronger retrieval layer, potentially through larger embedding models, improved document representations, or hybrid lexical-semantic retrieval.

Second, fragment-level matching is the most practically useful reranking signal among the local modules. It provides interpretable evidence of source reuse and is less fragile than the more experimental geometric components.

Third, the canonical-configuration module should currently be viewed as a research prototype rather than as a mature competition component. It is useful because it opens a path toward semantically invariant plagiarism detection, but in its present form it is not yet strong enough to drive ranking quality on realistic benchmark data.

More broadly, our submission should be understood as an exploration of how semantic geometry can be integrated into plagiarism retrieval, not as a finalized high-performance system. In that sense, the competition result is still valuable because it identifies the precise point where the current proof-of-concept stops being effective.

7. Conclusion

In this notebook, we presented the bedratiuk-lab submission to the PAN 2026 Generated Plagiarism Detection task. Our method combined dense retrieval with three local reranking signals: fragment matching, semantic attention, and canonical configuration similarity.

The final evaluation showed that the current system was not yet competitive on the main test set, and that dense retrieval remained the strongest practical signal by a wide margin. Nevertheless, the experiments also showed that local structural analysis is feasible within the PAN/TIRA setup and can serve as a platform for further development.

The main direction for future work is clear: stronger first-stage retrieval must be combined with better local structural modeling. In future versions, we plan to investigate larger retrievers, broader candidate sets, and more robust forms of canonization-inspired local comparison. We believe that semantically invariant plagiarism detection remains a promising long-term direction, but it requires a much stronger retrieval backbone than the one used in the present submission.

8. Software Availability

The implementation of our submission is publicly available at:

https://github.com/LeonidBed/pan26-generated-plagiarism-detection

The repository contains the Docker-based TIRA submission code, including the main execution script, dependency specification, and the final scoring configuration used in the submitted system.

Acknowledgments

The authors thank the PAN organizers and the TIRA team for providing the evaluation infrastructure and for their support during the submission process.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT for drafting, revising, and organizing parts of the manuscript and for supporting the analysis of experimental results. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.