Using Semantic Similarity and Overlap Ratio Optimized for Generated Plagiarism Detection

Notebook for PAN at CLEF 2025

Derui Mo, Huaiyu Zhang, Xiaojun Zhang and Leilei Kong*

Foshan University, Foshan, China

Abstract

This paper describes the method submitted to the PAN 2025 Generated Plagiarism Detection task. This task aim to identify unauthorized content reuse involving complex rewriting in texts. Addressing the limitations of existing baseline models in continuous fragment merging efficiency, interference from stop words, and long-fragment semantic verification capability, this paper proposes an improved multi-feature fusion plagiarism detection algorithm. The algorithm core consists of three modules: (1) Sentence overlap calculation based on word frequency statistics; (2) Continuous fragment merging strategy based on an adjacency matrix; (3) Result verification mechanism with multi-threshold constraints. By extracting deep semantic features of texts using the pre-trained GloVe.6B.300d model and combining them with traditional statistical features like word frequency overlap and adjacency relationships, a multi-dimensional detection framework is constructed, effectively enhancing the identification capability for generated plagiarism (especially texts deeply rewritten by Large Language Models (LLMs)).

Keywords

Generated Plagiarism Detection, GloVe, Semantic Similarity, Large Language Models

1. Introduction

Given a suspicious document `dsusp` and a source document collection `Dsrc`, the goal of generated plagiarism detection is to identify all continuous, maximal-length text fragments within `dsusp` that originate from `Dsrc`. These fragments may have undergone deep synonymic transformation by LLMs (e.g., sentence structure adjustment, complex synonym replacement, sentence restructuring) to conceal reuse traces. Traditional methods primarily rely on lexical overlap ratios or character edit distance [1]. While effective for detecting direct copying or simple rewriting, their ability to identify texts deeply rewritten by LLMs is limited. To mitigate this deficiency, this paper introduces the GloVe word vector model [2] to calculate semantic similarity, specifically employing the GloVe.6B.300d model to enhance the method's perception capability at the semantic level.

This method is built around text similarity analysis and machine learning classification. The core process includes:

1. **Text Preprocessing**
2. **Semantic Similarity Feature Extraction based on GloVe**
3. **Integrating the Semantic Similarity FeatureFusion and the Traditional Features**
4. **Fragment Detection and Dynamic Merging**
5. **Result Verification and Calibration**

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

*Corresponding author.

*This work was created by the yukino team. The source code can be obtained from The codes of PAN - CLEF25 Final.zip in (https://github.com/Yukinomo7/pan25-generated-plagiarism-detection2).

moderui44@gmail.com (D. Mo); zhanghuaiyu2005@gmail.com (H. Zhang); zhangxiaojun420410@gmail.com (X. Zhang); kongleilei@fosu.edu.cn (L. Kong)

0009-0000-3344-5426 (D. Mo); 0009-0003-8483-5397 (H. Zhang); 0009-0000-9521-2855 (X. Zhang); 0000-0002-4636-3507 (L. Kong)

2. Our method

This section clarifies the objective of the PAN 2025 Generated Plagiarism Detection task[5]: For a given suspicious document set `Dsusp`, it is necessary to retrieve and compare against the source document set `Dsrc` to identify all continuous, maximal-length text fragments within `Dsusp` that originate from `Dsrc` (including content deeply rewritten by LLMs). The final performance evaluation metric is Plagdet (the harmonic mean combining Recall and Precision), calculated based on the detected fragment's starting offset (Offset) and length (Length) within the documents.

2.1. Text Preprocessing

Preprocessing aims to convert raw text into a standardized format for feature extraction. Give a document, we prepross the text as follows.

Text Basic Cleaning: Uniform encoding (UTF-8), removal of HTML tags, punctuation (retaining inter-sentence separators) and consecutive whitespace; execution of lowercase conversion to eliminate formatting differences.

Tokenization and Lemmatization: Use of spaCy English tokenizer for fine-grained word segmentation; application of WordNet lemmatizer to convert vocabulary to base forms (e.g., "running" $\rightarrow$ "run"), improving feature consistency.

Stop Word Filtering: Removal of function words (e.g., "the", "and") based on the NLTK stop word list, retaining content words (nouns, verbs, adjectives) with substantial semantics.

Sentence-level Structuring: Utilization of spaCy sentence boundary detection model to segment text into sentence units. Each unit contains: (a) The original token sequence; (b) Position metadata (sentence number, starting character offset); (c) Statistical features (word count, character count, part-of-speech distribution, which includes the frequency distribution of POS tags such as nouns, verbs, and adjectives in a bag-of-tags manner rather than syntactic tree structures).

2.2. Semantic Similarity Feature Extraction based on GloVe

This module utilizes a pre-trained word vector model, GloVe, to capture deep semantic associations in text.

Given a preprocessed sentence, we obtain its sentence-level semantic vectors Svec by applying the average pooling: retrieve the GloVe embedding for each word in the sentence and average these embeddings to yield the sentence vector. This method effectively preserves the overall semantic tendency of the sentence. The details is shown in Eq (1).

$$S_{\text{vec}} = \frac{1}{n} \sum_{i=1}^{n} w_{i,\text{vec}} (1)$$

where $w_{i,\text{vec}}$ is the GloVe vector of the i-th word, and n is the number of words in the sentence.

For the GloVe, we use the publicly available GloVe.6B.300d model (trained on a 6-billion-word Wikipedia corpus), whose 300-dimensional word vectors contain rich syntactic and semantic information [2] to encode the word in a sentence. For the out-of-vocabulary (OOV) words, we handled using randomly initialized vectors.

Futhermore, we use **Cosine Similarity** to measure semantic association between sentences:

$$S_{\text{glove}} = \frac{S_{1,\text{vec}} \cdot S_{2,\text{vec}}}{\|S_{1,\text{vec}}\| \cdot \|S_{2,\text{vec}}\|} (2)$$

Its value range is $[-1, 1]$, with higher values indicating greater semantic similarity.

2.3. Integrating the Semantic Similarity FeatureFusion and the Traditional Features

To balance semantic understanding and exact matching, fuse GloVe semantic features with traditional statistical features.

We employed the following two sets of traditional text similarity features:

- **Word Frequency Overlap Features:**
  - Base Overlap (S_{base}): Number of common words in two sentences / Number of words in the shorter sentence (This is for the application of traditional methods, and this algorithm is not employed);
  - Stop Word Filtered Overlap (S_{filtered}): Number of common content words after removing stop words / Number of content words in the shorter sentence.
- **Adjacency Matrix Feature:** Construction of an inter-sentence co-occurrence relationship matrix to capture potential paragraph structural similarity (e.g., sequential consistency of consecutive sentences).

Then we integrate the semantic similarity feature and the traditional features by a linear weighted fusion

$$S_{\text{combine}} = \alpha \cdot S_{\text{glove}} + (1 - \alpha) \cdot S_{\text{filtered}}(3)$$

where α is the semantic feature weight. This fusion strategy embodies the detection logic of "semantics first, exact matching as a fallback".

2.4. Fragment Detection and Dynamic Merging

Based on semantic similarity graph adjacency analysis, we identify and merge consecutive similar sentences.

Firstly, we construct an adjacency relationship graph for further mining semantic similarity fragment. We represent sentences as nodes and S_{combine} as edge weights to form a graph $G = (V, E)$. An edge e_{ij} exists if $S_{\text{combine}}(i, j) \geq \tau_1$,where τ_1 is a preset threshold.

We use the **Breadth-First Search (BFS)** algorithm to traverse graph G, identifying connected subgraphs as candidate plagiarism fragments. Compared to edit distance-based greedy merging, BFS can effectively discover long-distance similar fragments spanning paragraphs.

Then we design the dynamic merging rules to obtain the candidate fragments:

- **Position Proximity:** The sentence order difference within the fragment $\leq \tau_2$ (τ_2 is a preset threshold)sentences in the original text (preventing erroneous connections across sections);
- **Semantic Coherence:** The average S_{combine} of sentences within the fragment $\geq \tau_3$ (τ_3 is a preset threshold, filtering fragments with semantic breaks);
- **Minimum Length Constraint:** Merged fragment word count ln (reducing noise interference).

2.5. Result Verification and Calibration

We employ a Result Verification and Calibration method through multi-layer threshold filtering and boundary optimization to enhance detection reliability.

We first apply strict conditions to initially detected fragments:

- **Character Length:** Fragment length $\geq Charln$ characters (avoiding false positives for short fragments);
- **Word Overlap Ratio:** $\text{OverlapRatio} = \frac{\text{Number of Co-occurring Words}}{\text{Number of Words in Suspicious Fragment}} \geq \tau_4$ (τ_4 is a preset threshold);
- **Comprehensive Similarity:** $S_{\text{combine}} \geq \tau_5$ (τ_5 is a preset threshold).

Then we perform local expansion search at fragment start and end positions, calculate the S_{combine} of the expanded fragment, and select the local maximum point as the final boundary, mitigating semantic break issues caused by improper sentence segmentation.

We employ the **Non-Maximum Suppression (NMS)** algorithm [8]to retain the candidate with the highest S_{combine} among overlapping fragments. Unlike traditional hard-threshold NMS used in object detection [8], we adopt a **soft-decay strategy** inspired by computer vision's Soft-NMS [6] to preserve semantically similar LLM-generated fragments. Specifically, fragments with $0.65 < S_{\text{combine}} < 0.75$ and overlap <50% are downweighted (via Gaussian decay) rather than discarded, which is critical for retaining long-distance semantic associations in rewritten text [7]. This approach aligns with the "feature fusion for irregular fragment matching" principle in image restoration tasks [7], demonstrating cross-domain applicability.

3. Experimental Results and Analysis

3.1. Datasets

The dataset is available via Zenodo. Enclosed in the train and validation corpora, two folders are found: (1) the text data and (2) the annotation data (postfix) _truths.

Text Data: contains a file which lists all pairs of suspicious documents (in the folder) and source documents (in the folder) to be compared. `pairssuspsrc`

Annotation Data: contains XML files for each pair in the file providing information about the locations and its source of reused texts. `pairs`

The annotation data contains the following information that should be used for training:

```
<document reference="suspicious-documentXYZ.txt">
    <feature
        name="plagiarism"
        this_offset="5"
        this_length="1000"
        source_reference="source-documentABC.txt"
        source_offset="100"
        source_length="1000"
        ...
    />
    <feature
        name="altered"
        this_offset="5"
        this_length="1000"
        source_reference="source-documentABC.txt"
        ...
    />
    ...
</document>
```

The `feature` specifies an aligned passage of text between `suspicious-documentXYZ.txt` and `source-documentABC.txt`, and that it is of length 1000 characters, starting at character offset 5 in the suspicious document and at character offset 100 in the source document. The other attributes are used to allow for a more detailed analysis of the results and can be ignored for training.

The `altered` feature specifies the location of paraphrased text that was not reused (no plagiarism). This allows to distinguish between genuine LLM generated texts and reused text. For the evaluation, only the `plagiarism` features need to be predicted.

For each pair `suspicious-documentXYZ.txt` and `source-documentABC.txt` in the `pairs` file, your plagiarism detector shall output an XML file which specifies the location of the plagiarism cases

detected within. The name of the feature should be `detected-plagiarism` and specify the offsets and lengths in the suspicious and the source document. No other attributes are evaluated. For example:

```
<document reference="suspicious-documentXYZ.txt">
    <feature
      name="detected-plagiarism"
      this_offset="5"
      this_length="1000"
      source_reference="source-documentABC.txt"
      source_offset="100"
      source_length="1000"
    />
    <feature ... />
    ...
</document>
```

For evaluation, the offset and length attributes of `detected-plagiarism` features will be compared against the `plagiarism` features in the annotation data. No other information will be evaluated[4].

3.2. Experimental setting

For the linear weighted fusion parameter α in Eq.(3), we optimized via grid search combined with 5-fold crossvalidation, experimentally determined to be 0.5.

τ_1 was experimentally determined to be 0.46.
τ_2 was experimentally determined to be 9.
τ_3 was experimentally determined to be 0.5.
ln was experimentally determined to be 16.
Charln was experimentally determined to be 190.
τ_4 was experimentally determined to be 0.36.
τ_5 was experimentally determined to be 0.47.

3.3. Experimental results and analysis

On PAN 2025 official dataset using Tira platform, evaluate this algorithm (Use Glove.6B.300d) and traditional baseline algorithms (pan12-baseline) [1]. Evaluation metrics follow PAN standard [1], including micro/macro average Plagdet, Recall, Precision, Granularity and Runtime. Experimental results:

Table 1
PAN 2025 Spot-Check Dataset Plagiarism Detection Results (Micro-average)

Corpus	Pairs	Micro-PlagDet	Micro-Recall	Micro-Precision	Granularity
Use GloVe.6B.300d	50	**0.598**	**0.777**	0.486	1.000
pan12-baseline	50	0.137	0.154	**0.554**	2.337

Table 2
PAN 2025 Spot-Check Dataset Plagiarism Detection Results (Macro-average)

Corpus	Pairs	Macro-PlagDet	Macro-Recall	Macro-Precision	Runtime (s)
Use GloVe.6B.300d	50	**0.541**	**0.737**	0.427	14.31
pan12-baseline	50	0.098	0.104	**0.554**	9.36

Table 3
PAN 2025 Validation Dataset Plagiarism Detection Results (Micro-average)

Corpus	Pairs	Micro-PlagDet	Micro-Recall	Micro-Precision	Granularity
Use GloVe.6B.300d	7976	**0.584**	**0.727**	0.487	1.000
pan12-baseline	7976	0.108	0.106	**0.569**	2.154

Table 4
PAN 2025 Validation Dataset Plagiarism Detection Results (Macro-average)

Corpus	Pairs	Macro-PlagDet	Macro-Recall	Macro-Precision	Runtime (s)
Use GloVe.6B.300d	7976	**0.541**	0.668	**0.455**	220.93
pan12-baseline	7976	0.077	0.073	0.509	1609.12

Table 5
Arithmetic mean of all evaluation measures per submission for the plagiarism detection alignment task.

Team	Score	System
chi-zi-zhi-xin-dui [81]	0.440	Sentence-BERT, MPNet, TF-IDF
jiruo [41]	0.263	E5 and MiniLM-L6
foshan-university [84]	0.400	TF-IDF and BERT classifier
yukino (our)[64]	**0.471**	Glove embeddings
Baseline PAN-12	0.233	Lexical near-duplicate detection
Baseline Llama-3.3 [1]	0.269	Llama-3.3 70B embeddings
Baseline Qwen2 [4]	0.375	Qwen2 7b Instruct embeddings

Experimental results indicate that the proposed algorithm demonstrates significant superiority over the baseline method [1] across multiple evaluation metrics. On the Spot-Check dataset (50 document pairs), the algorithm achieved a Micro-PlagDet of 0.598 (*+336%* improvement over the baseline's 0.137) and a Macro-PlagDet of 0.541 (*+452%*), as shown in Tables 1 and 2. The Micro-Recall of 0.777 captures *81% more plagiarized content* than the baseline (0.154), while the Granularity metric of 1.000 confirms precise identification of contiguous fragments compared to the baseline's 2.337.

On the larger Validation dataset (7,976 pairs), the algorithm sustained its with a Micro-PlagDet of 0.584 (*+441%*) and Macro-PlagDet of 0.541 (*+603%*), as reported in Tables 3 and 4. The Macro-Recall of 0.668 reflects consistent performance across diverse document pairs, while the *86% reduction in runtime* (220.93s vs. 1609.12s) demonstrates practical scalability for real-world applications.

Although the Micro-Precision slightly decreased (0.486 vs. baseline 0.554), the substantial Recall improvement drove a net gain in PlagDet. This trade-off is intentional and beneficial for detecting LLM-generated plagiarism, where deep semantic rewriting often reduces lexical overlap. The algorithm's superiority in Granularity (1.000 across datasets) confirms its ability to identify maximal-length fragments, aligning with the PAN task's emphasis on contiguous text reuse detection. These results validate the effectiveness of integrating GloVe semantic features with adjacency matrix-based merging for sophisticated plagiarism identification.

4. Conclusion

This paper proposes an improved multi-feature fusion plagiarism detection algorithm to address the limitations of existing baselines in continuous fragment merging efficiency, stop-word interference, and long-fragment semantic verification for generated plagiarism detection. The algorithm integrates GloVe-based semantic features with traditional statistical metrics (e.g., word frequency overlap and adjacency matrix analysis) through a linear weighting strategy, constructing a multi-dimensional framework to identify deeply rewritten fragments by LLMs. Experimental results on PAN 2025 datasets

demonstrate significant superiority: the method achieves Micro-PlagDet scores of 0.598 and 0.584 on the Spot-Check and Validation datasets, respectively, outperforming the baseline by 336These findings validate the algorithm's capability to enhance detection accuracy for LLM-generated plagiarism through semantic-syntactic feature fusion.

Acknowledgments

This work is supported by the National Social Science Foundation of China (Grant No. 22BTQ101).

Declaration on Generative AI

During the preparation of this work, the authors used Deepseek-R1 in order to: Drafting content and Text Translation. Further, the authors used ChatGPT-4 and Deepseek-R1 in order to: Grammar and spelling check. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] Potthast, Martin and Gollub, Tim and Hagen, Matthias and others, *Overview of the 4th international competition on plagiarism detection*, in *CLEF (Online Working Notes/Labs/Workshop)*, 2012.

[2] Pennington, J., Socher, R., & Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532-1543). Doha, Qatar: Association for Computational Linguistics.

[3] Bevendorff, Janek and Dementieva, Daryna and Fröbe, Maik and Gipp, Bela and Greiner-Petter, André and Karlgren, Jussi and Mayerl, Maximilian and Nakov, Preslav and Panchenko, Alexander and Potthast, Martin and Shelmanov, Artem and Stamatatos, Efstathios and Stein, Benno and Wang, Yuxia and Wiegmann, Matti and Zangerle, Eva, *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)*, edited by Jorge Carrillo-de-Albornoz and Julio Gonzalo and Laura Plaza and Alba García Seco de Herrera and Josiane Mothe and Florina Piroi and Paolo Rosso and Damiano Spina and Guglielmo Faggioli and Nicola Ferro, Springer, Berlin Heidelberg New York, September 2025, Madrid, Spain.

[4] Fröbe, Maik and Wiegmann, Matti and Kolyada, Nikolay and Grahm, Bastian and Elstner, Theresa and Loebe, Frank and Hagen, Matthias and Stein, Benno and Potthast, Martin, *Continuous Integration for Reproducible Shared Tasks with TIRA.io*, in *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Springer, Berlin Heidelberg New York, April 2023, Dublin, Irland, pp. 236–241.

[5] Greiner-Petter, André and Fröbe, Maik and Wahle, Jan Philip and Ruas, Terry and Gipp, Bela and Aizawa, Akiko and Potthast, Martin, *Overview of the Generative Plagiarism Detection Task at PAN 2025*, in *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, edited by Guglielmo Faggioli and Nicola Ferro and Paolo Rosso and Damiano Spina, CEUR-WS.org, September 2025, Vienna, Austria, in *CEUR Workshop Proceedings*.

[6] Bodla, N., Singh, B., Chellappa, R., & Davis, L. S. (2017). *Soft-NMS – Improving Object Detection With One Line of Code*. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. (https://doi.org/10.1109/CVPR.2017.364)

[7] Zhou, R., Xia, D., Zhang, Y., Pang, H., Yang, X., & Li, C. (2023). *PairingNet: A Learning-based Pair-searching and -matching Network for Image Fragments*. arXiv preprint arXiv:2312.08704. (https://arxiv.org/abs/2312.08704)

[8] Vedoveli, H. (2023). *NMS Unveiled: Elevating Object Detection Accuracy*. Medium. (https://medium.com/@henriquevedoveli/nms-unveiled-elevating-object-detection-accuracy-e40b8c690f8f)