A Hybrid Multimodal and Multi-Criteria System for Contextualized Plagiarism Detection in Academic Works

Elyah Frisco Andriantsialo[1], Volatiana Marielle Ratianantitra[1] and Thomas Mahatody[1]

[1] Laboratory for Mathematical and Computer Applied to the Development Systems, University of Fianarantsoa, Madagascar

Abstract. At present, there are many software applications and systems available for plagiarism detection, but most of them concentrate on global textual similarity or semantic likeness as regards the academic work without taking into account rich contextual properties. We remedy this with a hybrid for a system designed to serve the University of Madagascar. This approach integrates different cues, and it is a mixture of textual and visual analysis. Our approach uses vector representations produced by recent models (SentenceTransformer for text, CLIP for images) to interpret documents from more than one perspective. The system considers six aspects of similarity: the topic of the study, its places and methods, results as a whole, works done in the form of content obtained and inserted photos. It is given a weighted average to derive the total similarity score. This qualitative aspect of determining documents, (3), helps to distinguish between normal instances of overlap, for example shared research topics, from more questionable instances of reuse such as paraphrasing or figure replication. In our experiments, the model effectively reduced increases in false positives from contextual similarities and detected image-based plagiarism that is often missed by text-only tools. The main contribution of this work is a new methodological model and software architecture that addresses the current needs of multimodal analysis and AI-based writing.

Keywords: Plagiarism Detection, Semantic Analysis, Multi-Modal Systems, Artificial Intelligence, Embeddings, Cosine Similarity.

1 Introduction

The work presented in this paper sits at the interface of (computational) linguistics and AI, focusing on enforcement of academic honesty [1]. The digital revolution has transformed the scale and character of plagiarism in the past few decades [2]. Early detection systems, primarily based on lexical comparison [3], are no longer able to keep up with the more subtle types of misconduct such as paraphrasing or re-using ideas [4].

The recent emergence of generative AI has added a twist to the equation, alternatively dubbed algorithmic plagiarism [5], requiring more thorough semantic inspection [6]. This line of research is continuing in this direction and our work contributes to this new wave of detection research with a focus on University of Madagascar (UoM) where

we miss an institutional tool. There are several challenges to be addressed while developing a robust similarity detection model. Most such systems do not consider context at all and a more personalized index is required, which compensates for homonyms by inferring the semantic structure of a document, pushing or pulling its type 1 evidence depending on position in the document (e.g., Introduction vs. Results)

Another challenge is multimodal nature: some visual content, e.g.figures or graph can be missed by traditional tools [7]. Closing this semantic gap is similarly important [8, 9], because the intent here is to compare semantics and not superficial word formation. On the technical side, performance and scalability require techniques such as Approximate Near-est Neighbor (ANN) search and vector databases [10].

To address these correlated challenges, in this paper, we suggest a research plan for developing and implementing an antiplagiarism system based on deep semantic analysis and multimodal processing. The main output is a novel methodological, software framework and approach to approach developing threats brought by generative AI and complex forms of academic fraud.

2 Related Work

Plagiarism During the last decades, techniques for plagiarism detection have significantly improved, progressing with Natural Language Processing (NLP) [3]. They can be categorized broadly into three generations, built to overcome the limitations of the former generations.

The first generation of such systems were based predominantly on lexical comparison, and used techniques like n-gram matching and fingerprinting algorithms like Rabin-Karp to identify exact or near-exact text duplicates [3]. They did work well on large data-sets and did quickly uncover clear examples of copy-and-paste plagiarism. However, they only considered simple text similarity at the surface level and were unable to identify subtle forms of manipulation (e.g., rewording, minor changes in sentence orderings or paraphrasing) as academic misconduct has become indistinguishable which limited their effectiveness [4].

This was followed by the second generation that brought statistical and semantic analysis to overcome these limitations. Early models have also shown up in the form of the Vector Space Model (VSM) and LSA [11]. In VSM, documents are represented within a high dimensional vector space with each dimension being the term frequency-inverse document frequency (TF-IDF) of the term. VSM was able to cope with small changes in word order, but still had the problems of synonymy (car vs. automobile) and polysemy. By doing so, LSA had exposed underlying or implicit concepts/topics and provided a more sophisticated method for calculating meaning [11]. This improvement made it possible to detect conceptual similarities that purely lexical systems would overlook.

Citation Analysis and Stylometry also emerged around the same time. Citation Analysis looked for patterns of references that could indicate structural copying (see, for instance), and Stylometry employed features such as average word length or frequency

of function words to determine a change in writing style that might suggest changing authorship.

The third generation (current state of the art) is based on Deep Learning, and the transformer architectures [9]. The recent BERT model 9 takes NLP a leap further by producing contextualized embeddings that capture meaning beyond single words. Building on this technology, semantic-level plagiarism detection achieved significant progress, allowing to detect intricate paraphrasing and AI-enabled rewriting [12].

Yet, in the face of this progress, two large gaps remain. First, we believe the context of academic documents remains to be taken care of in most systems. Second, the visual aspect of research (figures, diagrams and charts) is still mostly unexplored [7].

But even with this headway, there remain two key lacunae. First, to this point, most systems do not yet integrate the structure of a paper into account. Second, the visual aspect of research is still largely not considered even though there are recent examples of multimodal models, e.g., CLIP that align text and image representations into a common vector space [13] - it is still uncommon to see visual analysis in a structured plagiarism-detection framework.

Table 1 provides a summary of various plagiarism detection methods used in the field.

Table 1. Summary of Plagiarism Detection Methods [16].

Method	Description
Fingerprinting	Represents the document in the form of fingerprints (n-grams) and utilizes algorithms such as "Rabin-Karp" for plagiarism detection.
Bag of Words	Utilizes a vector space model with vectors representing documents, calculating cosine similarity to measure the similarity between texts.
Stylometry	Utilizes statistical methods to quantify and analyze the writing style of an author based on features such as word frequencies.
BERT	Utilizes a pre-trained model on a large corpus of text, provides a deep understanding of the text but requires high computational power.
Neural networks	Achieves high performance on complex texts, detects paraphrases and similarities, but requires a high level of implementation complexity and massive amounts of data for training.
Citation Analysis	Analyzes citations within texts to detect similar patterns in citation sequences, adapted for academic and scientific texts

3 Methodology

At the University of Madagascar, like in many other large schools, originality in academic work is yet largely controlled by individual faculty members. While valuable as a form of pedagogical assessment, it is becoming increasingly unmanageable as the number of student submissions continues to scale. At the University of Fianarantsoa, where this system should be applied there are more than 30,000 students dispatched in many doctoral schools and departments.

Unfortunately, without an automatic institutional approach to plagiarism detection, it is very difficult to ensure academic honesty on such a scale. Currently available commercial and open-source tools are not reliable or trustworthy too, as they generally focus on text comparison at surface level. Such a differential is typically narrow, however and can produce misleading results when transferred to the actual academic environments.

For example, thematic repetition can occur when students are given a recurring theme (e.g. “Digitalization of Enterprises”) years after year. Ambiguity arises in the context of overlap when two students perform similar internships at different places and/or with different organizations, working on two related but separate datasets. Another instance is for research continuation, in which a project actually prolongs or follows on from the previous student work.

In all of these instances, conventional tools that compare the scriptural or structural likeness are susceptible to identifying the genuine work as a mere case of plagiarism. To address these issues, our method calls for a smarter detection mechanism. Instead of just matching words, it's using behind-the-scenes Artificial Intelligence to really understand what the text says. This makes it capable of detecting not only verbatim copying, but also subtle paraphrasing, complicated rephrasing and deep conceptual matching that traditional methods tend to overlook.

3.1 Contextualized and multi-criteria Model

The system we introduce in this paper is a hybrid and multi-criteria model designed to measure the similarity of scholarly papers more comprehensively under context. Rather than treating a document as one flat piece of text, the technique breaks it down into its main intellectual components. This breakdown is reflective of how academic writing is carried out in actuality, for instance according to the IMRD (Introduction, Methodology, Results and Discussion) format (Swales & Feak 1994) where each part also adds value to the research differently. In this way it makes the process of detection congruent with the logical and conceptual arrangement of scholarly work.

To assess originality, our framework examines six specific dimensions of analysis, each representing a key aspect of an academic paper’s content and structure:

- Theme / Subject: The core topic of the paper (essential for initial clustering).

- Study Location (or Context): The environment, company, or specific geographical/technical scope of the research (critical for distinguishing similar themes in different settings).
- Methodology and Objectives: The scientific approaches, experimental design, and goals of the study
- Obtained Results: The conclusions, findings, data, and original contributions.
- Global Document Content: The semantic similarity of the entire text body (the traditional measure).
- Visual Elements (Images): Schemes, tables, diagrams, and figures (addressing the issue of visual plagiarism).

The "Study Location" category was added to differentiate publications with a similar research focus but in different applied contexts (e.g. companies, institutions or geographical/theoretical environments). This is especially true with applied fields of knowledge like, for example, computer science, engineering or management where one aspect of originality involves being able to define part of the study enviroment. Nevertheless, the system is meant to be flexible; in fields, such as theory, where location is not a relevant characteristic, this constraint can be relaxed or replaced with different contextual information (for example the institution of experiments or mood of study).

Through the integration of text and image modeling, our approach can judge on each document in a holistic manner. Such multi-dimensional approach is intended to generate a comprehensive similarity score, which may weigh the contribution of each criterion subjectively based on its importance in an academic review. The work flow of this approach in general, and the details of the analysis steps are described subsequently and depicted in Figure 1.

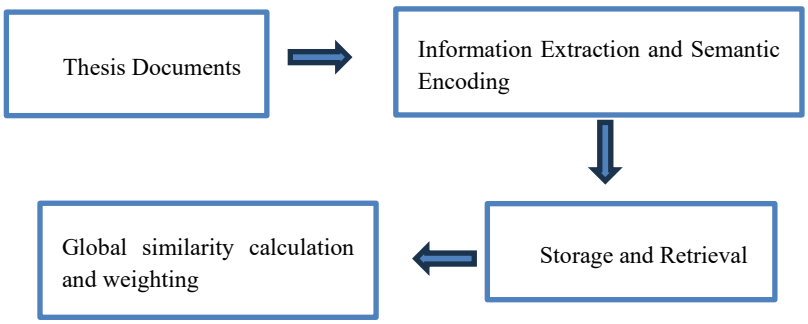

Fig. 1. The general process of our approach: from extraction to multi-criteria aggregation.

3.2 Data Preprocessing and Encoding

The quality of the output of the multi-criterion model we used, is dependent on how efficient and adequate data can be prepared. This part of the system converts raw PDF

papers to quantifiable numerical forms, referred to as embeddings, that incorporate both textural (which is one hot encoded) and visual information.

The first step in processing each document is preprocessing which involves extracting the 'full text' along with any embedded images [14]. The textual corpus is next transformed into dense semantic vectors to represent the overall analytical dimensions: THeme, LOcation, MEthodology, REsults and GLobal content (THLME-Gre schema). This is performed using a Transformer [12] based model allowing the embeddings to encode this deep semantic meaning of the text, which is required to be able to identify paraphrases with high precision. To make the pipeline more roboust, it deals with irregular PDF structures and bad scans by using PyPDF2 for text and pdf2image for images. Minimal error handling and filtering treat unreadable/empty regions of documents, to provide dependable embeddings despite diverse document formatting

In parallel, the visual elements are analyzed through a multimodal model such as CLIP [13]. This model translates image data into a semantic vector space, and the resulting image embeddings are combined to form a global visual representation of the document. The system adopts a hybrid storage structure allowing real-time querying and extensive analysis over the university corpus. Metadata and specialized embeddings (Theme, Location, Methodology, Results, Images) are stored in a Relational Database while the Vector database indexes Global Content embeddings for Approximate Nearest Neighbor (ANN) search [10]. The full data flow and the way these hybrids components interact are illustrated in Figure 2.

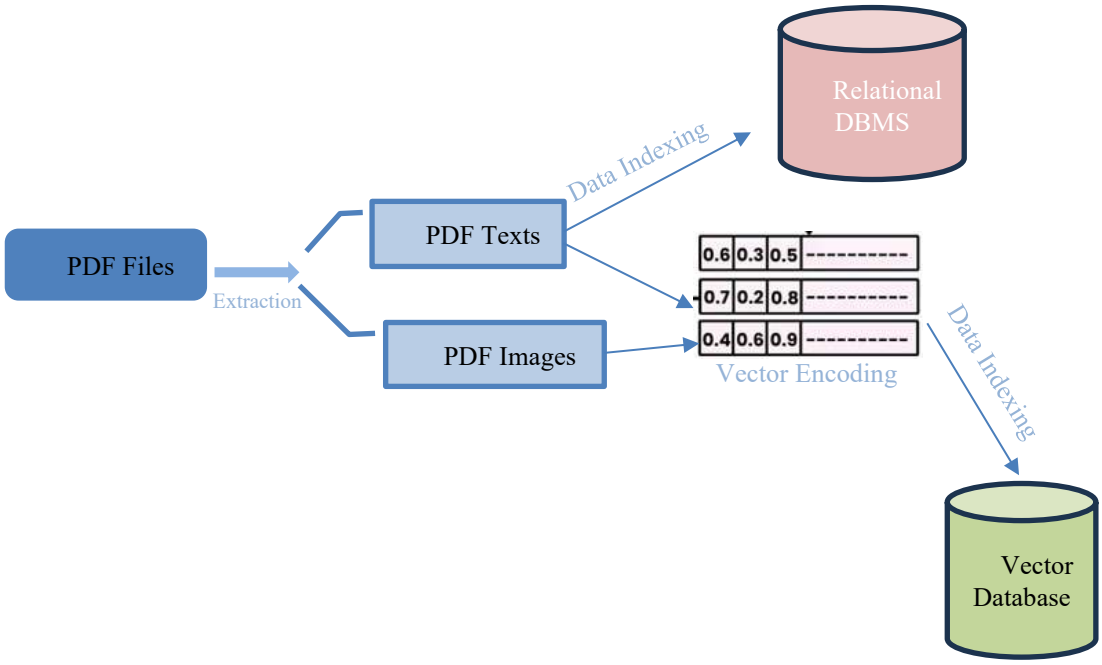

Fig. 2. System Architecture: Data flow and hybrid storage (Relational and Vector Database).

In terms of scalability, the research team used a vector-based retrieval design with ChromaDB and ANN search. This method guarantees that similarity queries will not be time consuming even with several thousands of academic works in the repository. The SentenceTransformer model selected generates small embeddings of dimensionality 384, which allow embedding storage using a real-time search algorithm while optimizing space.

3.3 Contextualized Similarity Calculation

To compare two papers, our model calculates the cosine similarity for each criterion, then aggregates these scores using an egalitarian weighting model to obtain a global score.

$$\text{similarity}(\mathrm{v}_1, \mathrm{v}_2) = \frac{\mathrm{v}_1 \cdot \mathrm{v}_2}{|\mathrm{v}_1| \times |\mathrm{v}_2|}$$

This formula represents the cosine similarity, which measures how close two vectors are in direction.
The numerator $v1 \cdot v2$ is the dot product, and the denominator $\|v1\| \times \|v2\|$ is the product of their magnitudes.

To calculate the overall similarity score, we adopt an Egalitarian Weighting Model, where each of the six criteria is assigned an equal weight (≈0.167). The global score is then computed as follows:

$$S_{global} = \frac{1}{6} \times \left(S_{theme} + S_{location} + S_{methodology} + S_{results} + S_{content} + S_{images}\right) \times 100$$

An equal weighting (1/6 for each criterion) was used as a starting point of neutrality, and this owed to the fact that textual, contextual and visual aspects were all equally crucial in academic originality. The framework was designed to be flexible, but not prescriptive, so that institutions or universities could change these weights based on their assessment policies (whether er academic domains or desired level of integrity). Additionally, future work will involve an empirical tuning of these coefficients through the use of validation data and optimization routines to better match this shaping scheme with real institutional situations.

For example, a value of 80% and above would imply that the risk of plagiarism is high, indicating substantial similarity in the contents either through copying or overreliance on them. A percentage between 60 and 80 indicates a high degree of suspicion as potential similarities are found, but they could also come from correct overlaps like similar study area or common academic expression. A score below 40% means that there is a low probability of plagiarism, meaning the document in question is highly original and not just what was taken from another source. These thresholds assist institutions in deciding when manual verification or further analysis is required to validate the originality and authenticity academic work.

4 Discussion and Evaluation

A plagiarism detector must be measured in both substance and form while the chance of false positives, and what is often more likely to occur, false negatives should be minimized. (Multi-criteria approach we also aim to stiffen this judgement by using discrete decision thresholds (0 hight: ≥ 80%, middle: 60% - 80%; low:<40%).

4.1 Protocol and Key Results

To demonstrate common examples, the Experimental Protocol presented selected controlled synthetic paper pairs (Table 2). The Validation Criterion is applied to compare our method with a content- only baseline approach.

The results (Table 2) validate the effectiveness of the method:

- False Positive Reduction (Case A): In the Case A scenario (Same theme, different locations and methodologies), both papers addressed "Conception et réalisation de site web" (Web design). A high thematic similarity (St_{heme}=0.92) was effectively neutralized by low scores in Location (0.10), Methodology (0.25), and Results (0.30). The resulting global score (35.33%) is below the suspicion threshold, successfully avoiding a false alert from the baseline (65%). This demonstrates the value of contextual criteria in discerning legitimate co-occurrence of common themes/jargon from actual fraud.

- Detection of Visual Plagiarism (Case B): This scenario, where text is rewritten but diagrams are copied, yields a very low score from the baseline (38%). However, our method detects a strong alert on the dedicated visual criterion (S_{Images}=0.92), pulling the global score into the moderate range (46.67%). This is a crucial validation of the multi-modal component, proving that it captures forms of plagiarism invisible to traditional tools.

- Reinforcement of Suspicion (Case C): For heavy semantic paraphrase (copying ideas), the global score is high (71.17%). This score falls within the "High Suspicion" zone (60%−80%), prompting mandatory manual review. The high convergence across intellectual criteria (Theme, Methodology, Results) supports the suspicion, leading to a targeted investigation rather than a blanket rejection.

- Nuanced Interpretation (Case D): When papers share only an identical university template/IMRD plan, the baseline might flag a similarity (60%). Our model, by isolating context, provides a more nuanced global score (52.00%52.00%), indicating institutional similarity rather than direct, fraudulent appropriation of core intellectual property.

Table 2. Summary of Plagiarism Detection Methods [16].

Scenario	Baseline (Content Only)	Our Method (Global Score)	Specific Alert	Key Interpretation
Case A (Similar Theme, Different Context)	65% (False Positive)	35.33%	-	False alert avoided by Location/Methodology criteria.
Case B (Visual Plagiarism, Rewritten Text)	38% (False Negative)	46.67%	S_{Images}=0.92	Detection of figure reuse; essential for AI.
Case C (Heavy Paraphrase)	68%	71.17%	-	High suspicion; convergence of intellectual scores.
Case D (Identical Template/Plan)	60% (Possible Alert)	52.00%	-	Nuance: Institutional similarity rather than direct plagiarism.

All the four validities sum to cancel its score, making the judgment stable and minimizing biases of structure, emphasizing cases with high image/results ratio. With the interpreted scores, processers can make informed judgment for mischievous activities (cheating or not cheating), which is important for academic integrity.

Even in the proposed approach, it considers text (SentenceTransformer) and Visual (CLIP) representation for measure document similarity. An ablation study will be conducted to substantiate the contribution from the visual part. In this work, we are interested in a comparison of the performance of full model (text + image) with a variant using only text. The comparison in terms of F1-score will enable us to see how much the accuracy changes when the modality images are stripped out from the model and demonstrate the actual effect of visual analysis on plagiarism detection. In one test case, two theses contained identical diagrams but different text, The text only otherwise failed (scoring 39%) by not detecting a reuse. Information obtained by the visual modality (CLIP) improved similarity up to 64%, achieving correct detection of visual plagiarism and confirming that it is an image-based analysis necessary.

Besides, a further comparison experiment with a state-of-the-art textual model (BERT) which is a strong baseline for semantic paraphrase detection was performed. While BERT is strong on pure text similarity, it does not have multimodal and context-disentangling ability. Our cross-modal approach (combining SentenceTransformer for text with CLIP for images) attained similar F1-scores on textual leading paraphrases, and significantly outperformed BERT on visual/compositional overlaps (e.g., shared figures or/and the same study settings). This work shows that through the multi-criteria and multimodal approach not only fits existing semantic models over text-based tasks but also enables to generalize them in richer academic contexts.

5 Conclusion and Future Work

This paper introduces a novel hybrid and multimodal approach for detecting plagiarism in the academic research domain, reflecting a considerable advancement over previous efforts. Integrating semantic embeddings from tasks/models like SentenceTransformer, with visual embeddings using CLIP's approach it brings together six unique dimensions: Theme Location Methodology Results Global Content Images This hybrid of visuals and text address the greatest shortcomings of strictly text-based tools. The main novelty of this study is in a contextual similarity estimation. As the case studies shown (Section 4), our model is capable of differentiating normal thematic overlap (Case A) and among complex plagiarism types, such as visual abuse or copy (Case D). This gives a more accurate and uniform rating of originality and authenticity. As such, this framework provides an implementable, scalable and academically robust structure for institutions wishing to enhance their integrity systems.

In order to enhance the system's rigor, completeness, and utility as decision-support tool, several research directions are envisaged. One aim is to improve the accuracy of evidence reporting by the system using Fine-Grained Source Traceability. Extending the citation graph pipeline by Gipp (2013) [15], during this stage highly probable segments are linked to potential sources and identified down to which exact sentence, figure or image region they contribute in terms of similarity. Such specificity is crucial for faculty members who do committee work and who need to distinguish what are truly uncited ideas from common turns of phrase. The authors are also looking into the possibility of connecting the system, in a secure and privacy preserving way, with open access repositories as well as academic databases beyond local institutional archive. Compare against larger base This will increase the comparison base and enhance reused content detection in external sources.

A third major direction is to make the system more transparent with using Explainable AI. The goal is to go beyond a single metadata score and generate clear, educational audit reports that illustrate the role of each of these six criteria, reveal major discrepancies in Methodology and Results across systems, and explain why the system found certain content visually similar e.g., by identifying matching axes or repeating data points in figures alongside textual evidence. We want to turn the system from being a black box detector into a transparent and interpretable partner that removes some uncertainty enhancing social trust on academic evaluation.

References

1. Howard, R. M. (2007). Understanding "Internet Plagiarism". Computers and Composition, 24(1), 3–15.
2. Park, C. (2013). New variant of plagiarism: a preliminary study of internet-related academic dishonesty. Assessment & Evaluation in Higher Education, 38(4), 393–405.
3. Clough, P. (2003). Old and new challenges in automatic plagiarism detection. National Plagiarism Advisory Service, JISC.
4. Martinelli, S., et al. (2018). A Taxonomy of Plagiarism: Unpacking the Nuances of Academic Misconduct. Journal of Academic Ethics, 17(4), 345–360.

5. Gao, L., Zhang, Y. (2023). The Algorithmic Ghostwriter: Generative AI and the Future of Academic Integrity. Journal of Academic Ethics, 21(3), 255–272.
6. Stein, B., Koppel, M. (2011). Delineating Plagiarism and Inspiration by Semantic Similarity. In: Proceedings of the 5th International Conference on Plagiarism Analysis, Patterns and Prevention.
7. Chowdhury, A. K., Chellappa, R. (2016). Visual Plagiarism: A New Challenge in Multimedia Forensics. IEEE Transactions on Information Forensics and Security, 11(8), 1709–1724.
8. Mikolov, T., et al. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv:1301.3781.
9. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL, pp. 4171–4186.
10. Johnson, J., Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.
11. Deerwester, S., et al. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.
12. Reimers, N., Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of EMNLP.
13. Radford, A., et al. (2021). Learning Transferable Visual Models from Natural Language Supervision (CLIP). In: Proceedings of the International Conference on Machine Learning (ICML).
14. Pérez, A. (2020). Practical PDF Data Extraction Techniques for AI and Machine Learning. Journal of Digital Document Processing, 12(4), 233–245.
15. Gipp, B. (2013). State-of-the-art in detecting academic plagiarism. International Journal for Educational Integrity, 9(1), 50–71.
16. Andriantsialo, E. F., Ratianantitra, V. M., Mahatody, T. (2024). Semantic Exploration of Textual Analogies for Advanced Plagiarism Detection. In: Proceedings 16th International Conference on Computational Processing of Portuguese (PROPOR), pp. 130–133.