Semantic Watermark for Detecting Thesis Title Reuse in Academic Repositories

Marisol Vázquez-Tzompantzi, Karen A. Neri-Espinoza, Francisco Gutiérrez-Galicia and Ricardo J. Sánchez-Quintal, *Member, IEEE*

***Abstract*— This paper proposes a semantic watermarking approach for thesis titles. Thesis titles are frequently used as the first key to search, classify, and monitor information in academic repositories. However, given the short nature of these, detecting title reuse or duplication in themes is not straightforward, especially when different titles share disciplinary terms or similar wording. The watermark is understood as an external semantic fingerprint generated from the lexical structure of each title. Each title was represented using TF-IDF vectors with unigram and bigram features, and similarity between titles was estimated through cosine similarity. The method was applied to an institutional corpus of 27,241 graduate thesis titles. Random title pairs showed an average cosine similarity of 0.02, while nearest-neighbor pairs reached an average value of 0.52. A threshold of 0.75 was then used to identify 3,300 pairs with high lexical-semantic similarity. Results indicate that semantic watermarking can be used as a simple first filter in order to support title-level analysis, metadata quality control, and the identification of potential thematic redundancy in academic repositories.**

I. Introduction

Academic repositories have become the main access point for theses, dissertations, and other institutional research products. Theses titles play a central role because they summarize the research topic in a very short textual form. However, this same brevity makes title-level analysis difficult, especially when different works share disciplinary terminology, similar methodological structures, or partially reformulated wording. Plagiarism detection platforms, such as Turnitin, are designed to compare full documents, citations or extended text corpus. Academic plagiarism detection has been widely studied, with approaches ranging from lexical matching and fingerprinting to semantic analysis and deep learning methods. Foltýnek et al. [1] reviewed computational approaches for plagiarism detection and demonstrated how the field has moved from surface text revision to the identification of paraphrased or conceptually related content. More recently, semantic similarity-based methods have been shown to improve detection of sophisticated forms of plagiarism by analyzing deeper relationships between terms and concepts beyond exact word overlap[2]. Nevertheless, most of these approaches are still oriented toward full-text analysis, while short academic data, such as thesis titles, remain less systematically studied.

Although plagiarism detection has advanced considerably for full-text documents [3], title-level analysis still presents specific challenges. The thesis titles are short, but they usually contain the research object, method, population, material, or application area. For this reason, two titles may look different at the surface level and still refer to closely related academic work. At the same time, titles from the same discipline often share common terms, which makes it necessary to distinguish between normal thematic coincidences and possible reuse or reformulation.

In this work, we propose a semantic watermarking approach for academic thesis titles [4]. The watermark is understood as a non-intrusive semantic fingerprint generated from the lexical structure of each title. No information is inserted into the original text; instead, the title is represented externally through a vector space model based on term frequency–inverse document frequency (TF-IDF). The similarity between titles is then estimated using cosine similarity. TF-IDF was selected because it is computationally lightweight and widely used in information retrieval and text mining. This makes it suitable as a filter for large academic repositories, where interpretability is important for institutional review.

Semantic fingerprints/watermarks have been used in related problems where documents need to be compared without relying only on exact text matching. For example, Zhai et al. [5], used semantic fingerprints generated from textual features for author name disambiguation in large academic repositories. In that case, the fingerprint helped represent each document in a compact form so that similarity could be analyzed beyond identical names or direct string coincidences. Pang et al. [6] proposed a text similarity method based on semantic fingerprints built from characteristic phrases extracted with TF-IDF, showing that this type of representation can be useful when the text has been partially modified or reformulated. These works address different problems from the ones studied here, but they provide a useful basis for representing academic text through external fingerprints. In our case, the fingerprint is applied to a much shorter unit of analysis: the thesis title. For

M. Vázquez-Tzompantzi is with the Dirección de Posgrado of the Instituto Politécnico Nacional Mexico City, 07320, MX. (corresponding author e-mail: mvazquezt@ipn.mx).

K.A. Neri-Espinoza is with the Unidad Profesional Interdisciplinaria de Ingeniería Campus Hidalgo of the Instituto Politécnico Nacional, Hidalgo, 42162 MX. (corresponding author e-mail: kneri@ipn.mx).

F. Gutiérrez-Galicia is with the Escuela Superior de Ingeniería y Arquitectura, Mexico City, 07320, MX. (fgutierrezga@ipn.mx).

R. J. Sánchez-Quintal is with the Facultad de Ingeniería of the Universidad Autónoma de Campehe, San Francisco de Campeche City, 24085, MX. (e-mail: rjsanche@uacam.mx).

that reason, we use the term semantic watermark in a non-intrusive sense. The watermark is generated as an external representation that allows title pairs to be compared within a common vector space.

The proposed method is an unsupervised strategy because labeled examples of title reuse, reformulation, or non-reuse are not usually available in institutional repositories. This approach is suitable for repository analysis, where the objective is not to make automatic accusations of plagiarism, but to identify candidate pairs for human review. The use of TF-IDF and cosine similarity keeps the process computationally simple and interpretable, which is important when the results must be examined by academic committees or institutional reviewers. From an academic management perspective, title-level similarity analysis can help to quality control, the detection of repeated formulations, and the exploration of thematic concentration across graduate programs. Similar titles do not necessarily imply improper reuse; they may also reflect common research lines, shared methods, or institutional trends. For that reason, this proposed semantic watermarking approach is intended to complement existing plagiarism detection systems and expert evaluations.

The paper is organized as follows: Section II describes the proposed semantic watermarking methodology and the construction of TF-IDF-based title fingerprints. Section III presents the results obtained from a corpus of 27,241 graduate thesis titles. Section IV discusses the scope, limitations, and institutional implications of the approach, and Section V summarizes the main conclusions.

II. Methodology

This study proposes a semantic watermarking mechanism for academic thesis titles based on vector space representations that can be graphically explained in Figure 1. The analysis was performed using a corpus of 27,241 graduate thesis titles from an institutional repository, previously subjected to a minimal lexical cleaning process. Each title was considered an independent unit of analysis.

The semantic watermark is defined as a normalized vector representation of the title, generated using the Term Frequency–Inverse Document Frequency (TF-IDF) scheme. The TF-IDF model was globally fitted using the entire corpus to construct a shared semantic space. Unigrams and bigrams were considered to capture both individual terms and short lexical patterns, and the representations were normalized using the L2 norm.

The semantic similarity between titles was assessed by calculating the cosine similarity between their TF-IDF vectors (see Figure 2). For each title, the most similar titles (top-k) were identified, excluding self-similarity. Additionally, a statistical baseline was established by calculating the cosine similarity between randomly selected pairs of titles, allowing for the characterization of the expected behavior for unrelated titles.

Once these values were obtained, a sensitivity analysis was carried out where the thresholds were tested in different ranges ranging from 0.60, 0.75, 0.80 and 0.90, and thus determine which range could be discretized and considered to decide when a title is a candidate for review.

A. Manual validation protocol

To address the need for empirical validation beyond descriptive similarity statistics, a manually reviewed sample of title pairs was created. The sample included three similarity groups: high-similarity pairs with a cosine similarity greater than or equal to 0.75; medium-similarity pairs with values between 0.60 and 0.75; and low-similarity control pairs with values less than 0.10. This manual validation phase was included to estimate the nature of the detected similarities, classifying them into five categories: exact duplicate, lexical reformulation, thematic similarity, false positive, or unrelated.

Broadly speaking, the proposed method does not use supervised learning or require labeled data; therefore, a traditional training-test partitioning was not performed. Validation was carried out using statistical separation analysis and empirical evaluation of semantic similarity, with an emphasis on robustness to lexical reformulations.

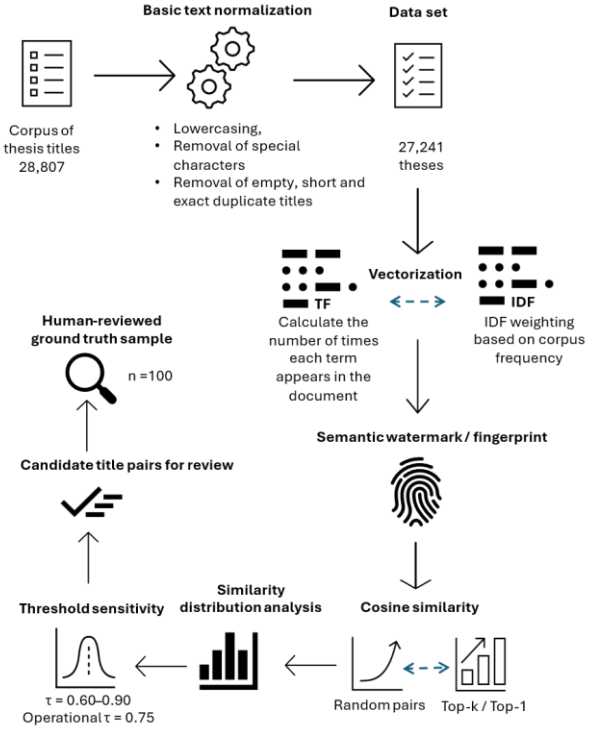

Figure 1. Diagram of the methodology applied for the semantic similarity analysis of thesis titles. The titles in the institutional repository undergo light preprocessing and are transformed into TF-IDF-based semantic fingerprints. These fingerprints are compared using cosine similarity in both random pairs and nearest neighbors, allowing analysis of the statistical separation between the two scenarios and the definition of a high similarity threshold for detecting semantically related titles. A threshold sensitivity analysis is applied using τ values from 0.60 to 0.90, with $\tau = 0.75$ selected as the operating threshold. The candidate title pairs are then evaluated using a reference sample for manual review.

B. Proposed institutional monitoring architecture

In addition to similarity analysis, the method is presented as a standardized institutional monitoring architecture for academic repositories. The workflow begins with the extraction of thesis titles to be validated, followed by minimal lexical cleaning and the generation of a TF-IDF-based semantic fingerprint for each title. These vectorizations are compared using cosine similarity to identify the most similar titles, which are then prioritized using predefined operational thresholds. Finally, the candidate pairs undergo human review and can be used to generate institutional reports for metadata curation, repository quality control, and monitoring for potential cases of title reuse or reformulation. In this architecture, similarity values function solely as indicators to support decision-making and not as automatic evidence of plagiarism or academic misconduct.

III. RESULTS

After minimal data cleaning, the final corpus consisted of 27,241 thesis titles. The cosine similarity distribution for randomly selected pairs of titles showed very low values, with an average similarity of 0.022 and a 95th percentile of 0.063 as Figure 2 shows graphically, indicating that unrelated titles rarely share significant semantic similarity.

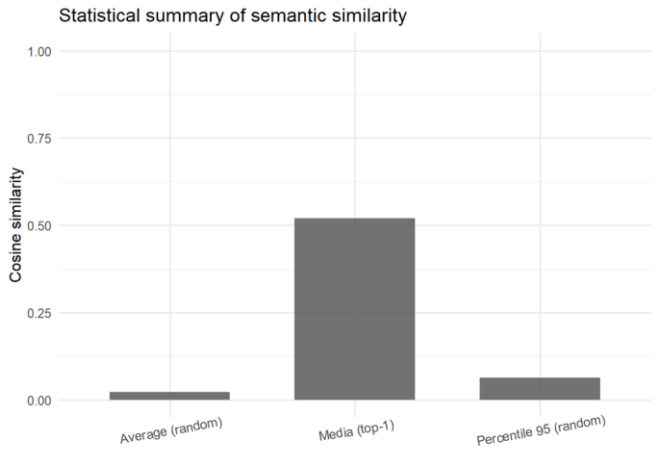

Figure 2. Comparison between random behavior and actual similarity (top-1).

Comparison with random pairs and Threshold sensitivity analysis

Random pairs function as a control group. The comparison between random pairs and nearest neighbors demonstrated that not all titles are similar by chance; rather, there is a group of titles that exhibit a much higher similarity than the repository's typical behavior. The nearest neighbor is the title most similar to another within the entire corpus, looking for shared terms or a similar semantic structure. Comparing both scenarios reveals a statistical separation between the similarity expected by chance and the potentially relevant similarity. This separation allows for the definition of an operational threshold for selecting pairs as candidates for human review.

Table 1 presents the sensitivity analysis of the threshold (cosine similarity). It shows how the number of candidate pairs detected decreases as the similarity threshold becomes more restrictive. With $\tau = 0.60$, 11,490 pairs were detected, indicating a broad similarity scenario with a higher probability of false positives. With $\tau = 0.90$, only 1,368 pairs were detected, corresponding to a highly restrictive scenario that primarily captures nearly identical titles or marked reformulations. The selected threshold of $\tau = 0.75$ identified 3,300 candidate pairs, representing an intermediate operating point between broad similarity and stricter, high-confidence detection.

Table 1. Sensitivity analysis of the threshold.

Threshold τ	Detecte candidate pairs	Interpretation
0.60	11, 490	Broad screening; higher risk of false positives
0.70	4711	Medium-high similarity; requires careful review
0.75	3300	Proposed operational threshold for high similarity
0.80	2444	Stricter detection; lower candidate volume
0.90	1368	Very strict; possible duplicates or near-identical titles

Figure 3 compares two scenarios of semantic similarity between thesis titles. The density representation highlights the statistical separation between random similarity and nearest-neighbor similarity, regardless of the corpus size. The blue (left) distribution represents the cosine similarity distribution obtained from randomly selected pairs of titles, which serve as a control and describes the expected behavior by chance. In this case, the similarity values are concentrated very close to zero, indicating that unrelated titles generally exhibit minimal semantic similarity.

On the other hand, the red (right) distribution corresponds to the cosine similarity between each title and its nearest neighbor (top-1) within the corpus, that is, the most semantically similar title. Unlike the random control, this distribution is clearly shifted toward higher similarity values, reflecting the existence of titles with similar thematic content or potentially reused content through lexical reformulation. The vertical line at $\tau = 0.75$ indicates the similarity threshold defined from the statistical analysis of both distributions. This threshold is well above random behavior (the 95th percentile is at 0.063) and marks the point from which the observed similarity can hardly be attributed to chance. In other words, the 0.75 threshold serves as a parameter to decide at what point a pair of titles should be marked as a case of high similarity (See Table 1).

In contrast, Figure 4 shows the similarity distribution corresponding to the nearest semantic neighbor (top-1) of each title showed considerably higher values, with an average similarity of 0.52 and a 95th percentile reaching values as high as 0.89. This pronounced difference demonstrates a clear statistical separation between unrelated titles and semantically related titles.

Based on these distributions, a similarity threshold $\tau = 0.75$ was established to identify cases of high semantic similarity.

Applying this threshold allowed the detection of 3,300 pairs of titles with strong semantic correspondence, which represent potential cases of title reuse, thematic convergence, or reformulation within the institutional repository.

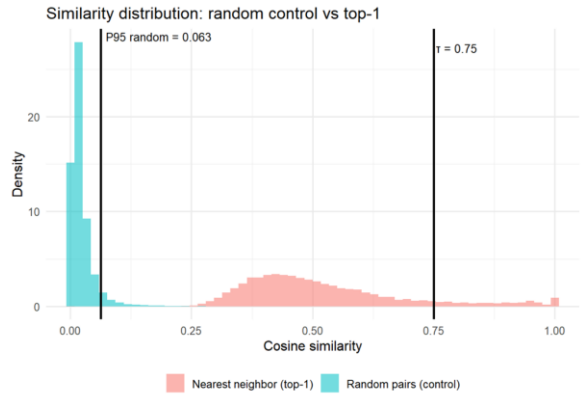

Figure 3. This shows the cosine similarity distribution (normalized) between thesis titles for two scenarios: randomly selected pairs of titles (control) and the nearest semantic neighbor (top-1) for each title in the corpus. In the case of random control, the distribution is concentrated in values close to zero, with an average similarity of 0.022 and a 95th percentile of 0.063, indicating that unrelated titles have a very low probability of semantic match.

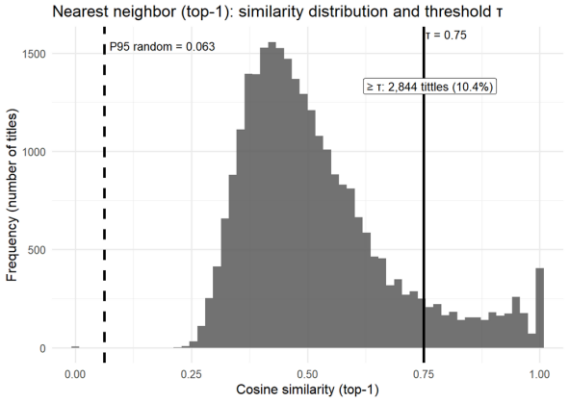

Figure 4. Nearest neighbor (top-1) for similarity distribution and threshold τ. The threshold defines the region of high similarity ("suspicious") and allows for the quantification of cases.

Consequently, title pairs that exceed this value are considered candidates for semantic reuse, which confirms that the adopted threshold is not arbitrary, but is based on a clear statistical separation between the two scenarios.

The methodology included the manual validation of the results, which is described in Table 2. Table 2 summarizes the results of the manual validation. In the high-similarity group, corresponding to the proposed operating threshold $\tau = 0.75$, 36 pairs were classified as exact duplicates and 40 as lexical reformulations. Therefore, 76% of the reviewed high-similarity pairs corresponded to exact duplicates or reformulated titles. Therefore, 76% of the reviewed high-similarity pairs corresponded to exact duplicates or reformulated titles.

Table 2. Manual validation results for the sampled title pairs.

Similarity group	Exact duplicate	Lexical reformulation	Thematic similarity	False positive	Unrelated	Total
≥ 0.75	36	40	14	10	0	100
0.60–0.75	0	2	42	56	0	100
< 0.10	0	0	0	0	59	59
Total	36	42	56	66	59	259

In addition, 14% were classified as thematic similarities requiring human review, while 10% were considered false positives. In the medium similarity group, most cases were classified as thematic similarities or false positives, indicating that lower thresholds increase coverage but also ambiguity. The low similarity control group was classified as unrelated, supporting the observed statistical separation between random or weakly related titles and high-similarity candidates.

In this work, contextual embeddings are not used; the semantic watermark is based on a TF-IDF representation with n-grams, which allows for an interpretable and scalable detection of semantic similarity at the title level.

IV. Discussion

The results obtained indicate that the proposed semantic watermarking approach effectively captures relevant semantic relationships between thesis titles, while maintaining low similarity values for unrelated titles. The clear separation between random similarity and nearest neighbor similarity supports the robustness of TF-IDF-based semantic fingerprints as invisible identifiers.

Sensitivity analysis also shows that $\tau = 0.75$ should not be interpreted as a universally optimal threshold. Rather, it represents an operational decision point for prioritizing high-similarity title pairs in an institutional review context. Lower thresholds can increase coverage but also increase the number of pairs requiring human review and the risk of false positives. Higher thresholds reduce the review workload but may overlook relevant lexical reformulations. Therefore, the threshold can be adjusted according to the repository size, institutional review capacity, and the desired balance between coverage and accuracy.

Manual validation confirms that the proposed approach is useful as a mechanism for selecting candidates for human review. In the high-similarity group, most of the reviewed pairs corresponded to exact duplicates or lexical reformulations, supporting the use of $\tau = 0.75$ as an operational threshold. However, the presence of thematic similarities and false positives also confirms that a high cosine similarity should not be interpreted as automatic evidence of plagiarism or misconduct. Similar titles can appear naturally in academic repositories due to shared topics, standard methodological expressions, or disciplinary conventions. Therefore, the proposed semantic watermark should be used as a decision-support tool for repository management and metadata quality control, not as an automatic punitive mechanism.

A key finding is that the pairs of titles detected with high similarity do not correspond solely to exact duplicates, but also

include cases of lexical reformulation, word order changes, and the use of synonymous expressions. This demonstrates that the method transcends superficial matching of text strings and allows for the identification of semantic reuse patterns that are difficult to detect using traditional keyword-based approaches.

From an institutional perspective, semantic watermarking constitutes a lightweight and scalable mechanism for monitoring academic repositories, tracking thesis production, and conducting exploratory analysis of thematic redundancies. Although this work focuses exclusively on thesis titles, the approach can be extended to other types of academic metadata or combined with contextual information in future work

A. Future work

Future research projects include extending the semantic watermarking approach to representations based on deep language models, such as contextual embeddings, to compare their performance against the TF-IDF scheme in more complex reformulation scenarios. The integration of additional contextual information, such as subject area, academic program, or time period, is also being considered to enrich the similarity analysis and reduce potential thematic ambiguities.

Another relevant line of research involves applying the method to other academic metadata, such as abstracts, keywords, or descriptors, which would allow for evaluating the scalability of the approach to longer texts. Similarly, the development of visual tools and interactive interfaces is envisioned to facilitate the exploration of semantic similarities by institutional managers and academic committees.

Finally, the proposed approach can serve as a basis for continuous repository monitoring systems, supporting quality control processes, trend analysis, and institutional decision-making related to academic output and knowledge management.

V. CONCLUSION

This paper presents a semantic watermarking approach for academic thesis titles, aimed at detecting semantic reuse and lexical reformulation and high thematic similarity in institutional repositories. Using a vector representation based on TF-IDF and cosine similarity, a non-visible semantic fingerprint was defined to represent the lexical structure of each title and support similarity-based structure of each title and, capable of capturing semantic relationships between titles that are not textually identical.

The evaluation performed on a real corpus of 27,241 postgraduate thesis titles demonstrated a clear statistical separation between randomly selected title pairs and nearest-neighbor title pairs. While the average similarity between random pairs remained close to zero, the similarity values corresponding to the closest semantic neighbors were significantly higher, supporting the usefulness of the proposed approach for candidate detection. The threshold sensitivity analysis showed that $\tau = 0.75$ provides an intermediate operating point for prioritizing high-similarity title pairs. At this threshold, 3,300 candidate pairs were identified, maintaining the method as a decision-support tool for human review or manual review, rather than an automatic plagiarism detection mechanism.

Manual validation of the selected title pairs further supported the usefulness of the proposed approach. Among the high-similarity pairs reviewed, 76% were exact duplicates or lexical reformulations, while other cases presented thematic similarities that required human evaluation. These findings confirm that the method is suitable for prioritizing candidate title pairs for institutional review, although it should not be interpreted as automatic evidence of plagiarism.

The method could be strengthened and made more reliable and robust by complementing or comparing it with other textual and semantic similarity models, such as SBERT or fuzzy matching. However, due to the scope and presentation of this work, it was not possible to include it.

Finally, the results suggest that the proposed semantic watermark/digital fingerprint is a viable, interpretable, and scalable tool to support metadata quality control, repository monitoring, and the analysis of thematic redundancies in institutional repositories. Its main value lies in reducing the search space and prioritizing potentially relevant title pairs for expert review, while preserving human judgment as a central component of the evaluation process.

ACKNOWLEDGMENT

The authors would like to thank the Dirección de Posgrado of the Instituto Politécnico Nacional (IPN) for providing access to the thesis dataset used in this study. The authors acknowledge financial support through project IPN-SIP-IND-2026-0780.

REFERENCES

[1] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic plagiarism detection: A systematic literature review," *ACM Comput. Surv.*, vol. 52, no. 6, 2020, doi: 10.1145/3345317.

[2] A. Amirzhanov, C. Turan, and A. Makhmutova, "Plagiarism types and detection methods: a systematic survey of algorithms in text analysis," *Front. Comput. Sci.*, vol. 7, no. Ml, 2025, doi: 10.3389/fcomp.2025.1504725.

[3] S. V. Moravvej, S. J. Mousavirad, D. Oliva, and F. Mohammadi, "A Novel Plagiarism Detection Approach Combining BERT-based Word Embedding, Attention-based LSTMs and an Improved Differential Evolution Algorithm," pp. 1–29, 2023, [Online]. Available: http://arxiv.org/abs/2305.02374

[4] N. S. Kamaruddin, A. Kamsin, L. Y. Por, and H. Rahman, "A Review of Text Watermarking: Theory, Methods, and Applications," *IEEE Access*, vol. 6, pp. 8011–8028, 2018, doi: 10.1109/ACCESS.2018.2796585.

[5] X. Zhai, H. Han, Z. Li, and Y. Ran, "Research on author name disambiguation based on fusion features and semantic fingerprints," *Journal of Physics: Conference Series*, vol. 1302, no. 2, p. 022013, 2019, doi: 10.1088/1742-6596/1302/2/022013.

[6] S. Pang, J. Yao, T. Liu, H. Zhao and H. Chen, "A Text Similarity Measurement Based on Semantic Fingerprint of Characteristic Phrases," *Chinese Journal of Electronics*, vol. 29, no. 2, pp. 233–241, Mar. 2020, doi: 10.1049/cje.2019.12.011.