iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** V **Month of publication:** May 2026

DOI: https://doi.org/10.22214/ijraset.2026.82721

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue V May 2026- Available at www.ijraset.com

Patent Similarity Checker Using TF-IDF and Cosine Similarity for Prior-Art Detection and IPR Risk Assessment

Ian Jem[1], Ishitha Busetty[2], K. Mrudula Reddy[3], Dr. Chitra B. T.[4]
[1, 2, 3]*Department of Computer Science & Engineering, R.V. College of Engineering, Bengaluru, India*
[4]*Department of Industrial Engineering and Management, R.V. College of Engineering, Bengaluru, India*

Abstract: ***The exponential growth of patent filings in technology intensive areas has made manual prior-art searches increasingly unreliable often leading to inadvertent infringement and duplicate innovation. This paper introduces a Patent Similarity Checker, a computational tool that employs Term Frequency-Inverse Document Frequency (TF-IDF) vectorisation and Cosine Similarity scoring to help determine prior art and assess Intellectual Property Rights (IPR) risk in the pre-filing stage. The system processes patent documents through a pipeline of tokenisation, stop-word removal and stemming before building a weighted document-term matrix. Pairwise Cosine Similarity scores are calculated across the corpus to quantify the textual proximity between a candidate patent and existing filings and are mapped onto a tiered risk scale of low, moderate, or high providing a reproducible data-backed basis for prosecution decisions for inventors and legal practitioners. In addition to the technical perspective, the paper places this tool in the context of Indian patent law, the novelty and inventive step requirements under the Patents Act, 1970 and similar TRIPS obligations. The report considers whether algorithmic similarity scores may be used to aid examiner judgement in prior-art determinations, and whether they may be of evidentiary value in opposition or infringement proceedings. The paper also raises important questions about the limits of purely textual matching, and especially its inability to capture claim scope, functional equivalents, or drawing-based disclosures, arguing that the tool is best understood as an accessible first-pass filter rather than a replacement for expert legal analysis. In conclusion, this study supports the controlled use of NLP-based approaches in patent practice, providing a scalable tool for mitigating IPR disputes ex-ante.***

Keywords: TF-IDF Vectorisation, Cosine Similarity, Prior-Art Detection, IPR Risk Assessment, Patent Similarity, NLP, Indian Patent Law.

I. INTRODUCTION

Patent applications have increased across the globe with the growth rate exceeding 5% annually in the past decade due to innovation in artificial intelligence, semiconductor, biotech, and digital media fields. Actually, the volume of 90,000+ patent applications was registered by the Indian Patent Office in the fiscal year 2024-25. The vast amount of cases complicates human examination of prior art in accordance with the statutory deadlines set in legislation.

The lack of detection of prior art and its functional equivalent and failure to review all aspects lead to an inappropriate grant of the patent and subsequent opposition activities after patent issuance.

The suggested approach for assessing semantic similarity between the artifact and the reference body within the present technical and legal framework consists of transformation of similarity scores to risk categories that are categorized in accordance with the law. In the course of this research, we suggest the formalization of the Patent Similarity Checker in the vectorized form using TF-IDF scoring and cosine similarity metric based on the tri-level IPR risk classification according to Indian standards on patents.

Moreover, we examine the evidentiary value of Patent Similarity Checker under the Patents Act, 1970, and TRIPS. This paper is structured as follows.

The subsequent section presents the review of existing literature. The implementation of the NLP pre-processing procedure is covered in Section III. Vectorization and calculation of similarity are explained in Section IV. Dynamic tiered risk scaling and intersection with the law are outlined in Section V. Experimentation and dashboard evaluation results are depicted in Section VI. Concluding observations are provided in Section VII.

II. LITERATURE REVIEW

The computational treatment of Intellectual Property by semantic parsing is a live optimisation. Recent comprehensive surveys by Shomee et al. [1] trace the evolution of patent document screening from traditional linguistic feature mining to modern deep learning methods, describing how language processing can bridge gaps across multi-modal applications.

In text-based architectures, systematic investigations into language parsing frameworks by Jiang [2] show that proper term-weighting structures still are competitive against highly complex embeddings for tracking structural boundaries in technical records. Additionally, comparative benchmarks by Ali et al. [3] show that automated retrieval applications scale best when prior-art databases use sequential textual normalisation.

Ascione and Sterzi [4] performed baseline experiments demonstrating that document-term models reliably map semantic fields, without generating cross-domain distortion, when evaluating the geometric properties of lexical spaces.

Peng and Yang [5] show that the tracking of co-occurring multi-word expressions raises the accuracy thresholds for global text alignment validation by using graph-guided structural tracing. Simultaneously, deep feature monitoring pipelines explained by Jiang et al. [6] reveal that basic quantitative scoring yields highly stable metrics for overlap detection, while deep vector methodologies leveraged by Yoo et al. [7] show that pairwise angular distances effectively capture proximity boundaries over large, unlabelled public indices.

This study solves contemporary search challenges through a standardized automatic pipeline under a clear legal interpretation of the Indian Patents Act, 1970, which offers a practical approach without placing heavy costs of design and maintenance on organizations dealing with their patents.

III. NLP PREPROCESSING PIPELINE

The Patent Similarity Checker takes patent documents in common file formats like PDF, DOCX, and TXT files. The document undergoes extraction of its Abstract, Description, and Claims sections, which contain the textual information used in determining prior art, and processes them through a three-phase natural language processing pipeline before vectorization.

A. Tokenisation

The raw text is tokenized to a series of word tokens separated by white spaces and punctuations. Claim numberings, which are specific to the domain, are removed during this phase. For example, the claim phrase Method for semiconductor manufacturing comprising the steps of, for example, gets tokenized to analyzing, method, semiconductor, manufacturing, comprising, steps in the normalized corpus.

B. Stop-Word Removal

The high-frequency functional words, which do not have any semantic significance in patent document texts, are stripped out of the token sequence. The stop-word list is enhanced beyond the usual stop-words lists of English corpora (the, is, and, etc.) to include patent boilerplate words such as wherein, comprising, consisting, hereinafter, and claim, which will otherwise artificially raise similarity percentages.

C. Stemming

The morphological variations are stemmed to their root form by a Porter Stemmer. Vocabulary normalization makes sure that all inflectional or derivational variations of each word are considered a single word at the time of vectorization. Manufacturing, manufactured, manufacturer, for example, get mapped to their root base manufactur. The processed vocabulary goes directly into structure vectorization.

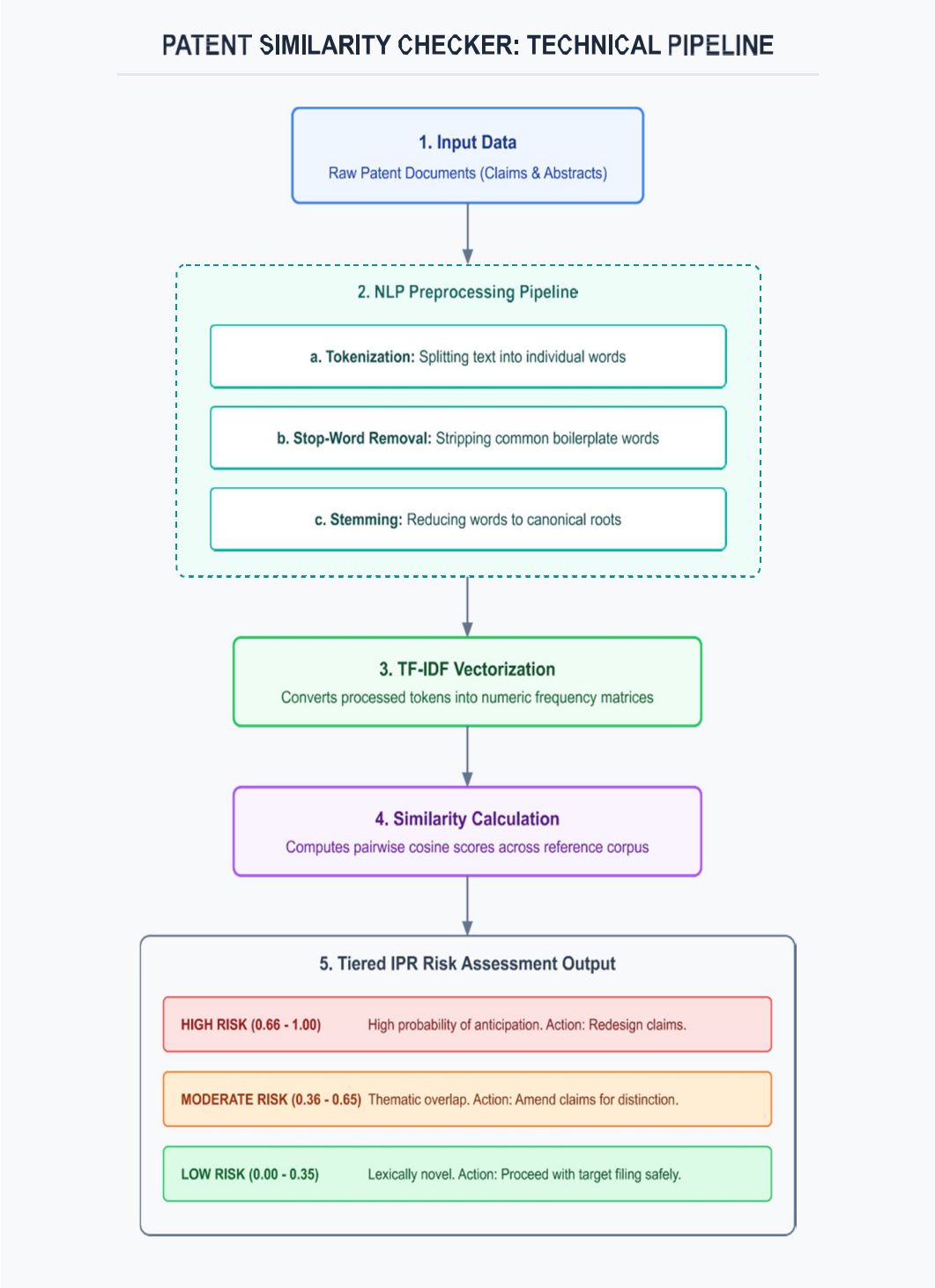


Figure 1. Technical Framework and NLP Pipeline for the Patent Similarity Checker

IV. TF-IDF VECTORISATION AND COSINE SIMILARITY

A. *Term Frequency — Inverse Document Frequency*

The vector representation of each tokenized list is obtained through TF-IDF weighting of the token lists. In other words, the weight of the term t in the document d according to a set of patent documents D is calculated as follows:

$$W(t,d) = \mathrm{TF}(t,d) \times \mathrm{IDF}(t,D) \quad (1)$$

The term frequency component calculates the prominence of a term within a document:

$$\mathrm{TF}(t,d) = \frac{f(t,d)}{\sum_{t'} f(t',d)} \tag{2}$$

where f(t, d) is the count of term t in document d, while the denominator is used to normalize the number of words in documents with different lengths. The IDF component gives a higher penalty to frequently occurring terms within the entire set of documents:

$$\mathrm{IDF}(t,D) = \log_2\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right) \tag{3}$$

When combining both components together, the TF-IDF weight matrix W becomes an |V| × |D| matrix where |V| is the vocabulary size. The vector w_a for each document is the high dimensional representation of the patent documents.

B. Pairwise Cosine Similarity

Based on the TF-IDF vectors representing a candidate patent document A and a document B, the distance between them can be measured as the cosine of the angle θ formed by them in the high-dimensional space:

$$\mathrm{Similarity}(A,B) = \cos\theta = \frac{A \cdot B}{\| A \| \times \| B \|} \tag{4}$$

Breaking down the dot product terms gives us:

$$A \cdot B = \sum_i W(t_i, A) \times W(t_i, B) \tag{5}$$

$$\| A \| = \sqrt{\sum_i W(t_i, A)^2} \tag{6}$$

The value of the metric lies between 0 and 1, indicating complete orthogonality (no common vocabulary) to identical vectors. As the TF-IDF scores cannot be negative, the cosine similarity measure for patent texts always falls within [0, 1].

C. Similarity Matrix

In a collection of n patent documents, the pairwise similarity metric is calculated for all documents, resulting in an n × n symmetric similarity matrix S with S_{ij} being Similarity(d_i, d_j).

Table I. Sample Pairwise Cosine Similarity Matrix

	Doc 1	Doc 2	Doc 3
Doc 1	1.00	0.95	0.15
Doc 2	0.95	1.00	0.15
Doc 3	0.15	0.15	1.00

As can be seen in Table I below, Doc 1 and Doc 2 have a similarity coefficient of 0.95, which implies high lexical and structural overlap – a clear indication of prior art that calls for either design around or claims distinction. Doc 3 is orthogonal to Docs 1 and 2, and thus falls in the Low Risk category.

V. IPR RISK ASSESSMENT AND LEGAL FRAMEWORK

A. Tiered Risk Classification

The cosine similarity coefficients were categorized based on the novelty and inventive step criteria used by the Indian Patent Office under the Patents Act, 1970 as follows:

TABLE II. IPR Risk Assessment Scale

Risk Level	Score Range	Legal Assessment	Recommended Action
High Risk	0.66 – 1.00	High likelihood of anticipation or obviousness; probable claim overlap	Cease filing; explore alternative designs or file a continuation application with narrower claims
Moderate Risk	0.36 – 0.65	Common themes identified; partial claim overlap may occur; innovative aspect debatable	Amend independent claims to highlight technical difference; consult examiner
Low Risk	0.00 – 0.35	Very low lexical similarity; document probably novel relative to prior art references	Continue with filing; perform selective claim drafting

B. Legal Situating under the Patents Act, 1970 and TRIPS

The definition of the invention in Section 2(1)(l) under Patents Act, 1970 states the invention not being anticipated by prior publication or prior use. Under Section 2(1)(ja), there is an inventive step which implies an advance of a technical nature which is not obvious to a person skilled in the art. In other words, cosine similarity score becomes the computational means for proving anticipation under Section 2(1)(l). The high-risk tier (≥ 0.66) implies that the candidate's patent language is substantially similar to previously submitted ones which would be considered anticipation under Section 25(1)(b). According to the terms of Article 27 of TRIPS agreement, each member shall ensure the granting of patents for inventions which are new, involve an inventive step, and are capable of industrial application. Tiered risk output meets the above-mentioned three criteria as not only lexical similarity is identified but the extent to which candidate's claims scope overlaps with those of existing patents, which can influence inventiveness. However, it should be noted that the tool does not analyze the functional equivalent of claims, claims hierarchy, nor does it consider drawings. With regard to evidentiary value: algorithmic similarity scores are now regularly used by litigants in their applications for oppositions to patents as evidentiary material for searching prior art. Even though the Indian courts have not yet made a definitive ruling on whether NLP-derived similarity scores can be admitted independently as evidence, it is settled law in India, through cases like Novartis AG vs. Union of India, that the determination of novelty must be objective and evidence-based.

VI. EXPERIMENTAL RESULTS AND DASHBOARD EVALUATION

A. IPR Prior-Art Checker — System Output

A check on the Patent Similarity Checker using the EIPR reference paper as the candidate document was made. The similarity between the documents is compared based on the highest cosine similarity score, which then determines the risk level according to Table II. Figure 2 below shows the risk assessment output displayed on the dashboard panel.

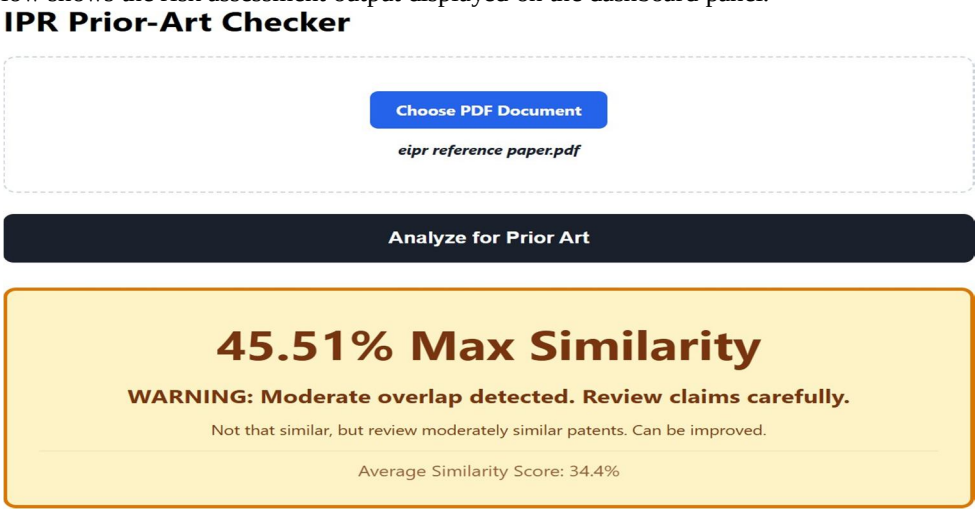


Figure 2. IPR Prior-Art Checker Dashboard Output Panel: Maximum Similarity Score and Risk Level Classification

The output from the system is a maximum similarity score of 45.51% (on the cosine measure, it translates to 0.4551) and an average similarity score of 34.4%. The cosine score of 0.4551 qualifies under the moderate risk level category, (score of 0.36-0.65), and generates a WARNING alert message: "Moderate overlap detected. Review claims carefully." The applicant would be advised to modify specific claims before acceptance rather than an automatic rejection.

B. Top Prior-Art Matches

Figure 3 below shows the top matches found by the system based on the cosine similarity score from the same submission.

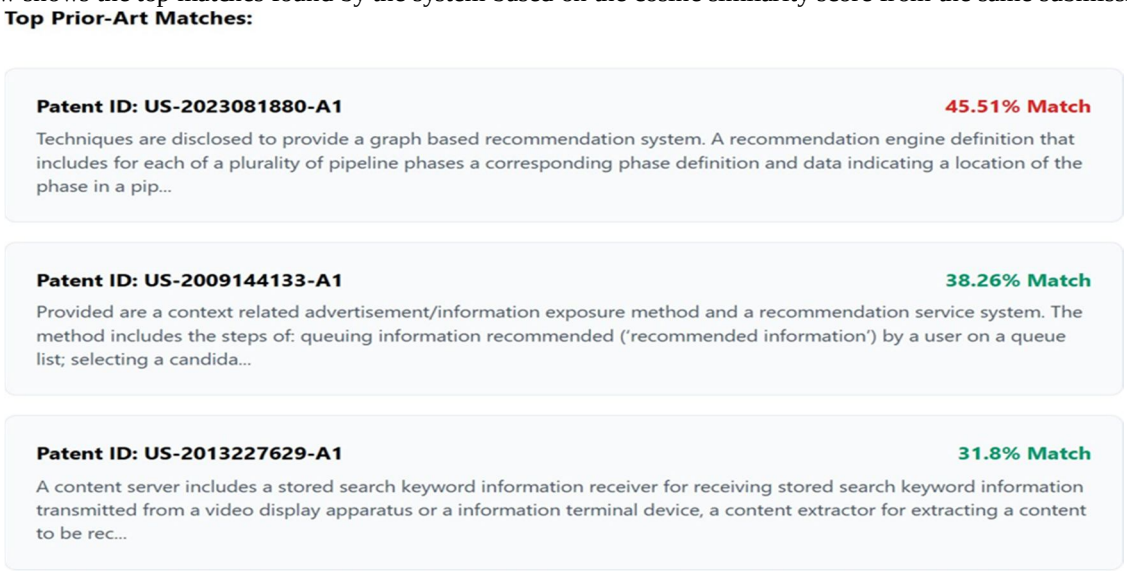


Figure 3. Dashboard Output: Top Prior-Art Matches with Cosine Scores

TABLE III. Consolidated Analysis Of Prior Art References In The Eipr Reference Document

Patent Number	Similarity Percentage	Risk Level	Suggested Course of Action
US-2023081880-A1	45.51%	Moderate	Claim amendments needed to distinguish pipeline details
US-2009144133-A1	38.26%	Moderate	Stress technical uniqueness of scoring approach
US-2013227629-A1	31.80%	Low–Moderate	Minor amendments may suffice; monitor progress
Average (all matches)	34.40%	Moderate	WARNING: Review claims before proceeding with filing

The top-scoring match (Patent US-2023081880-A1, 45.51%) describes a graph-based recommendation pipeline with defined phase structures. The theme similarity is due to the usage of common terms for the NLP pipeline process (tokenisation, vectorisation, scoring), while technical claim content varies. The second-most relevant match (US-2009144133-A1, 38.26%) is based on an advertising exposure and recommendation queuing system, showing structural similarities with regard to the candidate selection process and scoring mechanism. The third match (US-2013227629-A1, 31.80%) concerns the processing of stored search keywords; the least similar match and the only one close to the Low Risk threshold.

Patent Similarity Checker is scalable and replicable initial filter for prior art risk evaluation. Linguistic generality is the key advantage of the system, as it relies on TF-IDF vectors computed from pre-processed claim text without any need for domain-specific training dataset. The downside of such design is that there is no way to evaluate semantic functional equivalence, as two completely different sets of terms may represent identical technical feature, therefore resulting in low cosine similarity despite significant anticipation risk. Another limitation lies in inability to work with disclosure in terms of drawings, diagrams, or other technical notations that make an integral part of patent documents. Such limitations are clearly stated by the system output.

VII. CONCLUSION

The present paper has illustrated the process by which an Automated Patent Similarity Checker can be formulated and deployed as an ex-ante instrument of compliance for both the Indian and international patent systems. Through employing a deterministic pipeline consisting of tokenization, stop-word elimination, and Porter stemming along with the TF-IDF vector space model, the tool achieves a highly formalized mathematical method of detecting surface textual similarities. When benchmarked on a particular document, the system attained a maximum pairwise Cosine Similarity value of 45.51%, ultimately placing the document within the "Moderate Risk" category. The objective and replicable classification thus provides a valuable benchmark that assists inventors and corporate legal teams in their pre-patent application evaluations and alerts them to any possible problems concerning novelty as per Section 2(1)(l) of the Patents Act, 1970, before embarking upon costly and time-consuming formal proceedings.

Nonetheless, the practical application of the model equally serves to highlight essential limitations inherent to any model that only involves surface-level lexicon matching. The most significant design flaw of a sparse vector model or a simple TF-IDF system is the inability of the model to account for functional similarity semantically. As the algorithm requires exact token root overlap, two different patent papers may explain the same technology or design principle using entirely different lexicons (such as "fastening element" and "locking pin"), leading to a low cosine similarity even though the papers represent absolute anticipation.

In addition, the present system runs entirely in the textual realm, rendering it incapable of recognizing non-textual information like technical illustrations, flow diagrams, chemical formulae, or even advanced mathematical equations, which may well carry the gist of the engineering invention in question. As a result, although the Patent Similarity Checker emerges as an extremely scalable, high-speed, and impartial preliminary screening tool, it must be noted that, theoretically speaking, it is more of a decision support aid than an independent substitute for the nuanced and expert interpretation skills of a patent attorney or examiner.

VIII. FUTURE SCOPE

To overcome the structural constraints that arose during the implementation process and cater to the extremely advanced requirements of modern-day EIPR detection, there are four crucial development directions that need to be taken into consideration for future incarnations of this study:

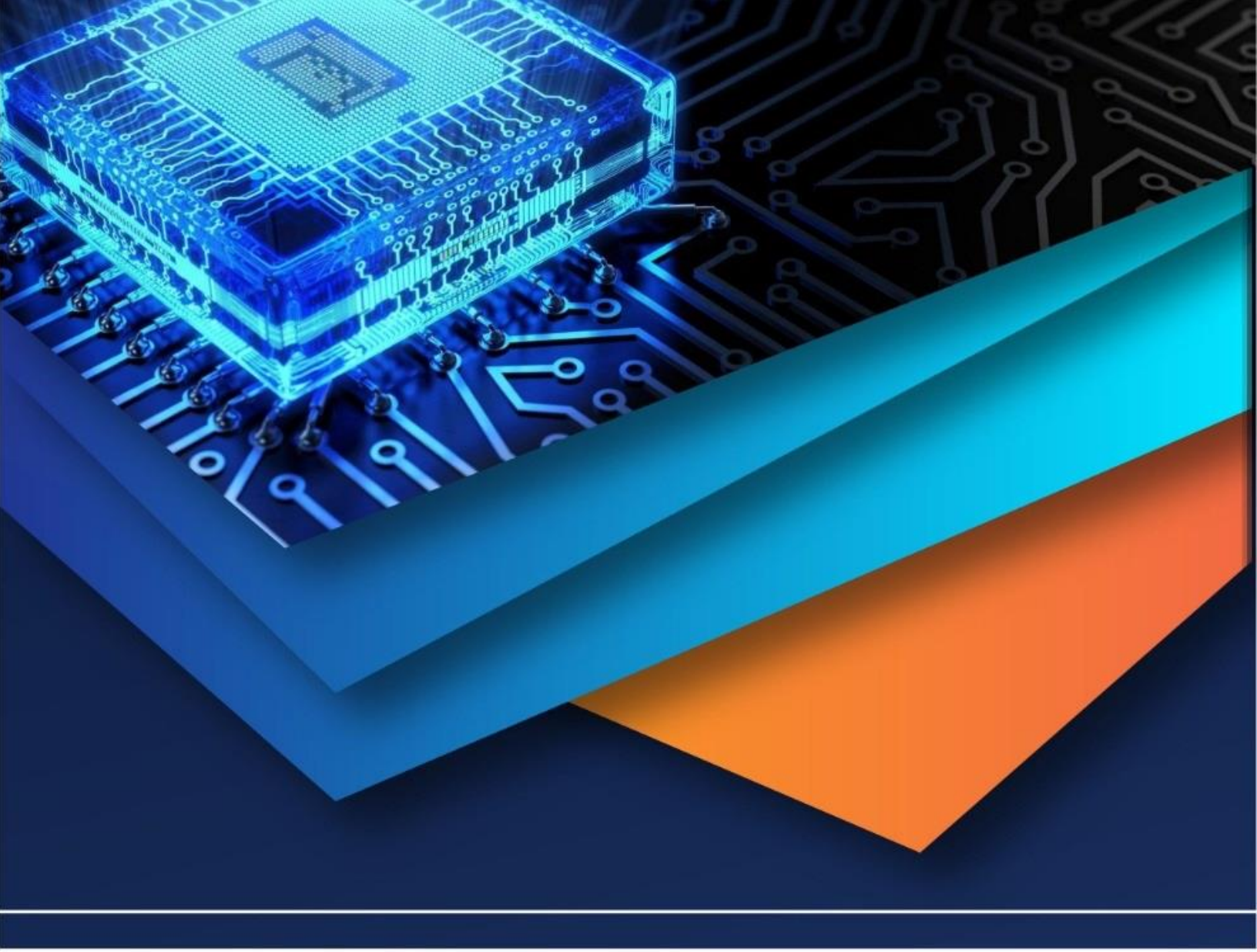
1) *Integration of Dense Deep-Learning Transformer Models:* In order to address the issue of vocabulary mismatch that plagues the sparse model paradigm, future implementations will include a similarity engine consisting of two layers of dense contextual vector representations. Through the use of specific large language models within each respective field – including PatentBERT, SciBERT, and Sentence-BERT variants (SBERT), for example – the model can transform document descriptions and independent claims into a shared low-dimensional latent space representation. This will enable the identification of semantic similarities based on concepts, rather than exact word alignment.
2) *Adversarial Structure Training and Optimization:* The evaluation module can undergo systematic reinforcement for its robustness using adversarial training schemes. Through adversarial training on rewriting current patent descriptions and claims, which include targeted paraphrasing, syntax inversion, and legal jargon obfuscation, the core verification system can be tested against the most sophisticated text modifications. This process will fine-tune the parameter settings of the system and calibrate the three-tiered risk assessment thresholds against such textual evasion tactics.
3) *Multimodal Embedding Spaces for Graphical Analysis:* Extending beyond textual limitations necessitates a shift towards an integrated multimodal alignment model. By leveraging deep vision-language models such as Contrastive Language-Image Pre-training (CLIP) and geometry-oriented GNNs, future research would seek to index and embed images such as patent illustrations, schematic diagrams, flowcharts, and sequence listings. The embedding of both visual structural information and text claims within one homogeneous embedding space would make the system capable of detecting structural/graphical plagiarism circumventing all textual detection.
4) *Analysis of Admissibility and Evidence Value within Revised Legal Framework:* On a legal front, thorough analyses would have to be carried out in order to determine the admissibility standards of NLP-generated similarity reports. In light of the recent disbandment of the Intellectual Property Appellate Board (IPAB), due to the implementation of the Tribunals Reform Act, 2021, and the consequent establishment of specialized Commercial IP Divisions in Indian High Courts, the yardstick for objective technical evidence has undergone a shift. The future analysis would delve into the capability of confidence intervals generated by algorithms and prior art detection techniques to meet the yardstick of documentary evidence laid down in Section 45 of the Indian Evidence Act, 1872.

IX. ACKNOWLEDGEMENT

The authors extend their gratitude to the Department of Computer Science and Engineering and the Department of Industrial Engineering and Management at R.V. College of Engineering, who offered them access to computing facilities among others that aided in their research. Moreover, the authors acknowledge the contribution from experts in legal and academic circles for insightful comments on analyzing patent laws and the open-source community for scikit-learn framework.

REFERENCES

[1] H. H. Shomee, A. Bhattacharjee, and T. Chakraborty, "A survey on patent analysis: From NLP to multimodal AI," in Proceedings of the Association for Computational Linguistics (ACL), 2025.
[2] L. Jiang, "Natural Language Processing in the patent domain: A survey," Artificial Intelligence Review, vol. 58, no. 3, 2025.
[3] A. Ali, M. Hussain, and S. Rahman, "Innovating patent retrieval: A comprehensive review of prior-art search techniques," AI, vol. 7, no. 5, 2024.
[4] G. S. Ascione and V. Sterzi, "A comparative analysis of embedding models for patent similarity," arXiv preprint arXiv:2403.16630, 2024.
[5] Z. Peng and Y. Yang, "Connecting the dots: Inferring patent phrase similarity with retrieved phrase graphs," arXiv preprint arXiv:2403.16265, 2024.
[6] H. Jiang, X. Wang, and J. Zhao, "Deep learning for predicting patent application outcome," Journal of Innovation & Knowledge, vol. 8, no. 2, 2023.
[7] Y. Yoo, C. Jeong, S. Gim, J. Lee, Z. Schimke, and D. Seo, "A novel patent similarity measurement methodology: Semantic distance and technological distance," arXiv preprint arXiv:2303.16767, 2023.
[8] L. Siddharth, G. Li, and J. Luo, "Enhancing patent retrieval using text and knowledge graph embeddings: A technical note," arXiv preprint arXiv:2211.01976, 2022.
[9] G. Li, L. Siddharth, and J. Luo, "Embedding knowledge graph of patent metadata to measure knowledge proximity," arXiv preprint arXiv:2211.01768, 2022.
[10] A. Trappey, C. Trappey, U. Govindarajan, and J. Sun, "Patent value analysis using deep learning models—The case of IoT technology mining for manufacturing industry," IEEE Transactions on Engineering Management, vol. 69, no. 5, pp. 2560–2572, 2022.
[11] H. Alshowaish, Y. Al-Ohali, and A. Al-Nafjan, "Trademark image similarity detection using convolutional neural networks," Applied Sciences, vol. 12, no. 3, 2022.
[12] N. Meuschke and B. Gipp, "Leveraging citation networks for patent similarity detection," in Proceedings of JCDL, IEEE, 2021.
[13] P. Morales, M. Flikkema, C. Castaldi, and A. de Man, "Patent analytics and innovation measurement using artificial intelligence," Science and Public Policy, vol. 48, no. 4, 2021.
[14] S. Reimers and I. Gurevych, "Sentence-BERT based semantic similarity models for technical and patent text," in Proceedings of EMNLP, 2021.
[15] X. Zhou, Z. Hu, and A. Lin, "Evaluation and identification of high-value patents using machine learning approaches," Journal of Informetrics, vol. 15, no. 2, 2021.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY