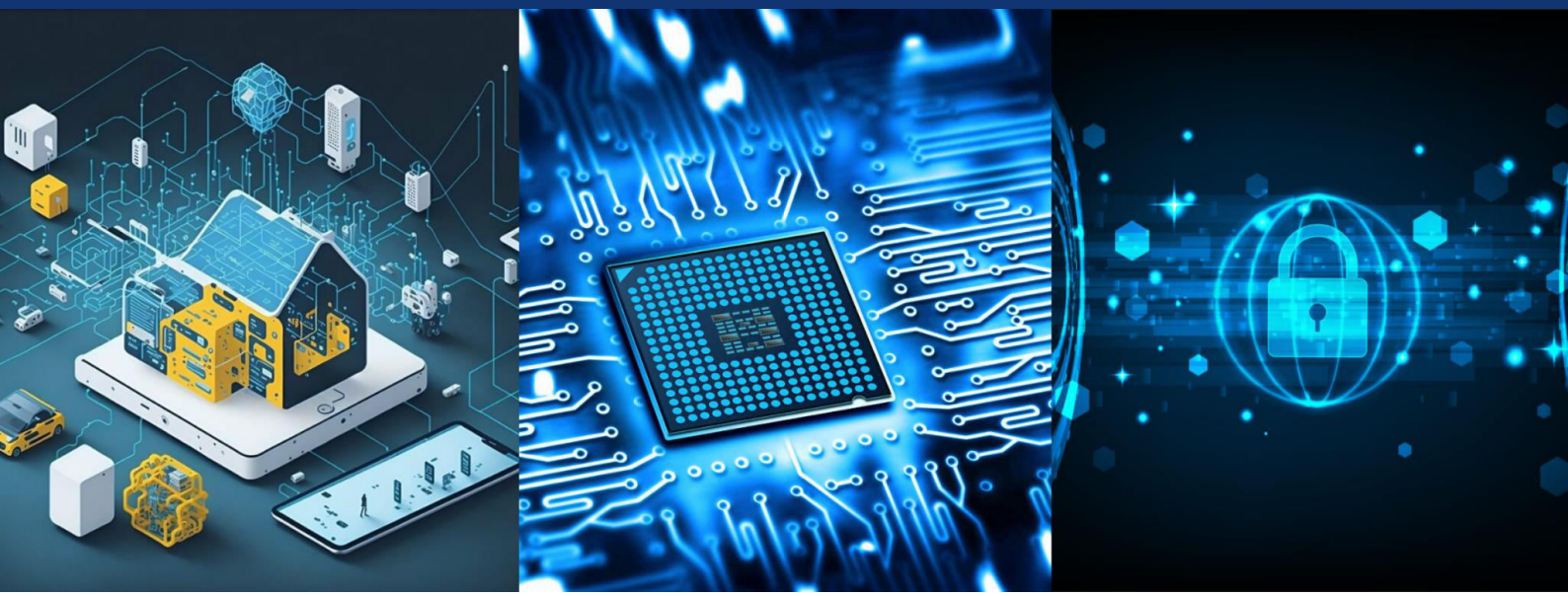
IJIRCCE

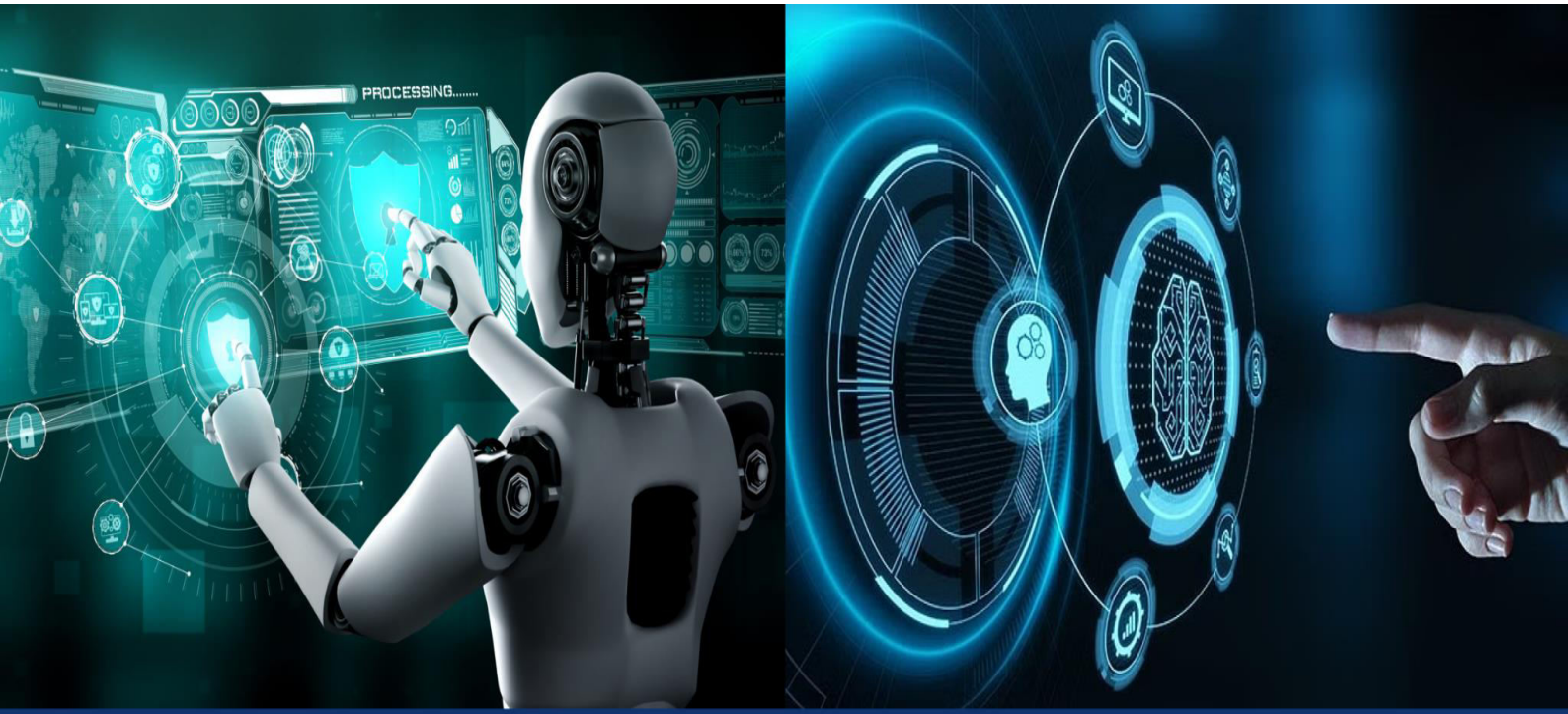
ISSN(O): 2320-9801

ISSN(P): 2320-9798



International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.771

Volume 13, Issue 4, April 2025

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Finding Similarities between Legal Case Documents

Tejal Gund, Purva Nidan, Supriya Madhawai, Anjali Wable, Prof.D.J.Perriera

Dept. of Computer Engineering, Government College of Engineering and Research, Avasari, Pune, India

ABSTRACT: Legal research often involves the difficult task of finding relevant case precedents, which can be complicated by the intricate, extensive, and unstructured nature of legal texts. This paper introduces a hybrid method for identifying similarities among legal case documents by combining TF-IDF vectorization with BERT-based embeddings. While TF-IDF focuses on termbased lexical similarities, BERT improves contextual understanding, allowing for more precise semantic comparisons. By merging these natural language processing (NLP) techniques, our system enhances the efficiency and accuracy of legal document retrieval. This approach aids lawyers and legal researchers in swiftly locating pertinent case precedents, thus streamlining the legal research process.

KEYWORDS: Legal document similarity, hybrid similarity, TF-IDF, BERT, NLP, case law retrieval.

I. INTRODUCTION

The increasing volume of legal case documents has made it more challenging for legal professionals to efficiently locate relevant precedents. The complexity, length, and specialized terminology found in these documents require advanced techniques for similarity detection to aid legal research. Traditional methods, such as keyword-based searches, often overlook the deeper semantic connections between cases, leading to incomplete or inaccurate retrieval results. As legal databases expand, there is a growing need for intelligent systems that can effectively analyze and compare legal documents.

This study presents a hybrid similarity approach that integrates TF-IDF vectorization with BERT-based embeddings to enhance the detection of similarities in legal documents.

A. TF-IDF Vectorization:

- Offers a statistical measure of term significance within a document.
- Enables effective lexical matching by pinpointing important words based on their frequency in a specific document compared to the entire corpus.

B. BERT-based Embeddings:

- Employs deep contextual embeddings that capture the meaning of words in their context.
- Improves semantic similarity detection by considering how words relate to one another within the broader context of the document.

By combining these techniques, our system aims to enhance the accuracy and efficiency of retrieving relevant legal case documents, providing a more thorough approach to case similarity detection.

Identifying similarities in legal cases is crucial for various legal functions, including:

- Precedent Analysis: Finding relevant past cases that inform legal arguments.
- Case Prediction: Estimating the likely outcomes of future cases based on historical data.
- Legal Recommendation Systems: Suggesting similar cases to lawyers and judges for better decision-making.

Efficient retrieval of similar cases can assist lawyers in constructing strong arguments, ensuring consistency in legal decisions, and reducing research time. However, the challenges posed by the unstructured nature of legal documents remain significant.

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

A. Finding Similar Legal Judgements Under the Common Law System

Authors: Sushanta Kumar, P. Krishna Reddy, V. Balakista Reddy, Malti Suri

Objective: This study aims to improve legal case similarity detection by using "paragraph-links" (PLs), which link paragraphs with similar content across judgments, enhancing traditional citation-based methods.

The approach combines link-based similarity with paragraph-level links. Judgements are segmented into paragraphs, tokenized, and vectorized using TF-IDF. Cosine similarity is applied between paragraphs to establish PLs, which are then combined with citation-based bibliographic coupling for similarity detection.Findings: The study found that using citations alone identified 62 similar judgments, while PLs improved detection to 145. The combined approach detected 280 similar cases, aligning well with expert evaluations.Limitations and Gaps: The dataset is limited to pre-1993 cases, which may not represent modern legal complexities. The approach is computationally intensive due to pairwise comparisons and assumes that each paragraph represents a distinct legal concept.Potential Improvements: Applying advanced NLP models, such as BERT, and expanding the dataset to broader jurisdictions could enhance accuracy and real-time capability. [1]

B. Measuring Similarity among Legal Court Case Documents

Authors: Arpan Mandal, Raktim Chaki, Sarbajit Saha, Kripabandhu Ghosh, Arindam Pal, Saptarshi Ghosh

Objective: The study compares advanced text-based methodologies, particularly embedding-based models like Word2Vec and Doc2Vec, with traditional approaches to improve legal document similarity detection.

Methodology: The approach focuses on text-based methods due to citation network sparsity. Various document representations, such as the entire document, summaries, paragraphs, and citation-specific sections, are evaluated. Multiple similarity measures, including TF-IDF, LDA, Word2Vec, and Doc2Vec, are used, with cosine similarity applied for document comparison.

Findings: Doc2Vec applied to the whole document showed the highest alignment with expert evaluations, outperforming baseline text-based and network-based hybrid models.

Limitations and Gaps: Performance depends on document length and complexity. The dataset is limited to Indian jurisdiction, affecting generalizability.

Potential Improvements: Combining citation information with embedding-based models and testing on datasets from diverse jurisdictions could enhance performance. [2]

C. A Text Similarity Approach for Precedence Retrieval from Legal Documents

Authors: D. Thenmozhi, Kawshik Kannan, Chandrabose Aravindan

Objective: The study aims to develop a text similarity approach for retrieving relevant prior cases based on similarity scores with a current case document.

Methodology: Three methods were developed: (1) TF-IDFbased similarity using only concepts (nouns), (2) an extended method incorporating relations (verbs), and (3) a Word2Vec embedding-based method combining concepts and relations. Cosine similarity was used for ranking prior cases.

Findings: Method-2, which includes concepts and relations with TF-IDF, outperformed others in Mean Average Precision (MAP) and Mean Reciprocal Rank (MRR). However, Method1 had better precision and recall.

Limitations and Gaps: Method-3 underperformed due to averaging vectors, suggesting that alternatives like Doc2Vec could enhance results. Only lexical features were considered, limiting semantic and contextual analysis.

Potential Improvements: Implementing Doc2Vec and incorporating additional contextual features could improve retrieval accuracy for legal precedence. [3]

D. Legal Document Similarity Matching Based on Ensemble Learning

Authors: Aman Fan, Shaoxi Wang, and Yanchuan Wang

Objective: The study improves legal case similarity prediction by utilizing ensemble learning, which merges representation-based and interaction-based text matching techniques.

Methodology: The model comprises two sub-networks: (1) A feature representation sub-network using BERT with contrastive learning to minimize the distance between similar cases and increase the distance between dissimilar ones. (2) A binary classification sub-network utilizing BERT with shared weights to assess case similarity through a binary classification framework.

The ensemble learning approach integrates the outputs from both sub-networks to enhance prediction accuracy. Cosine similarity and binary classification probabilities are used for similarity assessment. Data augmentation techniques include transformations based on commutativity, anti-symmetry, and reflexivity, as well as sampling from the latter sections of case documents to emphasize critical information.

Findings: The ensemble model achieved an accuracy of 74.53% on the CAIL2019-SCM dataset, surpassing existing methods. The contrastive learning network was particularly effective in distinguishing dissimilar cases, while the binary classification network bolstered prediction reliability. Only 13.54% of errors occurred in both sub-networks simultaneously, highlighting their complementary nature.

Limitations and Gaps: The models have limited use of prior legal knowledge and structured legal components. Performance enhancements are required to tackle complex legal reasoning tasks.

Potential Improvements: Integrating legal knowledge extraction, employing techniques like Doc2Vec, and expanding contrastive learning could improve results. Strategies to mitigate overfitting through diverse data augmentation methods should be explored. [4]

E. Analyzing Similarities Between Legal Court Documents Using NLP Approaches Based on Transformers

Authors: Raphael Souza de Oliveira, Erick Giovani Sperandio Nascimento

Objective: The study explores transformer-based models (BERT, GPT-2, RoBERTa) for clustering legal documents from the Brazilian judiciary based on their similarities.

Methodology: The dataset consists of 210,000 legal proceedings from the Brazilian Labour Court system. Several models were used: (1) Fine-tuned versions of BERT trained on Brazilian legal corpora. (2) GPT-2 and RoBERTa pre-trained on general Brazilian Portuguese corpora, further specialized for legal texts.

Data extraction and cleaning were performed using regular expressions, followed by word embeddings generation using BERT, GPT-2, and RoBERTa. Document embeddings were computed by averaging word embeddings with TF-IDF weighting. Clustering was conducted using k-means, with cosine similarity for evaluation, and dimensionality reduction via t-SNE for visualization.

Findings: RoBERTa achieved the best results in clustering quality and processing speed. Specialized models trained on legal-specific data underperformed compared to generalpurpose ones due to smaller training corpora. The study found that RoBERTa pt-BR processed 55.31 documents per minute, while GPT-2 pt-BR achieved 29.40 documents per minute. Models effectively clustered cases with high cosine similarity scores, improving legal precedent identification.

Limitations and Gaps: Specialized models trained on legal data did not outperform general-purpose ones, possibly due to smaller corpus size. A larger, more diverse legal dataset is needed for improved model specialization.

Potential Improvements: Creating second-level foundation models focused on legal language could enhance results. Further refining specialized models using larger legal corpora could better capture the nuances of legal text. Future work may explore generating decision drafts and automatic classification using these embeddings. [5]

F. Legal Case Document Similarity: You Need Both Network and Text

Authors: Paheli Bhattacharya, Kripabandhu Ghosh, Arindam Pal, Saptarshi Ghosh

Objective: Determining the similarity between two legal documents is essential for tasks such as retrieving prior cases and recommending citations. This research investigates both textual content and citation network structures to enhance similarity assessment.

Methodology: The study utilizes case documents from the Indian judiciary, with validation and test datasets annotated by experts from RGSOIPL and WBNUJS, India. Two models are used: PCNet (Precedent Citation Network), which represents case citations as a network, and Hier-SPCNet (Hierarchical Statute + Precedent Citation Network), an enhanced version incorporating statutes.

The processing pipeline involves constructing citation networks and extracting legal citations using regular expressions. Node embeddings are generated through Node2Vec and Metapath2Vec algorithms, followed by merging network-based and text-based embeddings to refine similarity estimation. The models are evaluated using Pearson correlation metrics.

Findings: The hybrid approach combining textual and network-based embeddings improved the accuracy of similarity detection compared to using text-based methods alone.

Limitations and Gaps: The study is limited to Indian legal documents, and network structures may not always capture legal nuances. Future improvements could involve expanding datasets to include legal cases from multiple jurisdictions and refining embedding techniques to incorporate more contextual legal knowledge. [6]

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

III. CRITICAL ISSUES

A. Challenges in Existing Approaches

1. Computational Complexity and Scalability: Transformer-based models such as BERT, RoBERTa, and GPT-2 require extensive computational resources for both training and inference, making it challenging to process large-scale legal document datasets efficiently. [3]

Legal citation networks, which are often sparse and contain large numbers of disconnected nodes, make it difficult for citation-based approaches to scale effectively when handling thousands of legal documents. [6]

2. Limitations of Text-Based Methods: Traditional textbased similarity techniques such as TF-IDF, Word2Vec, and Doc2Vec often fail to accurately capture the complex structure, semantics, and contextual meaning inherent in legal texts, leading to misleading similarity assessments. [1]

Standard natural language processing (NLP) models struggle with jurisdictional variations in legal language, as well as with the dynamic evolution of legal terminologies and interpretations, reducing their effectiveness in cross-jurisdictional applications. [4]

3. Limitations of Citation-Based Approaches: Citationbased networks, such as the Precedent Citation Network (PCNet), primarily focus on linking prior cases while often neglecting legal statutes and legislative provisions, resulting in incomplete similarity evaluations. [6]

Basic similarity metrics used in citation-based methods, including bibliographic coupling and co-citation analysis, fail to capture the nuanced relationships and dependencies that exist between different legal cases, leading to inaccurate similarity estimations. [6]

4. Challenges in Combining Textual and Network-Based Information: Although hybrid approaches that integrate textual similarity with citation networks have demonstrated improvements in legal document retrieval tasks, there is currently no standardized or widely accepted methodology for effectively fusing these two types of information. [1]

Many hybrid models tend to overestimate similarity scores due to the uniform weighting of citations, which does not adequately reflect the varying degrees of legal significance that different citations hold in judicial reasoning. [1]

Advanced fusion techniques, such as contrastive learning and multi-modal embeddings, are needed to achieve a balanced and context-aware integration of textual and citation-based legal similarity features. [3]

5. Accuracy and Generalization Issues: Legal language exhibits significant variations across different jurisdictions, making it difficult to develop models that generalize well across multiple legal systems without extensive domain-specific adaptations. [6]

Representation and interaction-based similarity models often show inconsistent performance across different legal datasets, limiting their ability to be applied universally across diverse legal corpora. [1]

Hierarchical citation models, such as Hier-SPCNet, frequently overestimate the similarity between documents when commonly cited legal statutes and constitutional provisions are involved, leading to inaccurate similarity rankings. [6]

B. Data Challenges

1.Imbalanced Datasets: Legal datasets often exhibit an uneven distribution of case types, where civil cases are significantly more prevalent than specialized legal domains such as tax law, intellectual property law, or criminal appeals, leading to biased predictions. [6]

Due to this imbalance, models trained on such datasets may overfit to dominant case types, which can result in high false positive rates for commonly occurring case categories while underperforming in less frequent legal domains. [5]

2.Limited Availability of Expert-Annotated Data: Highquality, expert-annotated legal datasets are scarce, as the process of labeling legal texts requires domain expertise and is both time-consuming and costly, making it difficult to build large-scale supervised learning models. [3]

Many existing legal datasets are jurisdiction-specific, which further limits their ability to be applied in cross-border legal research, as laws, precedents, and legal reasoning structures vary significantly across different legal systems. [4]

3.Handling Citations and References: Citation-based similarity detection is an essential component of legal document retrieval, but not all legal citations hold the same level of importance in judicial decision-making, making it challenging to weigh citations appropriately.

Frequently cited constitutional provisions and broad legal principles can distort similarity scores, leading to an overestimation of similarity between cases that reference the same general laws but differ in their specific legal context. [6]

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. Training Challenges

1. Computational Complexity and Resource Constraints: Transformer-based models, such as BERT and its legal domain variants (e.g., Legal-BERT), impose significant memory and processing requirements, making them unsuitable for realtime legal research or deployment in resource-constrained environments. [4]

Legal documents are often lengthy, and models like BERT, which have a token limit of 512, require long texts to be truncated or split into smaller sections, which can lead to a loss of critical contextual information necessary for accurate similarity detection. [1]

Fine-tuning large-scale transformer models for legal text similarity requires specialized hardware, such as high-end GPUs or TPUs, increasing the cost and accessibility barriers for legal institutions with limited computational resources. [4]

2. Hyperparameter Optimization Difficulties: Optimizing deep learning models for legal document similarity requires careful tuning of multiple hyperparameters, such as learning rate, batch size, and model depth, which makes the training process time-consuming and computationally expensive. [1] Domain-specific pretraining of language models on legal corpora introduces additional complexity, as it requires extensive trial-and-error experimentation to identify the most effective training configurations for different types of legal documents. [1]

3. Overfitting and Generalization Issues: Legal NLP models often suffer from overfitting, particularly when trained on specialized datasets from a single jurisdiction, making them ineffective for handling legal cases from different regions or legal traditions. [6]

Although regularization techniques such as dropout, weight decay, and data augmentation can help mitigate overfitting, they also increase the complexity of the training process and require extensive hyperparameter tuning to achieve optimal results. [1]

D. Accuracy and Reliability Issues

1) Balancing False Positives and False Negatives: Overreliance on text similarity measures can result in erroneous case retrieval, where legal cases that contain similar keywords or phrases are mistakenly identified as similar, despite addressing fundamentally different legal issues. [6]

Citation-based approaches, on the other hand, may fail to detect meaningful similarities between cases that reference different precedents but share common underlying legal principles, leading to a high rate of false negatives. [3]

Existing evaluation methodologies for legal document similarity lack consistency and robustness across different legal systems, making it difficult to ensure reliable performance in cross-jurisdictional legal research applications. [6]

IV. METHODOLOGY

A. Data Collection

The dataset used in this research consists of a total of 3,111 documents, which are categorized into case documents and statutory documents. These documents are sourced from the AILA (Artificial Intelligence for Legal Assistance) dataset available on Kaggle, and include various document types such as case laws, statutes, and legal articles.

The dataset can be broken down as follows: [6]

• Case documents in the directory Object_casedocs. The AILA dataset contains the following annotations and features:

Document Types: Case laws, statutes, and legal articles. Annotations: Titles, authors, dates, and legal citations.

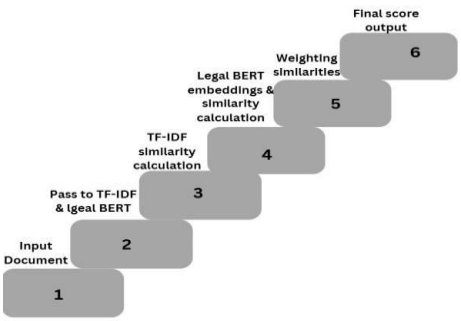


Fig. 1. Methodology

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

Categories: The dataset covers various legal categories, including criminal, civil, and constitutional law.
Language: The documents are in English.
Jurisdiction: The dataset primarily focuses on the Indian legal system.
Usage: This dataset is particularly useful for legal research, classification, and named entity recognition in legal AI projects.
The features in the dataset are as follows:
Case Documents (from Object_casedocs):
Case Text: Complete text of each legal case.Case Title/ID: Unique identifiers or titles for each case.Key Entities and Terms: Mentions of important names, legal terms, or entities found within cases.

B. Preprocessing

Preprocessing is a crucial step in preparing legal case documents and queries for similarity analysis. The following steps were followed to ensure uniformity and consistency in the text:
.Metadata Removal: Legal case documents often contain irrelevant metadata, such as court names, case identifiers, legal roles, date references, and specific legal terms, which do not contribute to the semantic similarity analysis. These elements are removed through the use of regex-based filtering. The following are examples of metadata elements that are removed:

- Court names (e.g., Supreme Court, High Court)
- Case identifiers (e.g., Writ Petition No. 1234)
- Legal roles (e.g., Petitioner, Respondent)
- Dates (e.g., 12 March 2021)

2.Text Cleaning: The text cleaning process includes several steps to ensure uniformity and consistency in the document text:

- Convert the text to lowercase to ensure uniformity.
- Remove non-alphanumeric characters (special symbols, punctuation) and numerical values, as these do not affect the semantic content of the document.

3.Tokenization: Tokenization is the process of splitting a document or query into individual words. In this case, we used NLTK's word tokenizer to perform efficient text processing. Tokenization breaks down the text into smaller units (tokens), making it easier to analyze and process the document.

4.Stopword Removal: Stopwords are common words in the English language, such as the, and, is, etc., which are removed during preprocessing. These stopwords are removed using NLTK's stopword list and Scikit-learn's ENGLISH_STOP_WORDS. Removing stopwords helps eliminate high-frequency words that do not contribute to the meaning of the document.

5.Lemmatization: Lemmatization is the process of reducing words to their base form. For example, running is reduced to run, and judgments is reduced to judgment. This helps in normalizing word variations, making feature extraction more effective. We used WordNet Lemmatizer for lemmatization in this process.

6.Query Preprocessing: Queries underwent the same preprocessing steps as the case documents. The query text was cleaned, tokenized, and lemmatized to ensure consistency between the case documents and queries. This ensures that the documents and queries are on the same level for similarity matching.

C. Passing Data through TF-IDF and Legal BERT

After preprocessing the legal documents, the next step is to convert them into numerical features suitable for similarity comparison. In this study, two feature extraction methods are employed: TF-IDF and Legal BERT.
TF-IDF Vectorization:
TF-IDF (Term Frequency-Inverse Document Frequency) transforms text into numerical vectors that reflect the importance of terms within a document relative to a corpus. It has two components:

1) Term Frequency (TF): Measures how often a term appears in a document:

$$TF(t,d) = \frac{\text{Frequency of term t in document d}}{\text{Total terms in document d}}$$

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

2) Inverse Document Frequency (IDF): Measures the importance of a term across the corpus:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right)$$

Where N is the total number of documents in the corpus and df(t) is the number of documents containing the term t.
The final **TF-IDF** score for a term t in document d is:

$$TF - IDF(t,d) = TF(t,d) \times IDF(t)$$

This creates a sparse vector representation, with each term weighted according to its importance.

Legal BERT Embedding:
Legal BERT generates dense, continuous vector representations (embeddings) of legal documents. Unlike TF-IDF, which captures term importance, Legal BERT encodes contextual meaning by considering the surrounding words in the document. The embedding for a document S is represented as:

$$Embed(S) = \mathbf{v}_S$$

These embeddings capture the semantic context of legal terminology, enabling more accurate document similarity detection.

D. TF-IDF Similarity Calculation

After transforming legal documents into TF-IDF vectors, the next step is to compute the similarity between them. Cosine similarity is commonly used for this purpose as it effectively measures the angle between two document vectors, providing a normalized similarity score.
Cosine Similarity Computation:
Cosine similarity between two TF-IDF vectors **A** and **B** is computed as:

$$\mathbf{A} \cdot \mathbf{B} \cos(\theta) = \underline{\|\mathbf{A}\|\|\mathbf{B}\|}$$

where:
- $\mathbf{A} \cdot \mathbf{B}$ is the dot product of the two vectors,
- $\|\mathbf{A}\|$ and $\|\mathbf{B}\|$ are the Euclidean norms of the vectors.

The cosine similarity score ranges from 0 to 1, where:
1 indicates identical documents, 0 indicates no similarity.
Application in Legal Document Retrieval: By computing cosine similarity between a query document and case documents, the most relevant legal cases can be retrieved based on their textual similarity. This method is computationally efficient and interpretable, making it well-suited for large-scale legal text analysis.

E. Legal BERT Embedding and Similarity Calculation

To improve the semantic understanding of legal text, Legal BERT is used to generate dense vector representations of legal documents. Unlike TF-IDF, which relies on term frequency, Legal BERT captures the contextual meaning of words based on their usage in legal cases.

Legal BERT Embedding Generation: Legal BERT,a domain-specific adaptation of BERT, transforms legal text into numerical embeddings. These embeddings encode the syntactic and semantic structure of legal documents, enabling more accurate similarity detection.

Semantic Similarity Computation: Once documents are converted into Legal BERT embeddings, their similarity is computed using vector-based similarity metrics. This allows for a more refined comparison by capturing contextual relationships between legal terms and phrases. Euclidean distance was used to assess the similarity between BERT-generated embeddings of legal case documents. This metric computes the straightline distance between two high-dimensional vectors, with a smaller distance signifying greater similarity. The formula for Euclidean distance is:

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

$$d(A,B) = \sqrt{\sum_{i=1}^{n}(A_i - B_i)^2}$$

where A and B are the embedding vectors of two legal documents, and n is the dimensionality of the embeddings.

Application in Legal Document Retrieval: Legal BERTbased similarity computation improves case retrieval by understanding legal terminology in context, reducing the limitations of keyword-based methods. This is particularly useful for handling variations in legal language and finding relevant precedents more effectively.

F. Weighting Similarity

To combine the strengths of TF-IDF vectorization and Legal BERT embeddings, a weighted hybrid similarity approach is utilized. This technique assigns weights to each similarity measure based on its variance, allowing for a balanced integration of both lexical and semantic features.

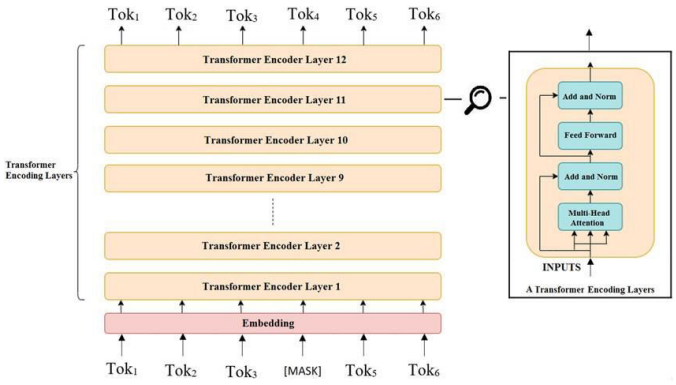


Fig. 2. BERT Model Architecture

1. Dynamic Weighting Approach: The hybrid similarity score is computed as a weighted sum of TF-IDF cosine similarity and similarity based on Legal BERT (assessed using Euclidean distance). The weight assignment is determined by the variance of each similarity measure:

$$\alpha = \frac{\sigma^2_{\text{TF-IDF}}}{\sigma^2_{\text{TF-IDF}} + \sigma^2_{\text{BERT}}} \qquad (1)$$

where $\sigma_{TF\text{-}IDF}^2$ represents the variance of TF-IDF similarity scores, σ_{BERT}^2 indicates the variance of BERT similarity scores, and α dynamically adjusts based on the relative contributions of each similarity measure.

With this adaptive weighting, the final hybrid similarity score is determined as follows:

$$S\text{hybrid} = \alpha \cdot S\text{TF-IDF} + (1 - \alpha) \cdot S\text{BERT} \qquad (2)$$

where $S_{TF\text{-}IDF}$ denotes the cosine similarity computed from TF-IDF vectors, S_{BERT} signifies the similarity derived from Legal BERT embeddings, and S_{hybrid} is the resulting hybrid similarity score.

2) Justification for Weighted Similarity: TF-IDF similarity focuses on lexical similarity by capturing word frequency and significance but lacks contextual understanding. In contrast, Legal BERT embeddings capture the semantic meaning of legal text, understanding relationships between terms beyond direct word matches. The variance-based weighting ensures that the similarity measure with more reliable variance has a greater impact, making the similarity

© 2025 IJIRCCE | Volume 13, Issue 4, April 2025| DOI:10.15680/IJIRCCE.2025.1304217

www.ijircce.com | e-ISSN: 2320-9801, p-ISSN: 2320-9798| Impact Factor: 8.771| ESTD Year: 2013|

International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

computation more robust and adaptive to different datasets. This weighting mechanism allows the system to leverage both traditional lexical-based similarity and contextual embeddings to enhance legal case similarity detection.

G. Results

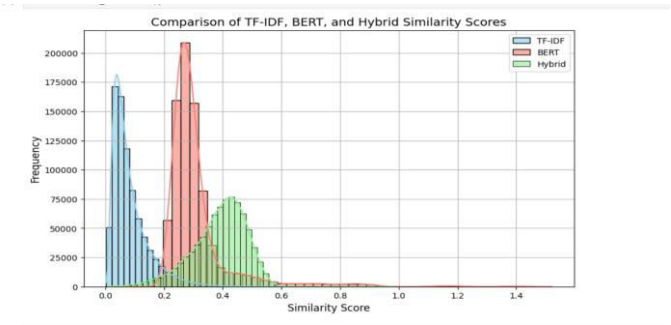


Fig. 3. Comparison

1) Comparison of Similarity Score Distributions for TFIDF, LegalBERT, and Hybrid Approaches: The comparison of similarity score distributions for TF-IDF, LegalBERT, and Hybrid Approaches reveals distinct characteristics. The TFIDF distribution exhibits a sharp peak at low similarity values, suggesting it mainly captures surface-level lexical overlaps without a deep semantic understanding. In contrast, LegalBERT displays a broader range of similarity scores, highlighting its capability to identify semantic similarities among legal documents. The Hybrid model smooths out the distribution, indicating that the combination of TF-IDF and LegalBERT results in a more balanced similarity measure that incorporates both lexical and semantic features.

Relying solely on TF-IDF may lead to false negatives, as it fails to capture the deeper relationships between legal texts. While LegalBERT enhances the detection of semantic similarities, its dependence on embeddings can result in higher computational costs. The Hybrid approach capitalizes on the strengths of both methods, offering a more dependable similarity measure for legal case retrieval.Future research could aim to dynamically fine-tune the weight between TF-IDF and BERT based on the complexity of cases. Furthermore, assessing the effectiveness of hybrid models against human-labeled legal case similarities could yield valuable insights. Exploring alternative distance metrics, such as Euclidean or Cosine similarity for BERT-based embeddings, might also improve performance.Key Takeaway: The hybrid approach successfully merges the strengths of TF-IDF and LegalBERT, resulting in a more balanced similarity distribution. This indicates that utilizing both lexical and semantic features can enhance the retrieval of similar legal cases, thereby improving the effectiveness of AI-driven legal research tools.

TABLE I

Query	Method	Top 1 Case (Score)	Top 2 Case (Score)	Top 3 Case (Score)
AILA Q1	TF-IDF	C2257 (0.4018)	C1699 (0.3670)	C1085 (0.3471)
	BERT	C347 (0.6261)	C516 (0.6184)	C1446 (0.6162)
	Hybrid	C2257 (0.5073)	C1874 (0.4892)	C1699 (0.4813)
AILA Q2	TF-IDF	C784 (0.3900)	C596 (0.3883)	C1136 (0.3773)
	BERT	C1377 (0.5814)	C1575 (0.5671)	C1616 (0.5566)
	Hybrid	C1616 (0.4226)	C1825 (0.3973)	C2067 (0.3939)
AILA Q15	TF-IDF	C1601 (0.3307)	C737 (0.3247)	C332 (0.3175)
	BERT	C798 (0.5959)	C1616 (0.5941)	C1575 (0.5787)
	Hybrid	C2067 (0.3934)	C1616 (0.3794)	C737 (0.3704)

COMPARISON OF TF-IDF, BERT, AND HYBRID METHODS FOR LEGAL CASE SIMILARITY

2) Observations from Similarity Score Distributions :

A. Performance of TF-IDF TF-IDF mainly identifies similar cases through lexical overlap. The similarity scores are generally lower than those of LegalBERT, with the highest score reaching 0.4018. Since TF-IDF does not account for semantic relationships, it may result in false negatives, missing conceptually similar cases that do not use the same wording.

B. Performance of LegalBERT LegalBERT, which utilizes contextual embeddings, yields significantly higher similarity scores, with a peak score of 0.6261. It effectively captures deeper semantic meanings, making it a more reliable method for retrieving legal cases. However, its dependence on dense embeddings results in higher computational costs compared to TF-IDF.

C. Performance of the Hybrid Approach The Hybrid model strikes a balance between lexical and semantic similarity by integrating TF-IDF and LegalBERT. It achieves similarity scores that are higher than TF-IDF but lower than LegalBERT, with a maximum score of 0.5073. This method improves retrieval accuracy by combining surface-level word matching with a deeper contextual understanding.

SUMMARY OF KEY FINDINGS

The findings indicate that the Hybrid approach is more effective than TF-IDF in identifying similarities, all while being less computationally intensive than using LegalBERT by itself. This highlights the significance of combining both lexical and semantic similarities to improve AI-powered legal research tools.

REFERENCES

[1] S. Kumar, P. Reddy, V. Reddy, " Finding similar Legal Judgements under common law system", Researchgate.net2013

[2] Arpan Mandal, Raktim Chaki, Sarbajit Saha,"Measuring Similarity among Legal Court Case Documents" ,ACMCOMPUTE2017

[3] D. Thenmozhi, K. Kannan, and C. Aravindan, "A Text Similarity Approach for Precedence Retrieval from Legal Documents," in Proceedings of FIRE 2017 – Forum for Information Retrieval Evaluation, 2017.

[4] Aman Fan, Shaoxi Wang, and Yanchuan Wang, "Legal Document Similarity Matching Based on Ensemble Learning," IEEE Access, 2024.

[5] R. S. de Oliveira and E. G. S. Nascimento, "Analysing Similarities Between Legal Court Documents Using Natural Language Processing Approaches Based on Transformers," arXiv preprint arXiv:2304.07182, 2023.

[6] P. Bhattacharya, K. Ghosh, A. Pal, and S. Ghosh, "Legal Case Document Similarity: You Need Both Network and Text," arXiv preprint arXiv:2205.14431, 2022.

[7] I. Chalkidis, M. Fergadiotis, P. Malakasiotis, N. Aletras, and I. Androutsopoulos, "LEGAL-BERT: The Muppets Straight Out of Law School," arXiv preprint arXiv:2010.02559, Oct. 2020.

[8] Karandikar, A. S. (2024). Building a highly resilient system for processing billions of events daily. International Journal of Research in Computer Applications and Information Technology (IJRCAIT), 7(2), 603-614.

[9] K. Sugathadasa, B. Ayesha, N. de Silva, A. S. Perera, V. Jayawardana, D. Lakmal, and M. Perera, "Legal Document Retrieval Using Document Vector Embeddings and Deep Learning," arXiv preprint arXiv:1805.10685, May 2018.

[10] S. Bithel and S. S. Malagi, "Unsupervised Identification of Relevant Prior Cases," arXiv preprint arXiv:2107.09207, Jul. 2021.

[11] D. Oniani, "Cosine Similarity and Its Applications in the Domains of Artificial Intelligence," May 14, 2020.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi crossref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details