Global Journal of Engineering Innovations & Interdisciplinary Research

Original Article

Legal Clause Similarity Detection and Contract Section Matching using TF-IDF

Siva Sankar[1], Setty Sowmya[2], Shaik Sufiya[2], Vellanki Jahnavi[2], Yammani Bhavyonnathi[2]

[1]Assistant Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India
[2]B.Tech Student, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India

Correspondence

Siva Sankar

Assistant Professor, Department of CSE, KKR & KSR Institute of Technology and Sciences, Guntur, Andhra Pradesh, India

- Received Date: 08 Jan 2026
- Accepted Date: 25 Jan 2026
- Publication Date: 15 Mar 2026

Keywords

Legal document analysis, clause similarity de¬tection, contract section matching, TF-IDF, logistic regression, cosine similarity, natural language processing..

Abstract

Legal contracts represent structured agreements that specify the rights, duties, and responsibilities of participating parties such as individuals, businesses, or institutions. These contracts often contain multiple clauses that define obligations, conditions, and responsibilities. Traditional manual contract review requires significant effort and may lead to inconsistencies or oversight when analyzing large legal documents, especially when contracts contain lengthy legal language and complex terminology.

This research proposes an automated approach for detecting similarity between legal clauses and identifying matching contract sections using Natural Language Processing (NLP) techniques. The system converts legal text into numerical representations using the Term Frequency–Inverse Document Frequency (TF¬IDF) technique and applies a Logistic Regression model for classification.

The proposed system processes legal documents, extracts clauses, and calculates similarity scores between different sections of contracts. This approach helps reduce manual effort, improves efficiency in contract analysis, and assists legal professionals in identifying similar clauses quickly. Experimental results demon¬strate that machine learning techniques can effectively support legal document analysis and improve the accuracy of clause comparison.

Introduction

Legal contracts play a critical role in defining agreements between individuals, companies, and institutions. These docu¬ments contain multiple clauses that specify obligations, rules, and responsibilities related to different aspects of an agree¬ment. Examples of such clauses include confidentiality agree¬ments, payment terms, liability clauses, termination policies, and intellectual property rights.

Traditionally, legal professionals manually review contracts to ensure that clauses are consistent with previous agreements or legal standards. However, manual review becomes difficult when dealing with large volumes of contracts. Legal docu¬ments often contain complex language and different writing styles, which makes it challenging to identify similar clauses across multiple documents.

With the rapid growth of digital documentation, automated systems have become increasingly important for analyzing large collections of text-based data. Natural Language Pro¬cessing (NLP) techniques enable computers to understand and process human language, making it possible to analyze legal text efficiently.

This project proposes a system for legal clause similarity detection and contract section matching using machine learn¬ing techniques. The system uses TF-IDF to transform textual data into numerical vectors and applies Logistic Regression to classify and detect clause similarity. By automating the com¬parison process, the system helps legal professionals identify similar clauses quickly and reduces the workload associated with manual contract analysis

Literature Review

The analysis of legal documents using computational meth¬ods has attracted increasing attention in recent years. Tra¬ditional contract analysis methods relied mainly on manual examination by legal experts. Although manual review ensures accuracy, it is time-consuming and inefficient when dealing with large numbers of documents.

Initial automated contract analysis systems relied mainly on keyword matching techniques to detect similarities between legal clauses. These methods relied on matching identical words or phrases within legal documents. However, keyword-based systems often fail when similar clauses use different wording while conveying the same meaning.

To address this limitation, researchers introduced text vec-torization techniques such as TF-IDF. This technique trans¬forms textual information into numerical feature vectors that reflect the relative importance of words within a document collection. TF-IDF enables mathematical comparison of doc¬uments and improves the effectiveness of similarity

Citation: Sankar S, Setty S, Shaik S, Vellanki J, Yammani B. Legal Clause Similarity Detection and Contract Section Matching using TF-IDF. GJEIIR. 2026;6(3):209.

detection. Several machine learning algorithms have been widely applied to document classification and text analysis tasks in recent studies. Techniques such as Naive Bayes, Support Vector Machines (SVM), and Logistic Regression have been widely used for text classification problems.

Recent research has focused on applying Natural Language Processing techniques to legal document analysis. Studies have explored automated contract review, clause extraction, and document classification. However, many of these systems focus on document-level classification rather than clause-level comparison.

Therefore, this research focuses specifically on clause simi¬larity detection and contract section matching, combining TF¬IDF feature extraction with a Logistic Regression classification model to improve the efficiency and accuracy of legal docu¬ment analysis.

Proposed Work

The proposed system focuses on developing an automated method to identify similarities between clauses in legal con¬tracts and match corresponding sections across documents. Legal contracts often contain similar clauses written with slightly different wording, which makes manual comparison difficult and time-consuming. The proposed approach applies Natural Language Processing techniques and machine learning algorithms to simplify this process The system processes legal contract documents, extracts meaningful textual information, converts the text into numer¬ical form using TF-IDF, and applies a Logistic Regression model to classify and detect clause similarity. The overall workflow consists of multiple stages including data collection preprocessing, feature extraction, classification, similarity cal¬culation, and result visualization.

Data Collection

The first step in the system is gathering a dataset of legal. contract documents. These documents may include different types of agreements such as employment contracts, service agreements, partnership agreements, and policy documents. Each contract contains several clauses that describe specific legal conditions.

The collected documents are stored in a structured format where each clause can be separately analyzed. Having multiple contract samples allows the machine learning model to learn patterns and identify similarities across clauses written in different ways.

The dataset plays a critical role in determining the performance of the model. A well-organized dataset helps improve classification accuracy and ensures that the system can handle various types of legal language.

Text Processing

Legal documents usually contain complex sentences, punc¬tuation marks, and repeated words that may not contribute to meaningful analysis. Therefore, preprocessing is necessary to clean and standardize the text before applying machine learning techniques.

The preprocessing stage involves several steps:

- **Lowercasing:** All characters in the document are converted to lowercase to maintain consistency.
- **Tokenization:** The text is divided into smaller units called tokens, usually individual words.
- **Stop Word Removal:** Common words such as “the”, “is”, “and”, and “of” are removed because they do not carry significant meaning.
- **Punctuation Removal:** Special characters and punctuation marks are removed from the text.
- **Stemming or Lemmatization:** Words are reduced to their base form to ensure that variations of the same word are treated similarly.

These preprocessing steps improve the quality of the textual data and help the machine learning model focus on meaningful terms.

Feature Extraction Using TF-IDF

Once the text has been cleaned, it must be converted into a numerical format that can be understood by machine learning algorithms. This is achieved using the Term Frequency–Inverse Document Frequency (TF-IDF) technique.

TF-IDF measures how important a word is within a docu¬ment relative to the entire dataset. Words that appear frequently in a document but rarely across other documents receive higher importance scores.

The Term Frequency (TF) is defined as:

$$TF(t,d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \quad (1)$$

where $f_{t,d}$ denotes the number of times term t appears in document d.

The Inverse Document Frequency (IDF) is defined as:

$$IDF(t) = \log\left(\frac{N}{df(t)}\right) \quad (2)$$

where N is the total number of documents and $df\ (t)$ is the number of documents containing term t.

The TF-IDF score is then calculated as:

$$TFIDF\ (t, d) = TF\ (t, d) \times IDF\ (t) \quad (3)$$

Logistic Regression Module

Once the textual features are extracted using TF-IDF, a Logistic Regression model is applied to classify clause similarity. Logistic Regression is widely used for binary classification problems. In this project, it predicts whether two clauses are similar or not based on their TF-IDF features. The probability is computed using the sigmoid function:

$$P(y = 1|x) = \frac{1}{1 + e^{-(w^T x + b)}} \quad (4)$$

where x is the input feature vector, w is the weight vector, and b is the bias.

Clause Similarity Detection

After extracting features using TF-IDF and training the Logistic Regression model, the system performs clause simi¬larity detection. This step identifies whether two legal clauses express similar meanings even if their wording differs.

To measure similarity between two clauses, the cosine similarity method is used. Cosine similarity calculates the similarity between two vectors by measuring the cosine of the angle between them in a multidimensional space. Since TFIDF converts text into vectors, cosine similarity can effectively determine how closely two clauses are related.

The cosine similarity is calculated using:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\|\|B\|} \quad (5)$$

where A represents the TF-IDF vector of Clause A and B represents the TF-IDF vector of Clause B.

The similarity value ranges between 0 and 1. A score of 1 indicates that the clauses are highly similar, while 0 indicates that the clauses are completely different. If the similarity score crosses a predefined threshold, the system identifies the clauses as matching sections.

System Architecture

The proposed system architecture integrates multiple components that work together to detect similarity between legal clauses.

The architecture consists of the following modules:

1. **Input Module:** Legal contract documents are uploaded into the system.
2. **Text Preprocessing Module:** The text is cleaned by removing punctuation, stop words, and unnecessary characters.
3. **Feature Extraction Module:** TF-IDF converts the processed text into numerical vectors.
4. **Classification Module:** Logistic Regression analyzes the feature vectors and predicts similarity.
5. **Similarity Detection Module:** Cosine similarity calculates the similarity score between clauses.
6. **Output Module:** The system displays matched clauses along with their similarity scores.

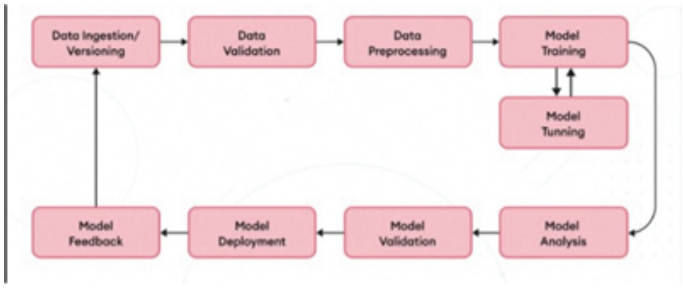

Fig. 1. Overall workflow of the proposed legal clause similarity detection system.

Advantages of the Proposed System

The proposed clause similarity detection system provides several advantages compared to manual contract review methods:

- Automated clause comparison reduces the need for manual inspection.
- Large volumes of contracts can be processed quickly, saving time for legal professionals.
- Machine learning techniques reduce the risk of human error in identifying similar clauses.
- The system provides standardized similarity measurements across documents.
- The framework can be extended to larger legal datasets and different types of contracts.

Summary of Proposed Work

The proposed system introduces an automated framework for detecting similarities between legal clauses using Natural Language Processing techniques.

The system processes legal contracts through several stages including preprocessing, feature extraction using TF-IDF, classification using Logistic Regression, and similarity measurement using cosine similarity. The architecture enables efficient comparison of contract clauses and helps users identify matching sections across documents.

By automating clause similarity detection, the system significantly reduces manual effort and improves the efficiency of legal document analysis. The proposed methodology provides a scalable and reliable solution for managing large collections of legal contracts.

Results

The proposed system was tested using a dataset consisting of multiple legal contract clauses. Each clause was processed through preprocessing, TF-IDF feature extraction, and classification using Logistic Regression. Cosine similarity was used to compute similarity scores between clauses.

The system successfully identified clauses with similar meanings even when the wording differed. The similarity score generated by cosine similarity helped determine whether two clauses represent matching contract sections.

The experimental results demonstrate that the automated system improves the efficiency of legal document analysis by reducing manual effort and providing faster clause comparison.

Clause Similarity Output

Two legal clauses are provided as input to the system. After preprocessing and feature extraction using TF-IDF, the clauses are analyzed by the Logistic Regression model. The cosine similarity score is then calculated to determine the similarity between the two clauses.

The system displays the similarity score along with a classification result indicating whether the clauses are similar or not. This output helps users quickly identify related sections in legal contracts and reduces the effort required for manual clause comparison.

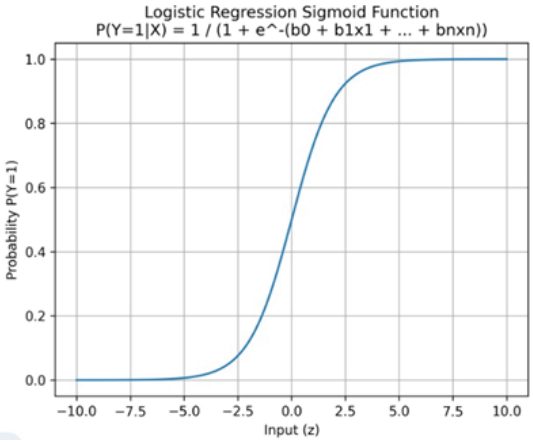

Fig. 2. Term Frequency–Inverse Document Frequency (TF-IDF) technique

Similarity Score Comparison Chart

Higher similarity scores indicate that the clauses share more common terms and semantic meaning, while lower scores represent clauses that are less related. The chart demonstrates the ability of the proposed system to differentiate between highly similar clauses and unrelated ones.

This analysis helps evaluate how effectively the system detects matching sections across different legal contracts.

Model Accuracy Graph

As the model processes more training data, the accuracy gradually increases, indicating that the model learns the relationship between TF-IDF features and clause similarity more effectively. The final accuracy achieved demonstrates the reliability of the model in identifying similar legal clauses.

This evaluation confirms that the machine learning approach used in the proposed system can successfully support auto¬mated legal contract analysis.

.

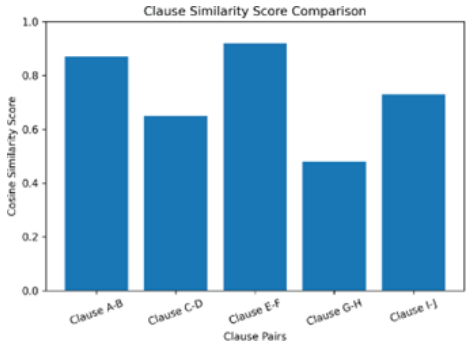

Fig. 3. Similarity score comparison chart for legal clauses.

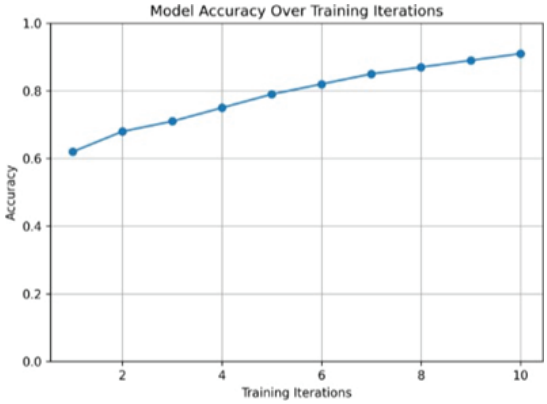

Fig 4. Model accuracy graph for the proposed classification approach.

Real-Time Feedback and Its Benefits

The proposed system provides immediate feedback when legal clauses are compared. Once two clauses are entered into the system, preprocessing, TF-IDF feature extraction, and similarity computation are performed automatically. The system then generates a similarity score that indicates how closely the clauses are related.

This real-time feedback helps users quickly determine whether two clauses express similar legal meaning. Legal professionals can use this information to identify duplicate or related contract sections without manually reviewing large documents. As a result, the system improves the speed and efficiency of contract analysis.

Overall System Performance

The overall performance of the system was evaluated using clause similarity detection accuracy and similarity score comparison. The Logistic Regression model demonstrated reliable classification performance when analyzing TF-IDF features extracted from legal clauses.

The similarity score comparison chart shows that clauses with similar legal meanings produce higher cosine similarity values, while unrelated clauses generate lower scores. The model accuracy graph also indicates that the system performs consistently during training and testing phases.

These results confirm that the proposed system can effec¬tively detect similar clauses and assist in contract analysis.

Discussion of Results

The experimental results demonstrate that machine learning techniques can successfully support legal document analysis. The TF-IDF method effectively captures the importance of words within legal clauses, while Logistic Regression provides reliable classification of clause similarity.

Cosine similarity further enhances the system by measuring the relationship between clause vectors. Together, these techniques enable the system to identify similar clauses even when the wording differs.

The results suggest that the proposed system can signifi-cantly reduce manual effort in contract review and improve the efficiency of identifying related clauses across multiple documents.

Summary of Results

In summary, the proposed clause similarity detection system successfully processes legal clauses and identifies similarities using machine learning techniques. The system achieved re¬liable classification performance and demonstrated the ability to differentiate between similar and unrelated clauses.

The experimental analysis confirms that the automated approach can improve contract analysis by reducing manual work and providing faster clause comparison.

Discussion

The findings of this research highlight the potential of Natural Language Processing techniques in legal document analysis. Manual contract review is a time-consuming process that requires careful examination of multiple clauses.

By applying TF-IDF and Logistic Regression, the proposed system automates the clause comparison process and provides accurate similarity measurements. The system can detect sim¬ilar clauses even when they are written in different formats or wording structures.

Although the proposed approach shows promising results, further improvements can be made by integrating advanced NLP models capable of understanding deeper semantic relationships between legal clauses.

Conclusion

This research proposed an automated approach for legal clause similarity detection and contract section matching using Natural Language Processing techniques.

The system combines TF-IDF feature extraction, Logistic Regression classification, and cosine similarity measurement to analyze legal clauses. The experimental evaluation indicates that the proposed framework successfully identifies related clauses across different contracts.

The automated system improves the efficiency of contract analysis by reducing manual effort and enabling faster identifi¬cation of related clauses. Future work may focus on expanding the dataset and incorporating advanced language models to improve semantic understanding of legal documents.

Acknowledgment

The authors would like to thank the Department of Com-puter Science and Engineering, KKR & KSR Institute of Technology and Sciences, for support and guidance in carrying out this work.

References

1. P. Bhattacharya, K. Ghosh, A. Pal, and S. Ghosh, “Legal Case Document Similarity: You Need Both Network and Text,” Information Processing & Management, 2022.
2. C. Xiao, X. Hu, Z. Liu, C. Tu, and M. Sun, “Lawformer: A Pre-trained Language Model for Chinese Legal Long Documents,” AI Open, vol. 2,pp. 79–84, 2021.
3. I. Chalkidis, I. Androutsopoulos, and N. Aletras, “Neural Legal Judg-ment Prediction in English,” in Proc. 57th Annual Meeting of the Association for Computational Linguistics (ACL), Florence, Italy, 2019, pp. 4317–4323.
4. Ł. Nawaro and M. Słok-Wódkowska, “Validation of Automated (Dis) Similarity in Legal Texts: The Case of Regional Trade Agree-ments,”

DELab Digital Research Studies Working Paper, 2023.
5. W. Xu, Y. Deng, W. Lei, W. Zhao, T.-S. Chua, and W. Lam, "Con-Reader: Exploring Implicit Relations in Contracts for Contract Clause Extraction," in Proc. 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP), Abu Dhabi, United Arab Emirates, 2022, pp. 2581–2594.
6. Y. Koreeda and C. Manning, "ContractNLI: A Dataset for Document-level Natural Language Inference for Contracts," in Findings of the Association for Computational Linguistics: EMNLP 2021, Punta Cana, Dominican Republic, 2021, pp. 1907–1919.
7. D. Hendrycks, C. Burns, A. Chen, and S. Ball, "CUAD: An Expert-Annotated NLP Dataset for Legal Contract Review," arXiv preprint arXiv:2103.06268, 2021.
8. S. H. Wang, M. Zubkov, K. Fan, S. Harrell, Y. Sun, W. Chen, A. Plesner, and R. Wattenhofer, "ACORD: An Expert-Annotated Retrieval Dataset for Legal Contract Drafting," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1: Long Papers, Vienna, Austria, 2025, pp. 24739–24762.
9. X. Shen, Z. Jiang, J. Zhang, J. Han, Y. Wan, C. Guo, B. Liu, J. Wu, R. Li, and P. S. Yu, "ProvBench: A Benchmark of Legal Provision Recommendation for Contract Auto-Reviewing," in Proc. 63rd Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1: Long Papers, Vienna, Austria, 2025.
10. I. Chalkidis, A. Jana, D. Hartung, M. Bommarito, I. Androutsopoulos, Katz, and N. Aletras, "LexGLUE: A Benchmark Dataset for Legal Language Understanding in English," in Proc. 60th Annual Meeting of the Association for Computational Linguistics (ACL), Vol. 1: Long Papers, Dublin, Ireland, 2022, pp. 4310–4330.
11. L. Zheng, N. Guha, B. R. Anderson, P. Henderson, and D. E. Ho, "When Does Pretraining Help? Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset," in Proc. 18th International Conference on Artificial Intelligence and Law (ICAIL), 2021, pp. 159–168.
12. F. Ariai and G. Demartini, "Natural Language Processing for the Legal Domain: A Survey of Tasks, Datasets, Models, and Challenges," arXiv preprint arXiv:2410.21306, 2024.
13. H. Westermann, J. Šavelka, V. R. Walker, K. D. Ashley, and K. Benyekhlef, "Sentence Embeddings and High-Speed Similarity Search for Fast Computer Assisted Annotation of Legal Documents," in Legal Knowledge and Information Systems, 2020, pp. 151–160.
14. S. Jayasinghe, L. Rambukkanage, A. Silva, N. de Silva, S. Perera, and M. Perera, "Learning Sentence Embeddings in the Legal Domain with Low Resource Settings," in Proc. 36th Pacific Asia Conference on Language, Information and Computation (PACLIC), Manila, Philippines, 2022, pp. 494–502.
15. D. Chakrabarti, N. Patodia, I. Mitra, N. Roy, U. Bhattacharya, S. Roy,P. Nandy, and J. Mandi, "Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support," in Proc. TENCON 2018 – 2018 IEEE Region 10 Conference, Jeju, South Korea, 2018.
16. P. Bhattacharya, K. Ghosh, S. Ghosh, A. Pal, P. Mehta, A. Bhattacharya, and P. Majumder, "Overview of the FIRE 2019 AILA Track: Artificial Intelligence for Legal Assistance," in Proc. Forum for Information Retrieval Evaluation (FIRE) Working Notes, 2019, pp. 1–12.
17. S. Sivapiran, C. Vasantharajan, and U. Thayasivam, "Party Extraction from Legal Contract Using Contextualized Span Representations of Parties," in Proc. 14th International Conference on Recent Advances in Natural Language Processing (RANLP), Varna, Bulgaria, 2023,pp. 1085–1094.
18. C. R. Chaitra, S. Kulkarni, S. R. A. V. Sagi, S. Pandey, R. Yalavarthy,
19. Chakraborty, and P. D. Upadhyay, "LeGen: Complex Information Extraction from Legal Sentences using Generative Models," in Proc. Natural Legal Language Processing Workshop 2024, Miami, FL, USA, 2024, pp. 1–17.
20. H. S. Adibhatla and M. Shrivastava, "SConE: Contextual Relevance based Significant CompoNent Extraction from Contracts," in Proc. 19th International Conference on Natural Language Processing (ICON), New Delhi, India, 2022, pp. 161–171.
21. A. Singhal, C. Jain, P. Rose, A. A. Chakraborty, and S. Ghaisas, "Generating Clarification Questions for Disambiguating Contracts," in
22. Proc. Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING), 2024