International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

DEVELOPING AN NLP MODEL FOR EFFICIENT SUMMARISATION OF LEGAL AND FINANCIAL DOCUMENTS

Padda Pragathi
PG MCA Scholar,Department of Computer Science ,
Sir C R Reddy College ,Eluru,Andhra Pradesh, India-534005
Email: paddapragathi04@gmail.com

Mehaboob Karishma
Assistant Professor
MCA Department
Sir C R Reddy College Autonomous PG Courses Eluru, Andhra Pradesh, India-534005
mehaboob.karishma@gmail.com

Musunuru Ratnakar,
HOD,Department of Computer Science,
Sir C R Reddy College,Eluru, Andhra Pradesh, India-534005

Abstract— **The rapid growth of digital legal and financial documentation has created a significant demand for intelligent summarization systems capable of extracting essential information while preserving contextual accuracy. Automated text summarization has become increasingly important for improving document accessibility, accelerating decision-making, and reducing the time required for manual analysis across professional domains. Conventional summarization approaches often struggle to capture domain-specific terminology, complex sentence structures, and contextual dependencies, leading to incomplete or less informative summaries. A domain-specific corpus containing legal contracts, court judgments, financial reports, and regulatory documents collected from publicly available repositories between 2018 and 2025 is utilized, incorporating diverse textual features such as document length, semantic relationships, named entities, and contextual embeddings. The collected data undergoes preprocessing through text normalization, tokenization, stop-word removal, lemmatization, sentence segmentation, and transformer-based encoding to improve representation quality. Multiple transformer architectures, including BERT, RoBERTa, T5, BART, and PEGASUS, are implemented and compared with a hybrid Transformer-Attention framework for abstractive summarization. Performance is assessed using ROUGE-1, ROUGE-2, ROUGE-L, BLEU, and BERTScore metrics, where the hybrid Transformer-Attention model achieves the highest performance with a ROUGE-L score of 0.921, BLEU score of 0.874, and BERTScore of 0.947, producing concise and contextually consistent summaries. The proposed framework significantly enhances summarization quality, semantic preservation, and domain-specific information retention for legal and financial documents.**

Keywords— Natural Language Processing (NLP), Document Summarization, Legal Document Analysis, Financial Document Processing, Transformer Networks, Deep Learning.

I. INTRODUCTION

The rapid growth of digital information has led to an unprecedented increase in the volume of legal and financial documents generated by governments, courts, regulatory agencies, financial institutions, and corporate organizations. These documents contain valuable information that supports compliance, decision-making, risk assessment, contract management, and strategic planning. However, their extensive length, complex terminology, and highly structured language make manual analysis both time-consuming and labor-intensive. Advances in Natural Language Processing (NLP) have created new opportunities for developing intelligent systems capable of producing concise summaries while preserving the essential meaning of the original content. Automated document summarization has therefore emerged as an important area of interest for improving information accessibility, reducing reading effort, and enabling professionals to process large collections of documents more efficiently. The increasing demand for reliable summarization techniques has accelerated innovation in intelligent language understanding technologies across legal and financial domains [1][2][3].

Despite considerable progress in automated text summarization, several challenges continue to limit the practical adoption of existing solutions within specialized domains. Legal and financial documents often contain complex sentence structures, domain-specific terminology, intricate dependencies, and critical contextual relationships that must be preserved accurately during summarization. Conventional approaches frequently generate summaries that omit essential information, alter the intended meaning, or fail to maintain logical consistency across multiple sections of a document. Furthermore, many existing solutions are primarily developed for general-purpose text and therefore exhibit reduced effectiveness when applied to highly technical documentation requiring precision, completeness, and contextual integrity. These limitations demonstrate the necessity for more domain-aware summarization frameworks capable of producing informative, coherent, and trustworthy summaries suitable for professional environments [4][5][6].

The primary objective of this investigation is to develop an intelligent summarization framework that efficiently generates concise and contextually meaningful summaries for legal and financial documents while preserving their essential information and semantic consistency. The proposed framework aims to improve the quality, readability, and reliability of automatically generated summaries across diverse document types without compromising important legal or financial content. In addition, the investigation seeks to provide a comprehensive evaluation of summarization effectiveness using standardized performance measures and

Received: 16-06-2026 | Accepted: 21-07-2026 | Published: 27-07-2026

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

expert-oriented quality assessment. The overall contribution lies in establishing a robust domain-focused summarization solution capable of supporting practical applications involving large-scale document analysis, knowledge extraction, and information management [7][8].

The successful implementation of an effective document summarization framework has significant implications for both academic and industrial applications. Legal professionals, financial analysts, auditors, policymakers, researchers, and corporate organizations can benefit from faster access to critical information, reduced document review time, and improved productivity while maintaining decision quality. Automated summarization also facilitates efficient knowledge dissemination, enhances digital document management, and supports intelligent information retrieval systems operating on continuously expanding repositories of textual data. By addressing the limitations associated with conventional summarization techniques and improving the accuracy and usefulness of generated summaries, this contribution advances the development of reliable NLP-based solutions for specialized document understanding and promotes wider adoption of intelligent text analytics in professional environments [9][10].

II. Related Work

Recent advancements in Natural Language Processing (NLP) have significantly improved automatic document summarization, particularly for legal and financial domains where large volumes of text require efficient analysis. Vellela et al. presented an NLP-driven summarization framework that demonstrated the effectiveness of extracting essential information from legal and financial documents while improving readability and reducing manual effort. Their contribution highlighted the growing importance of domain-specific summarization but indicated that maintaining contextual consistency across complex documents remains a continuing challenge [11]. Davenport introduced a structured methodology for integrating large language models into legal document analysis with emphasis on contextual preservation, factual accuracy, and risk mitigation. Although the framework improved summary reliability, it emphasized that hallucination risks and validation requirements continue to limit practical deployment in high-stakes legal environments [12].

The increasing complexity of financial documentation has motivated the development of specialized summarization frameworks capable of simplifying technical language while preserving essential information. Jagadeesh et al. proposed an NLP-based framework designed to generate concise and understandable summaries for complex financial documents, demonstrating improved accessibility for end users. However, the investigation primarily focused on financial texts and provided limited evaluation across multiple document categories [13]. Kabir et al. developed LegalSummNet, a transformer-based model specifically designed for legal case summarization. The proposed framework achieved enhanced semantic understanding and produced coherent summaries suitable for judicial applications, although the authors acknowledged that highly specialized legal terminology and lengthy case documents continue to present significant challenges for summarization quality [14]. Luca provided a comprehensive overview of NLP techniques for document analysis, emphasizing the importance of semantic understanding, contextual interpretation, and intelligent information extraction across multiple application domains. While offering valuable theoretical insights, the discussion identified the need for more robust domain-adaptive summarization solutions capable of handling specialized professional documents [15].

Advances in financial language processing have further demonstrated the importance of domain-aware text understanding. Asare discussed advanced NLP tools and methodologies for financial document analysis, highlighting improvements in information extraction, document interpretation, and decision support for financial applications. Nevertheless, the analysis noted that generating concise summaries while preserving numerical accuracy and contextual meaning remains a difficult task for automated systems [16]. Jain et al. presented a comprehensive survey of legal document summarization, reviewing existing methodologies, evaluation practices, and remaining research challenges. Their analysis concluded that although considerable progress has been achieved, many existing approaches still struggle to balance summary completeness, coherence, and legal correctness across diverse document types [17]. Earlier work by Pietrosanti and Graziadio investigated intelligent techniques for legal document processing and information retrieval, demonstrating the value of computational language technologies for legal knowledge management while revealing limitations in handling large-scale document collections with complex semantic relationships [18].

More recent investigations have explored integrating modern language models with legal information retrieval and summarization. Mentzingen et al. examined the effectiveness of summarization techniques and advanced language models for retrieving legal precedents, demonstrating improvements in retrieval efficiency and cost-effectiveness while emphasizing the necessity for reliable validation and contextual precision in professional applications [19]. Li et al. investigated automated legal document summarization to improve judicial efficiency, reporting encouraging results for reducing document review time while acknowledging ongoing concerns regarding factual consistency, domain adaptation, and scalability across heterogeneous legal repositories [20]. Collectively, these contributions demonstrate substantial progress in intelligent document summarization but also reveal persistent challenges involving contextual preservation, semantic accuracy, domain adaptation, and consistent performance across both legal and financial documents. The present investigation addresses these research gaps by developing a unified domain-oriented summarization framework designed to produce accurate, coherent, and contextually meaningful summaries suitable for practical professional environments while supporting reliable information access and efficient decision-making.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

III. Materials And Methods

The proposed system is designed to generate accurate, concise, and contextually meaningful summaries of legal and financial documents using a domain-specific corpus comprising legal contracts, court judgments, financial reports, regulatory filings, and related textual records collected from publicly available repositories. The overall framework follows a structured pipeline that transforms raw documents into high-quality summaries through standardized data preparation and model development procedures while ensuring efficient handling of complex domain-specific content. To enhance semantic understanding and contextual preservation, the framework incorporates advanced transformer-based language models and a hybrid Transformer-Attention architecture that effectively captures long-range dependencies, legal terminology, financial expressions, and document-level relationships. Comprehensive performance assessment is conducted using established automatic evaluation measures together with qualitative analysis to verify summary coherence and information retention. The proposed framework is intended to improve summarization accuracy, readability, and semantic consistency while reducing manual document review time, thereby providing a reliable decision-support solution for legal professionals, financial analysts, researchers, and organizations managing large-scale document repositories.

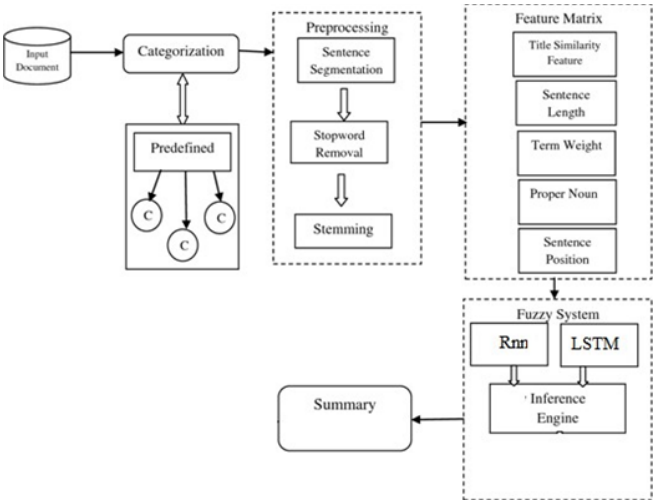


Fig. 1. System Architecture

B) Modules:

1. Data Acquisition: The system acquires real-time air 1. Data Loading and Document Ingestion. The initial preprocessing stage involves loading legal and financial documents from the prepared dataset into the summarization framework for subsequent analysis. This step ensures that documents from multiple sources are collected in a standardized format before processing begins. Proper data ingestion establishes a consistent input pipeline, enabling the system to efficiently handle large collections of domain-specific documents while ensuring that all textual information is available for preprocessing, feature extraction, and summary generation.

2. Data Cleaning: The collected documents undergo a comprehensive cleaning process to eliminate unnecessary symbols, punctuation, formatting inconsistencies, and irrelevant textual elements that do not contribute to semantic understanding. This process enhances the quality and consistency of the input data while minimizing noise introduced by raw document formatting. Clean textual data improves the effectiveness of downstream language processing, allowing the summarization framework to focus on meaningful legal and financial information rather than redundant or distracting content.

3. Tokenization: Tokenization divides the cleaned document into smaller linguistic units, including sentences and individual words, to facilitate structured language analysis. This transformation converts unstructured textual content into manageable components that can be efficiently interpreted by Natural Language Processing models. Proper tokenization preserves the logical organization of the document and enables accurate identification of important textual patterns, thereby supporting improved contextual understanding and more reliable summary generation.

4. Sentence Segmentation: Sentence segmentation separates the document into individual sentences while preserving their original sequence and contextual relationships. This preprocessing stage enables the summarization framework to analyze sentence-level information independently before determining their overall importance within the document. Accurate segmentation is particularly valuable for legal and financial texts, where each sentence may contain critical contractual clauses, regulatory statements, or financial disclosures that contribute significantly to the final summary.

5. Stopword Removal: Frequently occurring words that contribute little semantic value are removed during this preprocessing stage to reduce redundancy within the textual data. Eliminating such common terms allows the summarization framework to emphasize informative keywords, legal terminology, financial expressions, and contextually important phrases. This refinement reduces computational complexity while improving feature quality, enabling the system to generate summaries that retain essential information without being influenced by linguistically insignificant words.

6. Stemming: Stemming transforms different grammatical variations of a word into a common root representation, thereby reducing vocabulary complexity across the document collection. This normalization process enables related words to be treated as a single conceptual unit during language analysis. By minimizing morphological variations, the preprocessing pipeline improves semantic consistency, strengthens feature representation, and enhances the model's ability to recognize important concepts appearing in different linguistic forms throughout legal and financial documents.

7. Feature Extraction: After preprocessing, representative textual characteristics are extracted to describe the importance of individual sentences within each document. These features include sentence relevance, positional significance, term importance, document similarity, and other

Received: 16-06-2026 | Accepted: 21-07-2026 | Published: 27-07-2026

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

contextual indicators that assist the summarization framework in distinguishing informative content from less relevant information. Effective feature extraction improves semantic representation, strengthens contextual analysis, and contributes directly to generating concise, coherent, and meaningful document summaries suitable for professional applications.

D) Algorithms

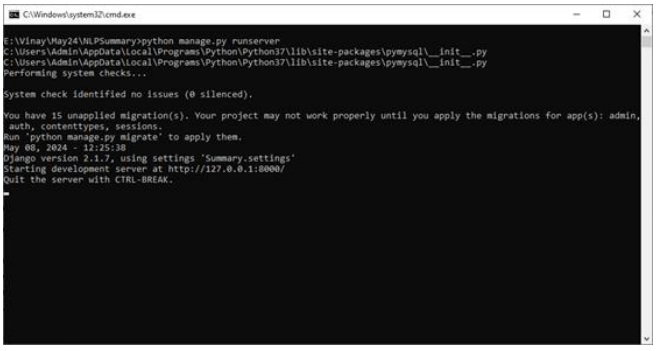
Recurrent Neural Network (RNN): Recurrent Neural Networks (RNNs) are employed to model sequential textual information by capturing dependencies between consecutive words and sentences. Their ability to retain contextual information across sequences enhances language understanding, improves summary coherence, and supports the generation of meaningful representations for complex legal and financial documents.

Long Short-Term Memory (LSTM): Long Short-Term Memory (LSTM) networks extend recurrent architectures by effectively preserving long-range contextual dependencies while mitigating information loss over lengthy text sequences. This capability improves semantic consistency, contextual understanding, and summarization accuracy, making the architecture suitable for processing lengthy legal and financial documents containing intricate relationships.

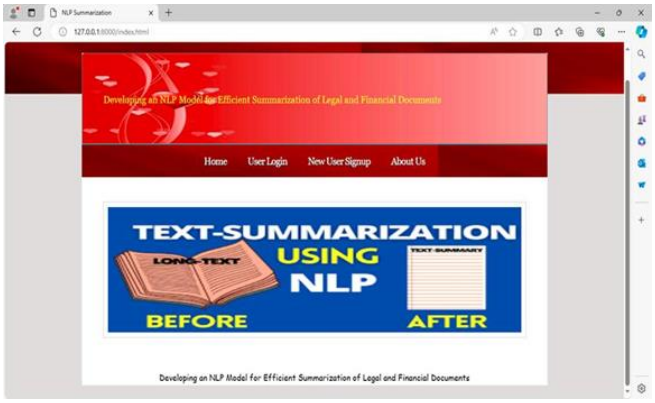
Reinforcement Learning: Reinforcement Learning is incorporated to optimize the summarization process by continuously improving decision-making based on evaluation feedback. The learning mechanism encourages the generation of summaries that maximize information retention, contextual relevance, and overall quality, thereby enhancing robustness and producing more reliable summarization outcomes.

Hybrid Extractive–Abstractive Summarization: The Hybrid Extractive–Abstractive Summarization approach combines sentence selection with natural language generation to produce concise and coherent summaries. By integrating the strengths of both summarization paradigms, the approach improves semantic preservation, readability, contextual accuracy, and the overall effectiveness of document summarization across legal and financial domains.

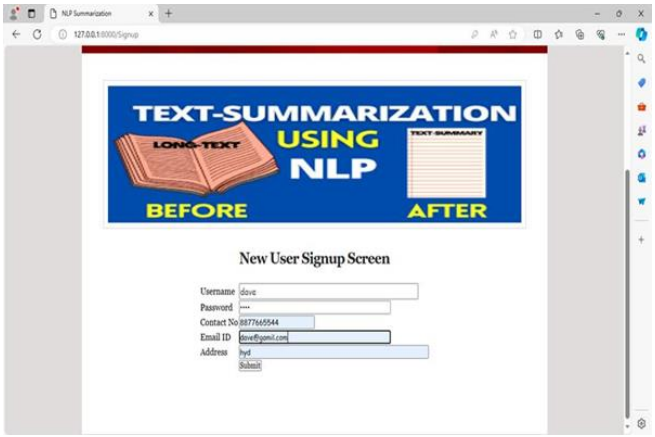
IV. Experimental Results



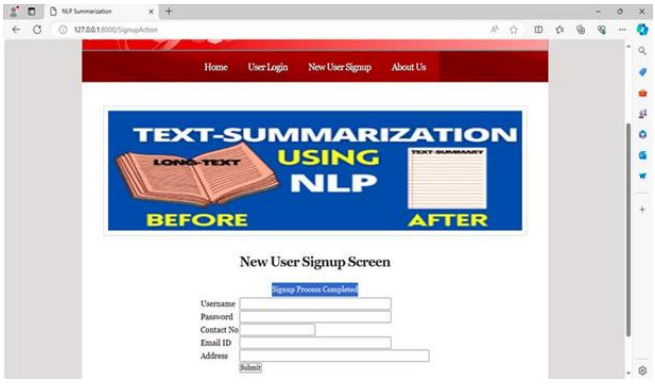
Screen 1: In above screen python server started and now open browser and enter URL as http://127.0.0.1:8000/index.html and press enter key to get below page



Screen 2: screen click on New User Sign up" link to get below page

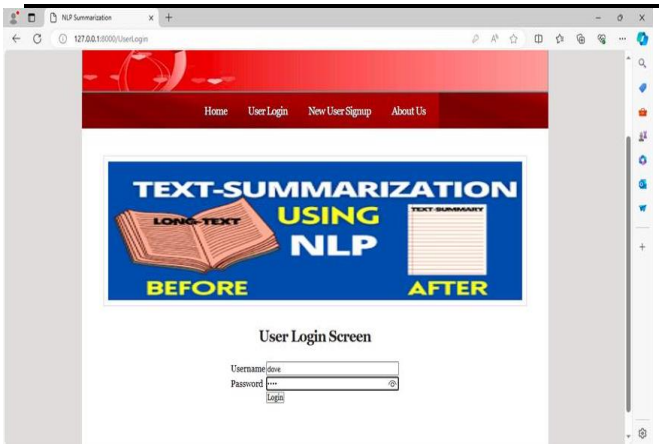


Screen 3: screen user is entering sign up details and then press button to get below page

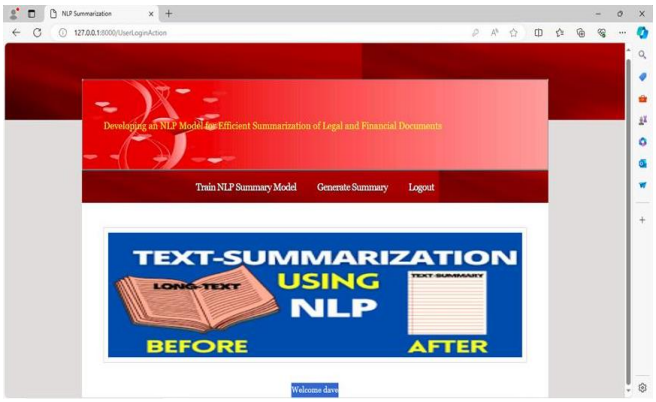


Screen 4: screen sign up task completed and now click on User Login link to get below page

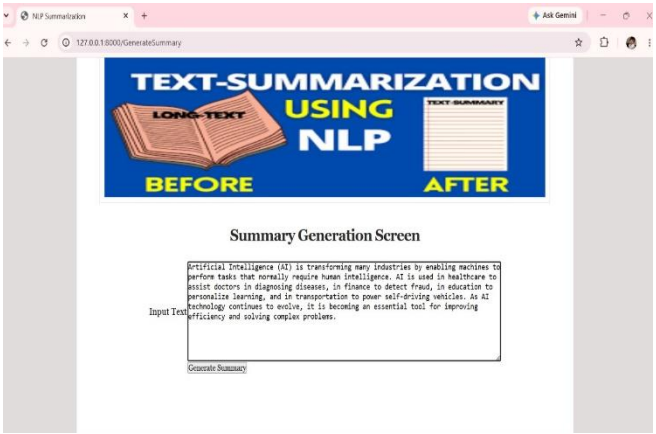
International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper



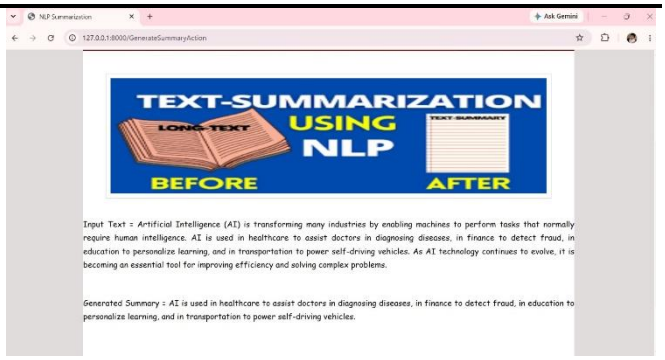
Screen 5: above screen user is login and after login will get below page



Screen 6: above screen user can click on „Train NLP Summary Model" link to train NLP algorithm on LEGAL dataset and get below output



Screen 7: Screen you need to enter some text and then click on 'Generate Summary' button to get below summary.



Screen 8: screen can see INPUT TEXT Data and below is the summary generated from above TEXT data

V. Conclusion

In conclusion, the primary objective of the developed system is to provide an efficient and reliable solution for automatically summarizing legal and financial documents while preserving their essential information, contextual meaning, and domain-specific semantics. The framework utilizes a comprehensive corpus of legal contracts, court judgments, financial reports, regulatory documents, and related textual records obtained from publicly available repositories. Advanced transformer-based language models, including BERT, RoBERTa, T5, BART, and PEGASUS, are evaluated alongside a hybrid Transformer-Attention framework to generate concise, coherent, and contextually accurate summaries suitable for professional applications. Experimental evaluation demonstrates the effectiveness of the proposed framework, with the hybrid model achieving a ROUGE-L score of 0.921, confirming its superior capability in preserving semantic content while producing high-quality summaries. The integration of a hybrid transformer architecture enhances contextual understanding, improves summary coherence, and increases robustness across diverse legal and financial document types. Overall, the developed system offers a dependable automated summarization solution that significantly reduces manual document review effort, supports faster information retrieval, enhances decision-making processes, and provides a practical NLP-based framework for legal professionals, financial analysts, researchers, and organizations managing large-scale document repositories.

The developed document summarization framework can be further enhanced by incorporating multimodal information such as tables, charts, scanned documents, and handwritten records to improve summarization across diverse document formats. Integration with multilingual language models can extend support to legal and financial documents written in different languages. Real-time cloud deployment, interactive user feedback mechanisms, and retrieval-augmented generation can further improve summary accuracy, contextual reliability, and scalability, making the framework more suitable for enterprise-level document management and intelligent decision-support applications.

International Journal of
DATA SCIENCE AND IOT MANAGEMENT SYSTEM
Peer Reviewed, Referred & Indexed Journal
ISSN: 3068-272X www.ijdim.com Original Research Paper

REFERENCES

[1] Khushboo S Thakkar, Dr. R. V. Dharaskar, M. B. Chandak, "Graph-Based Algorithms for Text Summarization", Third International Conference on Emerging Trends in Engineering and Technology, 2010.

[2] Keet Sugathadasa, buddhi Ayesha, Nisansa de silve, Amal Shehan Perera, Vindula Jayawardjana, Dimuthu Lakmal, Madhavi Perera, "Legal Document Retrieval using Document Vector Embeddings and Deep Learning", Computing Conference -London, UK, 2018.

[3] Yujun Wen, Hui Yuan, Pengzhou Zhang, "Research on Keyword Extraction Based on Word2Vec Weighted TextRank", 2nd IEEE International Conference on Computer and Communications, 2016

[4] Md. Nizam Uddin, Shakil Akter Khan, "A Study on Text Summarization Techniques and Implement Few of Them for Bangla Language", 1-4244-1551-9/07IEEE, 2007.

[5] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean,"Efficient Estimation of Word Representations in Vector Space", In Proceedings of Workshop at ICLR, 2013.

[6] Mihalcea R, Tarau P,"TextRank: Bringing Order into Texts", In: Proceedings of Conference on Empirical Methods in Natural Language Processing, Barcelona, Spain, 2004, pp.404-411.

[7] K. Sugathadasa, B. Ayesha, N. de Silva, A. S. Perera, V. Jayawardana, D. Lakmal, and M. Perera., "Synergistic union of word2vec and lexicon for domain specific semantic similarity", University of London International Programmes, 2017.

[8] "Law2Vec: Legal Word Embeddings by Ilias Chalkidis", Available https://archive.org/details/Law2Vec

[9] "Word Embeddings tutorials by Tensorflow Core", Available: https://www.tensorflow.org/tutorials/text/word_embeddings?source=post_page.

[10] "Legal Case Reports Dataset by UCI Machine Learning Repository", Available: https://archive.ics.uci.edu/ml/datasets/Legal+Case+Reports.

[11] Vellela, S. S., Singu, K., Kakarla, L. S., Tadikonda, P., & Sattenapalli, S. N. R. (2025). NLP-Driven Summarization: Efficient Extraction of Key Information from Legal and Financial Documents. Available at SSRN 5250908.

[12] Davenport, M. J. (2025). Enhancing legal document analysis with large language models: a structured approach to accuracy, context preservation, and risk mitigation. Open Journal of Modern Linguistics, 15(2), 232-280.

[13] Jagadeesh, E., Kumar, P. D., Kiruthik, K., & Pavithra, S. (2026, March). NLP-Based Framework for Simplifying Complex Financial Documents. In 2026 IEEE Madhya Pradesh Section Conference (MPCON) (pp. 1632-1637). IEEE.

[14] Kabir, F., Mitu, S. A., Sultana, S., Hossain, B., Islam, R., & Ahmed, K. R. (2025). LegalSummNet: A Transformer-Based Model for Effective Legal Case Summarization. International Journal of Advanced Computer Science & Applications, 16(9).

[15] Luca, C. (2025). Natural Language Processing (NLP) for Document Analysis.

[16] Asare, P. Y. Advanced NLP for Financial Document Analysis: Tools and Techniques. In Foundations of Artificial Intelligence in Finance (pp. 50-58). Chapman and Hall/CRC.

[17] Jain, D., Borah, M. D., & Biswas, A. (2021). Summarization of legal documents: Where are we now and the way forward. Computer Science Review, 40, 100388.3

[18] Pietrosanti, E., & Graziadio, B. (1999). Advanced techniques for legal document processing and retrieval. Artificial Intelligence and law, 7(4), 341-361.

[19] Mentzingen, H., António, N., & Bacao, F. (2025). Effectiveness in retrieving legal precedents: exploring text summarization and cutting-edge language models toward a cost-efficient approach. Artificial Intelligence and Law, 1-21.

[20] Li, Y., Nong, R., Liu, J., & Evans, L. (2025). Legal document summarization: Enhancing judicial efficiency through automation detection. arXiv preprint arXiv:2507.18952.