An Automated Summarization Framework for Complex Financial and Legal Texts

Bandi Pranavi[1], Charan[2], Geerish Chennale[3], Gulla Naresh[4], Mrs. Mercy Joycce[5]

[1,2,3,4]students, Department of Computer Science and Engineering, Sree Dattha Group of Institutions, Sheriguda, Ibrahimpatnam, Telangana,India

[5]Assistant Professor, Department of Computer Science and Engineering, Sree Dattha Group of Institutions, Sheriguda, Ibrahimpatnam, Telangana,India

To Cite this Article

Bandi Pranavi, Charan, Geerish Chennale, Gulla Naresh, Mrs. Mercy Joycce, "An Automated Summarization Framework for Complex Financial and Legal Texts", Journal of Science Engineering Technology and Management Science, Vol. 03, Issue 05, May 2026, pp: 370-375, DOI: http://doi.org/10.64771/jsetms.2026.v03.i05.pp370-375

Submitted: 19-03-2026 Accepted: 25-04-2026 Published: 01-05-2026

***Abstract*— AI-powered framework designed to summarize complex legal and financial documents using modern Natural Language Processing techniques. The system combines two advanced models, COLBERT and T5, to improve both the relevance and readability of generated summaries. COLBERT is used to understand the contextual meaning of the document and rank important sentences effectively, while T5 generates concise and meaningful abstractive summaries from the selected content. Unlike traditional extractive methods, this approach focuses on capturing the true intent of the text and expressing it in a simplified manner. The model was trained using a dataset consisting of legal and financial documents along with their summaries. Due to computational limitations in Google Collab, a smaller dataset was used, containing 2,000 training samples and 300 testing samples. Despite these constraints, the system achieved a Rouge-1 score of 30%, indicating promising performance. The results highlight that even with limited resources, transformer-based models can produce useful summaries for domain-specific texts. This framework offers a scalable and cost-effective solution, with potential for significant improvement when trained on larger datasets & powerful hardware.**

***Keywords*— AI-powered summarization, Natural Language Processing, COLBERT, T5 transformer, abstractive summarization, legal documents, financial texts, contextual understanding, sentence ranking, Rouge score evaluation, deep learning, text generation.**

I. INTRODUCTION

The rapid growth of digital information in legal and financial domains has created a strong need for efficient document processing systems. Legal documents, contracts, case reports, and financial statements are often lengthy, complex, and filled with domain-specific terminology, making manual analysis time-consuming and error-prone. As a result, researchers have increasingly focused on applying Natural Language Processing (NLP) and deep learning techniques to automate tasks such as document retrieval, classification, and summarization. Early work in this area explored document representation using vector embeddings and deep learning methods to improve retrieval accuracy in legal corpora [1]. Similarly, the use of word embeddings has enhanced full-text search capabilities by capturing semantic relationships within legal documents [2].

With the advancement of neural networks, several studies have investigated the use of deep learning models for legal text classification and analysis. Comparative studies have shown that neural architectures outperform traditional methods in classifying legal documents due to their ability to

capture contextual information [3]. Deep learning techniques have also been applied in large-scale legal document review processes, demonstrating improved efficiency and accuracy [4]. Additionally, convolutional neural networks have been successfully used for document type classification in judicial systems, highlighting their effectiveness in handling structured legal data [5].

Recent research has expanded beyond classification to more complex tasks such as translation, summarization, and multi-task learning. Multi-task deep learning frameworks have shown promise in simultaneously performing multiple legal text processing tasks, including summarization [6]. Other studies have explored unsupervised approaches to analyze relationships in legal language, further enhancing understanding of document structure [7]. Sequence-based models such as LSTM and GRU have also been applied to predict judicial decisions and evaluate case outcomes, demonstrating the potential of deep learning in legal reasoning [8][14].

Moreover, identifying logical components within legal texts and extracting key contract elements have become important research areas, supported by neural network-based approaches [9][10]. Classification of linguistic features such as modality in legal language has also contributed to improved interpretation of regulatory documents [11]. Active learning strategies and ranking-based methods have further optimized legal document review and question-answering systems [12][13].

Despite these advancements, generating concise and meaningful summaries of legal and financial documents remains a challenging task due to their complexity and length. This project aims to address this challenge by leveraging advanced transformer-based models to produce accurate and context-aware summaries, thereby improving accessibility and decision-making efficiency in professional domains.

II. RELATED WORK

Sugathadasa et al. [1] proposed a deep learning approach for improving legal document retrieval using vector embeddings. Traditional keyword-based search methods often fail to capture the true meaning of queries, leading to less relevant results. To overcome this limitation, the authors introduced document embeddings, where legal texts are represented as numerical vectors that preserve semantic relationships. This allows the system to understand context rather than relying only on exact word matches. By applying deep learning techniques, the model can identify similarities between documents even when different terminology is used. The study demonstrated that embedding-based retrieval significantly improves accuracy and relevance in legal searches. Their work highlights the importance of semantic understanding in handling complex legal language. This approach has influenced further research in legal NLP tasks such as classification and summarization, as it provides a strong foundation for building systems that require deeper contextual analysis and improved information retrieval efficiency in large document.

Landthaler et al. focused on enhancing legal document search by incorporating word embeddings into full-text search systems. Traditional search engines rely heavily on keyword matching, which limits their ability to identify documents with similar meanings but different wording. The authors addressed this issue by representing words as vectors in a semantic space, allowing the system to capture relationships between terms. This enables more flexible and accurate search results, especially in legal contexts where similar concepts may be expressed differently. Their approach improves retrieval performance by identifying contextually relevant documents rather than exact matches. The study showed that integrating word embeddings leads to better search quality and user satisfaction. This work is important because it demonstrates how semantic understanding can enhance information retrieval systems. It also laid the groundwork for more advanced NLP techniques, encouraging the use of deep learning models in legal applications such as summarization, classification, and intelligent document analysis. [2]

Undovia et al. conducted a comparative study on legal document classification using neural network models. The aim was to evaluate how effectively deep learning techniques can categorize legal texts compared to traditional machine learning approaches. Legal documents are often complex and require accurate classification for efficient organization and retrieval. The authors experimented with different neural network architectures and found that deep learning models

perform better in capturing contextual and semantic information. This leads to improved classification accuracy and better generalization. The study also highlighted challenges such as the need for large labeled datasets and handling class imbalance. Despite these issues, neural networks proved to be more reliable than conventional methods. The findings emphasize the growing importance of deep learning in legal text processing. This research supports the adoption of advanced AI models for handling complex documents and contributes to the development of systems for summarization.[3]

Wei and Qin presented an empirical study on applying deep learning techniques to legal document review, focusing on text classification tasks. Legal document review involves analyzing large volumes of text, which can be time-consuming and labor-intensive. The authors proposed using neural networks to automate this process and improve efficiency. Their approach enables the system to learn patterns and relationships within the text, resulting in more accurate classification. The study compared deep learning models with traditional methods and found that deep learning achieved better performance in terms of precision and recall. It also demonstrated the ability to handle large datasets effectively. However, the research pointed out challenges such as the need for significant computational resources and high-quality training data. Overall, the study highlights the potential of deep learning in transforming legal document review by reducing manual effort and improving accuracy. It also supports the use of AI-based solutions in complex text processing tasks.[4]

Da Silva introduced a convolutional neural network-based approach for classifying legal documents from Brazil's Supreme Court. The study aimed to improve the organization and management of judicial data by accurately identifying document types. Traditional classification methods often struggle with the complexity of legal language, making automated solutions necessary. The proposed CNN model processes text by extracting important features through convolutional layers, which helps in identifying patterns within the data. This approach allows the system to effectively distinguish between different categories of legal documents. The results showed that the CNN model achieved higher accuracy compared to traditional techniques. The study demonstrates the effectiveness of deep learning in handling structured and unstructured legal text. It also highlights the potential of CNNs for other applications such as summarization and information extraction. This work contributes to the advancement of AI in legal systems by providing a reliable method for document classification and analysis.[5]

III. PROPOSED METHODOLOGY

The proposed system adopts a hybrid architecture that integrates COLBERT and T5 models to perform efficient and accurate document summarization. The process begins with preprocessing the input text, where unnecessary symbols and noise are removed to improve data quality. The cleaned text is then passed to the COLBERT model, which uses contextual embeddings to analyse and rank sentences based on their importance within the document. This step ensures that only the most relevant content is selected for further processing. Once the important segments are identified, they are forwarded to the T5 model. T5, a transformer-based sequence-to-sequence model, converts the selected content into a concise and coherent summary. Unlike extractive methods, T5 generates new sentences, making the summary more natural and easier to understand. The system is implemented in Python using Google Colab, making it accessible and cost-effective. It also includes a simple user interface where users can sign up, log in, train the model, and generate summaries for custom input text. This structured approach enhances both usability and performance, making the system suitable for real-world legal , financial applications.

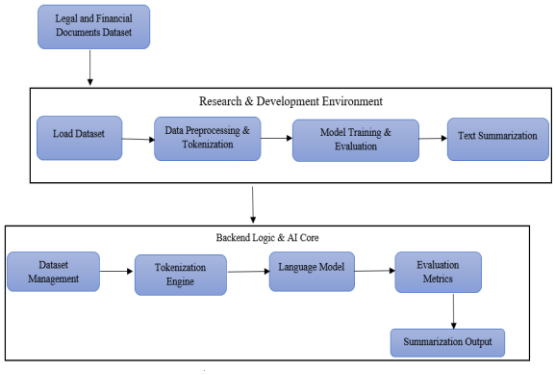

Figure [1]: System Architecture

Figure [1] illustrates a specialized text summarization pipeline tailored for legal and financial documents, divided into two distinct execution layers. At the top, the Research & Development Environment handles the initial heavy lifting by loading the raw dataset, executing pre-processing and tokenization, and training the model to handle complex domain-specific text. This phase flows directly into the Backend Logic & AI Core below, which represents the operational system. In this production-oriented layer, managed data is fed through a tokenization engine and processed by the language model, where it is measured against specific evaluation metrics to ultimately produce the final summarization output.

IV.RESULT AND DISCUSSION

The performance of the proposed system was evaluated using standard metrics such as Rouge scores, which measure the similarity between generated summaries and reference summaries. The model achieved a Rouge-1 score of approximately 30%, indicating that the generated summaries captured a reasonable portion of the key

information from the original text. In some test cases, the system also demonstrated high recall values, suggesting that important content was successfully retained during summarization. The results show that combining COLBERT with T5 improves both sentence selection and summary generation. COLBERT effectively identifies relevant parts of the document, while T5 transforms them into meaningful summaries. This combination addresses the limitations of traditional methods, which often fail to capture deep contextual meaning. However, the performance is somewhat limited by the size of the dataset and computational resources. Training on a larger dataset with better hardware would likely improve accuracy and fluency. Additionally, domain-specific fine-tuning could further enhance performance for complex legal language.

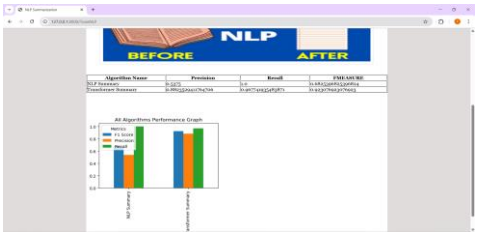

Figure [2]: NLP Summary & Transformer Summary

Figure [2] illustrates This image displays the Model Performance Comparison screen for the summarization framework. It presents a table and a corresponding bar chart evaluating two approaches: "NLP Summary" and "Transformer Summary". The algorithms are compared using three standard statistical evaluation metrics: Precision, Recall, and F-Measure (F1 Score), with the Transformer model demonstrating higher overall balanced performance.

Figure [3]: Summary Generation Screen

Figure [3] illustrates Summary Generation Screen of the web application. It features a text area labelled "Input Text" where users can paste their long-form documents. A paragraph discussing machine learning models and cybersecurity has been entered into the field, and a "Generate Summary" button is located directly beneath it to trigger the AI processing.

Figure [4]: Output Screen

Figure [4] illustrates Output Screen of the "AI-Powered Document Summarization Framework for Legal and Financial Texts." It displays the original "Input Text" that was submitted in the previous step, now laid out on the page. In a full view, the generated concise summary produced by the system would typically appear directly below this section.

DISCUSSION

The performance of the proposed system was evaluated using standard metrics such as Rouge scores, which measure the similarity between generated summaries and reference summaries. The model achieved a Rouge-1 score of approximately 30%, indicating that the generated summaries captured a reasonable portion of the key information from the original text. In some test cases, the system also demonstrated high recall values, suggesting that important content was successfully retained during summarization. The results show that combining COLBERT with T5 improves both sentence selection and summary generation. COLBERT effectively identifies relevant parts of the document, while T5 transforms them into meaningful summaries. This combination addresses the limitations of traditional methods, which often fail to capture deep contextual meaning. However, the performance is somewhat limited by the size of the dataset and computational resources. Training on a larger dataset with better hardware would likely improve accuracy and fluency. Additionally, domain-specific fine-tuning could further enhance performance for complex legal language. Overall, the system demonstrates promising results and validates the effectiveness of transformer-based approaches in document summarization tasks

V. CONCLUSION

This project successfully demonstrates the use of advanced AI models for summarizing legal and financial documents. By combining COLBERT and T5, the system achieves a balance between accurate content selection and natural language generation. The approach overcomes many limitations of traditional summarization techniques, particularly in handling complex and lengthy texts. The implementation on Google Collab shows that effective models can be developed even with limited computational resources. Although the dataset used for training was relatively small, the system still produced meaningful summaries with acceptable performance metrics. This highlights the potential of transformer-based models in real-world applications. Future improvements can focus on increasing the dataset size, optimizing model parameters, and using more powerful GPUs to enhance performance. Additional features such as multilingual support and domain-specific customization can further expand the system's usability. In conclusion, the proposed framework provides a strong foundation for intelligent document summarization. It offers a practical solution for professionals dealing with large volumes of legal and financial data, saving time and improving information accessibility.

REFERENCES

[1]. Sugathadasa, K.,etal.:"Legal document retrieval
using document vector embeddings and deep learning". In:Arai,K.Kapoor,S.,Bhatia,R.SAI 2018.
[2]. Landthaler. J., Waltl,B.,Holl.P.,Mathes,F.":Extending full text search for legal document collections using word embeddings" (2016)
[3]. Undovia,S.,Meyers,A.,J.E: "A Comparative study of classifying legal documents with neural networks". (2018).
[4]. Wei,F.,Qin,H.":Empirical study of deep learning for text classification in legal document review". (2018).
[5]. Da Silva, N.C.:Document type classification for Brazil's supreme court using a Convolutional Neural Network. (2018).

[6]. Elnaggar, A., Gebendorfer,C.,Glaser,I., Matthes, F.: "Multi-task classification. Multi-task deep learning for legal document translation, summarization" (2018).
[7]. Wang, Y.: "An unsupervised approach to relatedness analysis of legal language". Master's thesis, University of Waterloo (2018).
[8]. Li, S.,Zhang, H., Ye,L.,Geo,X.,Fang,B.:"Evaluating the rationality of judicial decision with LSTM-based case modeling" (2018).
[9]. Son, N.T., Nguyen, L.M., Quoe, H.B., Shumazu, A.:" Recognizing logical parts in legal texts using neural architectures". (2016)
[10]. Chalkidis, I., Androutsopoulos, I: "A deep learning approach to contract element extraction". (2017)
[11]. Neil, J.O., Buitelaar, P., Robin, C., Brien, L.O.: "Classifying sentential modality in legal language: a use case in financial regulations acts and directives". (2017).
[12] "Empirical evaluation of Active learning strategies in legal document review," by Rishi Chahtwal.
[13]. Do, P.K., Nguyen, H.T., Tran, C.X., Nguyen, M.T., Nguyen, M.L.: Legal question answering using ranking SVM and deep convolutional neural network. (2017).
[14]. Kowsrihawat, K., Vateekul, P., Boonkwan, P.: Predicting judicial decisions of criminal cases from thai supreme court using bi-directional GRU with attention mechanism. In: 5th Asian Conference on Defense Technology (ACDT), pp. 50–55. IEEE, Vietnam (2018)
[15]. Hashianov, A., Alomava, I., Nurhumbetvo, G.: lawyers intellectual tool for analysis of legal documents in Russian. (IC-AIAI) International Conference on Artificial Intelligence Applications and Innovations (2018)
Babburi, S. Lightweight Distributed Provenance Framework for Edge and IoT Data Systems.

Mahimalur, R. K., Vasgam, M., & Manoharan, D. Devops Lifecycle Management And Cloud Migration Assessments: A Security-Driven CICD Perspective.

Bajarang Bhagwat, V. (2023). Optimizing Payroll to General Ledger Reconciliation: Identifying Discrepancies and Enhancing Financial Accuracy. JOURNAL OF ADVANCE AND FUTURE RESEARCH, 1(4). https://doi.org/10.56975/jaafr.v1i4.501636

Manoharan, D. (2025). An ETL-centric quality engineering approach for healthcare claims reconciliation. International Journal of Humanities Science Innovations and Management Studies, 2(3), 32-43.

Poojari, R. INTELLIGENT SYSTEMS+B108 AND APPLICATIONS IN ENGINEERING
.
Maturi, S. Y. (2021). Blockbond hardening: Securing pooled-hash protocols against traffic tampering, MITM hash-rate hijacking, and template coercion. International Journal of Communication Networks and Information Security, 13(3), 718–728.

Gaddam, S. From Fixed Specifications to Self-Adapting Systems: A Machine Learning Perspective on Software Engineering.

Maturi, S. Y. (2023). Crowdsourced frontier: Unveiling autonomous adversarial cybercapabilities via open AI competition. International Journal of Intelligent Systems and Applications in Engineering, 11(1s), 275–284.

Gummadi, V. P. K. (2026). Infrastructure optimization techniques for enterprise integration platforms: A comprehensive analysis. Computer Fraud & Security, 37-44. https://doi.org/10.52710/cfs.875

Shashank, A. (2025). Self-Healing Data Pipelines for Enhanced Reliability: A Paradigm Shift in Enterprise Data Management. Journal of Computer Science and Technology Studies, 7(8), 1097-1104.

Adabala, P. K. (2024). Utilizing predictive analytics to improve efficiency and decision-making in ERP-connected supply chains. International Journal of Intelligent Systems and Applications in Engineering, 12(22s), 2465

Venkata Ramana, P. (2024). AI-driven predictive analytics in ERP systems for proactive supply chain optimization. International Journal of Research in Information Technology and Computing, 8(4).

Srikanth Kavuri. (2023). Machine Learning Approaches for Security Vulnerability Detection in Software Testing. Computer Fraud and Security. https://doi.org/10.52710/cfs.837

Shashank, A. (2025). AI-Enhanced ETL Processes: Leveraging Artificial Intelligence for Optimized Data Integration Systems. Journal Of Multidisciplinary, 5(8), 219-225.

Gummadi, V. P. K. (2023). MuleSoft batch processing: High-volume streaming architecture. Computer Fraud & Security, 50-57. https://doi.org/10.52710/cfs.886

Harshitha, G. K., Nandigama, C., & Thiripalu, P. (2026). An exploration into identification of opportunities and challenges of establishing and running an enterprise in the area of biofuels. Minnesota Journal of Business Law and Entrepreneurship, 2026(1), 1159–1168.

Susarla, R. S., Boyapati, P. K., & Kandula, S. T. R. (2025, July). Cloud-Based Secure Data Storage in Smart Cities Using Central-Smoothing Hypergraph Neural Networks. In 2025 IEEE 4th World Conference on Applied Intelligence and Computing (AIC) (pp. 279-284). IEEE.

Boyapati, P. K. Building a centralized data operations hub for healthcare enterprise integration. IJSAT-Int. J. Sci. Technol. 16 (2). https://doi.org/10.71097/IJSAT.v16.i2.3219

Kolla, S. H., Nagabhyru, K. C., & Kummari, D. N. (2026). Letter to the Editor re:“CurvAssist: An AI assisted pipeline for penile shaft segmentation/curvature measurement in children with hypospadias”. Journal of Pediatric Urology.

Kumar, Anuj. "Optimization of Machining Parameters in Turning Operations for Surface Roughness and Tool Wear." International Journal of Engineering Research and Science & Technology, Vol. 12, No. 4, 2016, pp. 47-64

Chawla, N., Nandini, M. R., Pokala, H. K., & Siddartha, A. (2026, April). Resource Scheduling in Cloud-Assisted Communication Systems Using Enhanced Reinforcement Learning Algorithm. In 2026 2nd International Conference on Intelligent Systems and Computational Networks (ICISCN) (pp. 1-6). IEEE.

Kumar, Anuj. "Battery Thermal Management Systems for Electric Vehicles Using Phase Change Materials." International Journal of Engineering, Science and Mathematics, Vol. 10, Issue 3, 2021, pp. 202-211