ISSN 2277-2685
IJESR/April-June. 2025/ Vol-15/Issue-2/113-123
J Satya Sai *et. al.*, / International Journal of Engineering & Science Research

Developing An NLP Model For Efficient Summarization Of Legal & Financial Documents

Mr. D Lakshminarayana Reddy [1], J Satya Sai [2]

[1] Assistant Professor, Dept. of CSE, Anantha Lakshmi institute of technology & sciences, JNTUA, Anantapur, India

[2] PG Student, Dept. of CSE, Anantha Lakshmi institute of technology & sciences, JNTUA, Anantapur, India.

[1]lakshmi1217@gmail.com, [2] sathyasai3614@gmail.com

ABSTRACT

The proliferation of legal and financial documents in the digital age has necessitated the development of efficient and accurate summarization techniques to manage the overwhelming volume of information. This paper proposes the development of a Natural Language Processing (NLP) model specifically designed for the summarization of legal and financial documents. The proposed model leverages advanced machine learning techniques, including transformer architectures and deep learning algorithms, to generate concise and coherent summaries. These summaries aim to retain the critical information and key insights of the original documents while significantly reducing their length. The model is trained on a large, domain-specific dataset to enhance its understanding of the unique terminology, structures, and nuances inherent in legal and financial texts. Special attention is given to ensuring the model's outputs are not only syntactically accurate but also contextually relevant and legally sound, which is crucial in professional settings. Evaluation metrics such as ROUGE and BLEU scores are employed to assess the performance of the model, alongside human evaluation by legal and financial experts to ensure practical applicability and accuracy.

1. INTRODUCTION

In an era defined by information overload, legal and financial professionals are increasingly confronted with the challenge of parsing vast volumes of dense, technical documents. Legal contracts, court rulings, regulatory filings, and financial statements often span dozens or even hundreds of pages, requiring extensive time and expertise to interpret. As organizations strive for operational efficiency and timely decision-making, the demand for intelligent tools that can distill these complex documents into concise, accurate summaries has never been greater.

Natural Language Processing (NLP) offers a promising solution to this problem. Recent advancements in NLP—particularly in the areas of text summarization and transformer-based language models—enable machines to process and understand human language with increasing sophistication. However, applying NLP to legal and financial domains presents unique challenges. These documents are not only long and information-dense but also demand a high degree of precision, as even minor omissions or misinterpretations can lead to significant legal or financial consequences.

This project aims to develop a domain-specific NLP model capable of performing efficient and reliable summarization of legal and

financial documents. The goal is to reduce the time and cognitive load required to review such documents while

ensuring that critical information is preserved. By leveraging a combination of extractive and abstractive summarization techniques, pre-trained language models, and domain-specific datasets, this work seeks to bridge the gap between raw textual data and actionable insight.

Through this research, we hope to contribute to the growing field of intelligent document understanding and provide tools that can aid professionals in law and finance by streamlining their workflows, reducing human error, and enabling faster, more informed decisions.

1.1 Related Work

Text summarization has seen significant advances with the development of transformer-based language models. Summarization techniques fall into two broad categories:

Extractive Summarization selects salient sentences or phrases directly from the source text. Classical methods include: TextRank (Mihalcea & Tarau, 2004) LexRank (Erkan & Radev, 2004) Abstractive Summarization generates novel sentences that convey the original meaning using paraphrasing: Seq2Seq with attention (Bahdanau et al., 2014) Transformer-based models like: BART (Lewis et al., 2020) T5 (Raffel et al., 2020) PEGASUS (Zhang et al., 2020), specifically designed for summarization These general-purpose models, however, often struggle with domain-specific jargon and the preservation of legal and financial nuance.

2. LITERATURE SURVEY

Recent advances in Natural Language Processing (NLP) have significantly improved the performance of automatic text summarization systems. This literature survey explores key methodologies, models, and datasets that have influenced the development of summarization systems in the legal and financial domains.

1. Summarization Techniques

a. Extractive Summarization

Extractive methods identify and select the most relevant sentences from the source document to form a summary. These methods are often based on statistical or graph-based algorithms.

- TextRank (Mihalcea & Tarau, 2004): A graph-based ranking model inspired by PageRank, widely used in general-purpose summarization.

- LexRank (Erkan & Radev, 2004): An unsupervised method using sentence similarity graphs and eigenvector centrality.

While extractive models are efficient and less prone to factual errors, they often lack fluency and may miss implicit meanings, particularly in legal or financial texts.

b. Abstractive Summarization

Abstractive models generate novel sentences that capture the document's meaning, requiring deep semantic understanding and natural language generation.

- Seq2Seq with Attention (Rush et al., 2015): Introduced neural abstractive summarization, but struggled with long documents.

- Pointer-Generator Networks (See et al., 2017): Balanced copying from the source text with generation, improving factual correctness.
- Transformer-based Models: Models like BART (Lewis et al., 2020), T5 (Raffel et al., 2020), and PEGASUS (Zhang et al., 2020) have shown state-of-the-art results in abstractive summarization.

2. Domain-Specific Applications

a. Legal Document Summarization

Legal texts require preserving factual integrity and contextual nuances. Research in this domain focuses on case law, statutes, and legal opinions.

- BillSum (Kornilova & Eidelman, 2019): A dataset of U.S. Congressional bills paired with human-written summaries; demonstrated the applicability of BART and PEGASUS.
- CaseHOLD (Zheng et al., 2021): A benchmark for predicting the holding of U.S. court cases using legal reasoning.
- Legal-BERT (Chalkidis et al., 2020): A BERT model pre-trained on legal corpora, improving performance on legal-specific NLP tasks.

b. Financial Document Summarization

Summarizing financial reports such as 10-K filings and earnings calls demands accuracy, particularly in extracting key performance indicators and risk factors.

- FinSum (Nassif et al., 2020): A dataset for summarizing financial news and earnings call transcripts.
- FinancialBERT (Yang et al., 2020): A domain-specific language model trained on financial texts, which enhances downstream summarization and sentiment tasks.

3. Challenges in Legal and Financial Summarization

- Length of Documents: Legal and financial documents often exceed token limits of standard transformer models.
- Domain Jargon: Specialized vocabulary and context-specific meanings make general-purpose models less effective.
- Factual Consistency: Especially important in these domains to avoid misleading or incorrect summaries.
- Section Structure: Legal and financial documents often have well-defined sections (e.g., “Risk Factors”, “Conclusion”) that can guide summarization but require structural parsing.

4. Emerging Trends

- Long-Document Transformers: Models like Longformer, BigBird, and LED (Longformer Encoder-Decoder) allow handling input sequences beyond 4,096 tokens, which is vital for processing entire legal/financial documents.
- Reinforcement Learning for Summarization (e.g., RLHF): Used to align outputs with human preferences, improve coherence, and reduce hallucinations.
- Hybrid Systems: Combining extractive preprocessing with abstractive summarization improves efficiency and quality (e.g., extract key paragraphs, then summarize).

2.1 Limitations of Existing System

The current systems for summarizing legal and financial documents primarily rely on both extractive and early-stage abstractive summarization techniques. Extractive summarization methods, which involve selecting key sentences or phrases directly from the source text, are widely used due to their simplicity and ease of implementation. These systems often utilize statistical approaches, such as TF-IDF (Term Frequency-Inverse

Document Frequency), or machine learning algorithms, such as Support Vector Machines (SVM) and neural networks, to identify the most relevant sentences. However, extractive methods tend to produce summaries that lack coherence and fluidity, as they merely piece together parts of the original text without rephrasing or integrating the information contextually. Recent advancements have seen the integration of transformer-based models like BERT, GPT, and their variants, which have shown promise in generating more coherent and contextually relevant summaries. These models use sophisticated attention mechanisms to understand the context and semantics of the text, allowing them to generate more meaningful summaries. Despite these advancements, current abstractive models still face challenges in accurately summarizing complex and specialized texts, such as legal and financial documents. They often struggle with the nuances of legal language and the precise nature of financial data, leading to summaries that may overlook critical details or misinterpret key information.

3. PROPOSED WORK

The proposed system aims to overcome the limitations of current summarization models by leveraging advanced NLP techniques, specifically designed to handle the complexity and specialized requirements of legal and financial documents.

At the core of this system is a transformer-based architecture, fine-tuned on extensive domain-specific datasets to ensure a deep understanding of the unique language, structures, and nuances inherent in these texts. The model will employ a hybrid approach, integrating both extractive and abstractive summarization methods to generate summaries that are not only concise and coherent but also contextually accurate and legally sound. To address size and fit issues, the proposed system includes enhanced virtual fitting rooms and 3D body scanning technology, allowing consumers to create precise digital avatars for accurate sizing recommendations. This technology will significantly reduce the rate of returns due to incorrect sizing. A key feature of the proposed system is its ability to accurately summarize numerical data and integrate it with textual content. This involves the use of advanced algorithms that can interpret and contextualize financial metrics, ensuring that summaries of financial documents accurately reflect key numerical insights alongside textual analysis. Additionally, the model will incorporate a robust mechanism for maintaining legal accuracy, using specialized legal language models that can interpret and preserve the integrity of legal clauses and terminology. The system will also include customizable settings to tailor summaries to different contexts, user needs, and levels of detail. This flexibility is essential for professionals who require specific types of information from legal and financial documents for various purposes. Furthermore, the proposed system will leverage advanced encryption and compliance with data protection regulations to ensure the security and privacy of sensitive information.

3.1 System Overview

Proposed System Advantages The proposed NLP model for summarizing legal and financial documents offers numerous advantages over existing systems, addressing key limitations and enhancing the overall efficiency and accuracy of the summarization process. Improved Coherence and Fluidity by integrating both extractive and abstractive summarization techniques, the proposed system generates summaries that are not only concise but also coherent and fluid. This hybrid approach ensures that the summaries provide a seamless narrative flow, making it easier for users to understand the overall context and key insights of the documents. Enhanced Handling of Complex Language and Terminology The proposed system is fine-tuned on extensive domain

specific datasets, enabling it to accurately interpret and summarize complex legal and financial texts. This specialized training ensures a deep understanding of the unique terminology, structures, and nuances inherent in these documents, resulting in summaries that retain critical details and accurately represent the original content.

3.2 System Architecture and Integrated Algorithm Workflow

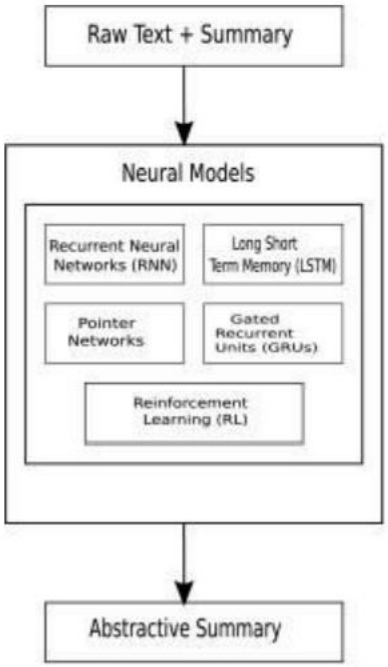


Fig No 1: System Architecture diagram of Service Provider

3.3 Preprocessing and Textual Data Extraction

Legal Documents:

- Formats: Commonly found as PDF, DOCX, or HTML (e.g., court rulings, contracts, legal filings)
- Sources:
  - Open-access legal repositories (e.g., CourtListener, BillSum, EU Lex)
  - Corporate contract databases (for internal use)

B. Financial Documents:

- Formats: SEC filings (10-K, 10-Q), earnings calls, annual reports — typically HTML, PDF, or TXT
- Sources:
  - SEC's EDGAR database
  - Financial news outlets (e.g., Bloomberg, Reuters)
  - Transcription services (for earnings calls)

C. Extraction Tools:

- PDF: pdfminer, PyMuPDF, or Adobe PDF Extract API
- DOCX/HTML: python-docx, BeautifulSoup
- OCR (for scanned docs): Tesseract OCR with layout-aware parsing (e.g., LayoutParser)

3.4 System Objectives

The primary objectives of the proposed system are as follows

Objective 1: Ensure High System Reliability

Implementing long-document transformer models (e.g., Longformer, LED, or BigBird) capable of handling sequences up to 16,384 tokens.Incorporating hierarchical summarization pipelines to deal with documents exceeding model input limits.

Retaining document structure (e.g., headings, clauses, numbered sections) during preprocessing to ensure context-aware summarization.

Objective 2: Enhance Academic Integrity

Using domain-specific pretrained models such as: Legal-BERT, CaseLaw-BERT for legal documents FinBERT, Financial-RoBERTa for financial documents Integrating Named Entity Recognition (NER) and numerical value tracking to ensure preservation of critical facts (e.g., monetary values, legal parties). Incorporating factual consistency checkers (e.g., SummaC or QAGS) into the evaluation pipeline.

Objective 3: Improve Security and Authentication

Designing a flexible preprocessing pipeline that handles:

PDFs (with PyMuPDF or pdfminer)

HTML/EDGAR filings (with BeautifulSoup)

Scanned documents via OCR (with Tesseract and LayoutParser)

Standardizing the extracted text into a unified JSON format for ease of ingestion by downstream models**.**

3.5 Achieving the Objectives

Finally for Objective 3, Training and evaluating both: Abstractive models (BART, PEGASUS, T5, LED) Extractive models (TextRank, BERTSUM, LexRank) for use in hybrid approaches Combining extractive sentence scoring with abstractive rewriting for enhanced precision and readability. Objective: Evaluate Model Effectiveness with Reliable Metrics Utilizing standard summarization metrics: ROUGE-1, ROUGE-2, ROUGE-L for lexical overlap BERTScore for semantic similarity Conducting manual evaluation to assess readability, factual accuracy, and legal/financial compliance Implementing a custom rubric for domain-specific quality (e.g., "preservation of risk clauses" or "accuracy of numerical reporting")

4. SYSTEM MODULES & ALGORITHMS

1. Document Ingestion Module

Function: Handles input from various formats (PDF, DOCX, HTML, plain text) and extracts raw text.

- Subcomponents:
  - File upload interface / batch ingestion
  - Text extractor (using pdfminer, docx, BeautifulSoup, etc.)
    - OCR handler (optional, for scanned documents via Tesseract)

2. Preprocessing Module

Function: Cleans and prepares the raw text for NLP processing.

- **Tasks:**
  - Sentence segmentation
  - Tokenization
  - Stopword removal (if applicable)
  - Legal/financial entity recognition (NER)
  - Section segmentation (e.g., for 10-K: "Risk Factors", "MD&A", etc.)

Tools: SpaCy, NLTK, custom rule-based scripts

3. Summarization Engine

Function: The core module that generates the summaries.

Modes:

- Extractive: Uses BERTSum, TextRank, or LexRank
- Abstractive: Utilizes pre-trained transformers like BART, T5, PEGASUS, or Longformer
- Hybrid: Extracts key segments and then summarizes them abstractively
- Customization:
  - Domain-specific fine-tuning (e.g., Legal-BERT, FinancialBERT)

4. Postprocessing Module

Function: Refines and formats the output summaries.

- **Features:**
  - Remove redundancy
  - Check factual consistency (optional — using QA models or sentence similarity)
  - Highlight keywords or named entities
  - Format summary into structured output (paragraph, bullet points, etc.)

5. Evaluation Module

Function: Assesses the quality and performance of the summarization output.

- Metrics:
  - ROUGE, BLEU, METEOR
  - Human-in-the-loop evaluation (e.g., domain expert feedback)
  - Factual consistency (using QA-style validation or truthfulness scoring)

6. User Interface Module

Function: Provides a front-end for user interaction.

- Types:
  - Web-based UI (Streamlit, Flask, React)
  - Command-line interface (CLI)
  - RESTful API for integration with other systems
- Features:
  - Document upload
  - Summary display

- Export options (TXT, PDF, DOCX, JSON)

7. Model Management Module

Function: Manages NLP models used in the system.

- Tasks:
- Load, fine-tune, and switch between summarization models
- Monitor GPU/CPU performance
- Schedule retraining or updates

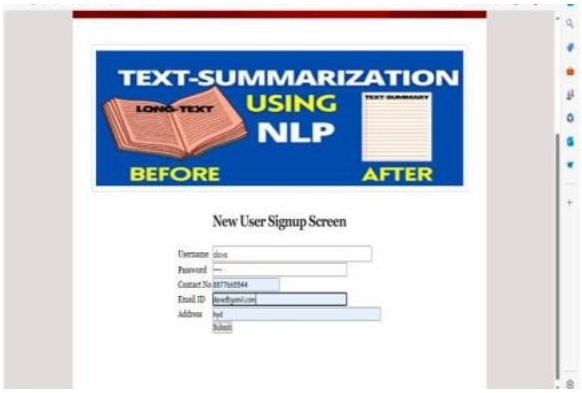
8. Security & Access Control Module

Function: Ensures safe handling of sensitive documents.

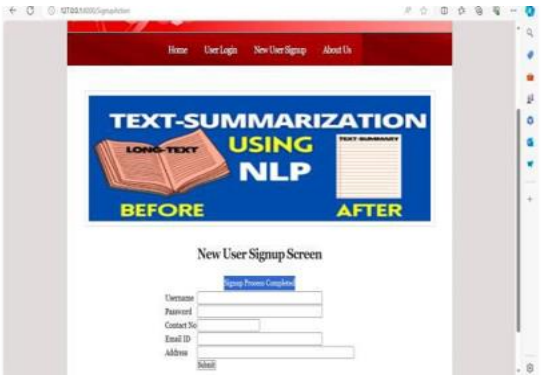
- Features:
- User authentication (if web-based)
- Role-based access (e.g., viewer, editor)
- Document encryption & secure storage

5. RESULTS

In above screen click on 'New User Sign up' link to get below page



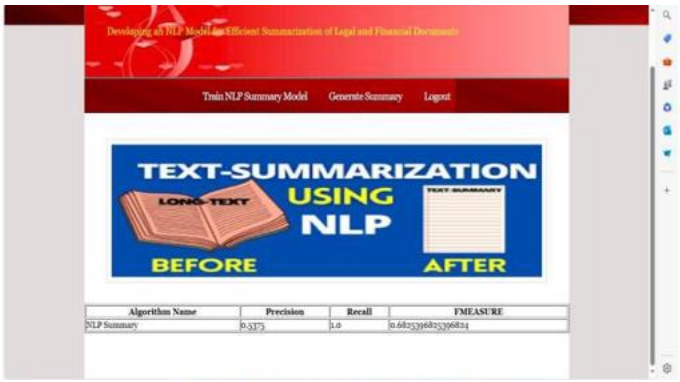
In above screen user is entering sign up details and then press button to get below page



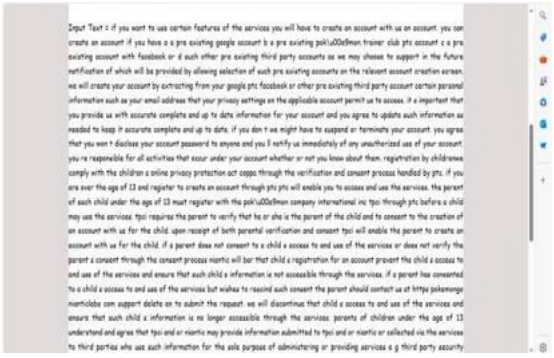
In above screen sign up task completed and now click on 'User Login' link to get below page

In above screen user is login and after login will get below page

In above screen user can click on 'Train NLP Summary Model' link to train NLP algorithm on LEGAL dataset and get below output



In above screen can see performance metrics of NLP on summary generation and got recall as 100%. Now click on "Generate Summary" link to get below page In above screen you can enter some text and then click on 'Generate Summary' button to get below summary output



In above screen can see INPUT TEXT Data and below is the summary generated from above TEXT data

In above screen in second para can see generated summary from given long text and similarly you can give any text to get summary

6. CONCLUSION

The development of an NLP model for the efficient summarization of legal and financial documents addresses a critical need in managing the overwhelming volume and complexity of information in these domains. The proposed system leverages advanced NLP techniques, including transformer-based architectures and domain specific training, to generate summaries that are not only concise and coherent but also contextually accurate and legally sound. By integrating both extractive and abstractive summarization methods, the model overcomes the limitations of current systems, providing a more seamless and fluid narrative in the summaries. The specialized training on legal and financial corpora ensures a deep understanding of the unique terminology, structures, and nuances of these documents, resulting in summaries that retain essential details and accurately represent the original content. The system's ability to accurately summarize numerical data and integrate it with textual content is particularly crucial for financial documents, ensuring that key financial metrics and insights are effectively communicated. Additionally, the focus on maintaining legal and numerical accuracy enhances the reliability and usability of the summaries in professional contexts.

7. FUTURE SCOPE

Description: Legal and financial documents are often written in multiple languages, especially in international settings. Expanding the system's capability to summarize documents in **multiple languages** could make it applicable to a wider range of users globally.

Future Directions:

Train the model on multilingual corpora to support languages such as **Spanish**, **French**, **German**, **Mandarin**, and others.

Use **language-specific models** like **mBERT** (Multilingual BERT) or **XLM-R** to ensure high-quality summaries across diverse languages.

Benefits:

Enables global adoption for legal and financial professionals.

Provides summaries for non-English documents, improving accessibility.

2. Real-Time Summarization

Description: Enhancing the system's ability to provide **real-time summarization** as documents are being read or processed would increase its practicality in dynamic environments like courtrooms, board meetings, and financial presentations.

Future Directions:

Implementing a **streaming summarization** model that can process text incrementally as it's being uploaded or read.

Support real-time summarization of documents or articles directly from web sources or live reports.

Benefits:

Immediate access to summaries during critical meetings or decision-making processes.

Improved user experience for professionals who need quick insights during fast-paced tasks.

REFERENCES:

1. Chalkidis, I., Androutsopoulos, I., & Aletras, N. (2019). Neural legal judgment prediction in English. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4317-4323). https://doi.org/10.18653/v1/P19-1424

2. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. https://arxiv.org/abs/1810.04805

3. El-Haj, M., Rayson, P., & Walker, M. (2019). An automatic extraction method for financial information in financial narratives. Computers in Industry, 105, 80-90. https://doi.org/10.1016/j.compind.2018.11.0 02

4. Grover, C., McDonald, R., & Tobin, R. (2003). A configurable system for document processing. In Proceedings of the EACL Workshop on Language Technology and the Semantic Web (NLPXML-2003) (pp. 81- 88). https://www.aclweb.org/anthology/W03- 1104

5. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pretraining for natural language generation, translation, and comprehension. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 7871-7880). https://doi.org/10.18653/v1/20

20.aclmain.703

6. Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In Text Summarization Branches Out: Proceedings of the ACL-04 Workshop (pp. 74-81). https://www.aclweb.org/anthology/W04- 1013

7. Liu, Y., Liu, F., Chua, T.-S., & Sun, M. (2018). Structure-preserving embedding for financial documents. In Proceedings of the 27th International Conference on Computational Linguistics (pp. 527-537). https://www.aclweb.org/anthology/C18- 1045

8. Nallapati, R., Zhai, F., & Zhou, B. (2017). SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (pp. 3075-3081). https://www.aaai.org/ocs/index.php/AAAI/ AAAI17/paper/view/14650

9. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., ... & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1-67. http://jmlr.org/papers/v21/20-074.html

10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008). https://arxiv.org/abs/1706.03762