Machine Learning for Human Intelligence
2025, Vol. 3, Issue 2

Research Article

Advancing Legal Text Summarization with Deep Learning Architectures

Junjie Zhang[1⋆] and Gu Mingoy[2]

[1,2]College of Computer Science, Beijing Normal University, Beijing China.; Email: junjiezh@bnu.edu.cn ; gumingoy@bnu.edu.cn
⋆Corresponding author: Junjie Zhang (junjiezh@bnu.edu.cn)

Article History

Academic Editor:
Dr. Muhammad Sajid
Submitted: March 11, 2025
Revised: May 19, 2025
Accepted: September 1, 2025

Keywords:
Hierarchical Transformers, Legal information retrieval, Text summarization, BERT, Extractive summarization

Abstract

The computational requirements for swiftly and precisely comprehending lengthy legal documents pose a considerable barrier. The advancement of effective automated summarizing methods is essential for resolving these challenges. Extractive summarization, a widely utilized technique, emphasizes the selection of prominent sentences to abridge extensive publications. The intrinsic subjectivity of this work and the difficulties of obtaining contextual information inside extensive legal documents render it problematic, even for human specialists. To tackle these issues, we suggest an innovative method based on hierarchical transformers. Our concept is based on the stacked transformer encoder design of BERT. In light of the quadratic escalation in computing expense linked to the self-attention mechanism of transformer models as sequence length increases, which constrains their use with lengthy documents, we integrate the Longformer. The Longformer's attention mechanism, which scales linearly with sequence length, enables the analysis of texts with thousands of tokens or more. This mechanism substitutes conventional self-attention with a hybrid method that integrates task-specific global attention and local windowed attention. The suggested model was assessed using two recognized benchmark datasets for long-sequence transformers: BillSum and FIRE. Our experimental findings indicate that it surpasses leading approaches on both datasets. On the BillSum dataset, it attains Rouge-1, Rouge-2, and Rouge-L scores of 47.11, 33.02, and 42.19, respectively. In the FIRE dataset, the respective scores are 58.43, 44.31, and 41.54. These results highlight the enhanced efficacy of our suggested approach relative to current leading techniques.

1 Introduction

The extensive information environment of the modern World Wide Web is marked by exponential data expansion, with many websites currently producing more content every day. This surge of information frequently includes a significant volume of unnecessary, redundant, and unproductive content, which obscures essential facts. Consequently, users often encounter the need to navigate through numerous papers and websites to obtain the requisite information [1]. A document summary offers a feasible solution to address this problem. The incorporation of succinct summaries on individual webpages can markedly enhance user efficiency and elevate website traffic. Nonetheless, manually generating

summaries for the vast number of webpages available online is an impractical task [2]. Summarizing legal case judgments is an important challenge due to their extensive length and the increasing internet accessibility of legal information. Legal experts may struggle to navigate and understand case judgments that span tens or hundreds of pages filled with complex legal terminology [3]. Thus, the automated summary of legal documents has emerged as an essential and intricate endeavor.

Moreover, the absence of uniform structure in case rulings in certain jurisdictions, such as India, compared to legal papers in Europe or the United States, amplifies the need for and intricacy of automated summarization [4]. The everyday production of legal knowledge is considerable, including statutes, case law, regulations, and legal journals. Access to this information is essential for courts, attorneys, legislators, and citizens in maintaining justice [14]. The emergence of the internet has enabled access to legal information; yet, the vast quantity of data renders navigation and analysis a laborious and complex task [15]. Legal practitioners must invest substantial time and proficiency in understanding and condensing extensive and complex legal documents [6]. The majority of legal documents, marked by specialist terminology, intricate language, and hierarchical organization, considerably hinder effective information retrieval. The manual examination of these documents is an extended process prone to human mistakes, which may result in significant oversights with legal consequences [7]. Additionally, inexperienced users often endeavor to ascertain the precedents of analogous court cases [8]. Thus, automated text summarizing provides advantages to several parties. It can aid human summarizers by pinpointing essential elements for incorporation in case summaries [9]. Attorneys frequently must devise multiple plausible defenses to bolster their arguments and respond to case-related questions, currently depending on human-generated summaries, which can create unnecessary reliance and delays, especially in urgent or temporary situations [10].

Primarily depending exclusively on human-generated summaries, experts in the field can autonomously seek pertinent legal papers and assess their suitability for an automated summary. This enables them to focus more on fundamental legal matters instead of the laborious task of finding corroborative evidence. Automated text summarization is very advantageous for non-experts and inexperienced users [11]. The growing accessibility of legal papers in the public domain enables straightforward access. The enormous volume and complex legal terminology inherent in court documents can hinder understanding of the provisions. An automated method for generating case summaries markedly improves a user's capacity to access legal documents [12]. Recent improvements in natural language processing have facilitated the automation of more complex language comprehension tasks [13]. The automated summarizing of legal papers is a vital application of this technology, designed to extract the most relevant information while maintaining the nuances and legal context [14].

The present research focuses on extractive summarization because of its computational efficiency relative to abstractive approaches and its capacity to produce semantically and grammatically coherent sentences straight from legal texts [15]. Both generative and extractive summarization methods encounter difficulties in handling lengthy texts due to the computational intricacies linked to the encoder network. The extractive summarization method produces a summary by identifying the most salient sentences from the original document [16]. This procedure entails first encoding the input document and subsequently calculating the sentence scores included within it. Sentences with high scores are selected to create the summary, and these sentences are generally organized according to their respective scores [17]. This study presents a neural summarization model engineered to effectively generate extractive summaries by analyzing various input texts [18]. Our methodology enhances the Transformer architecture [14] with the integration of multi-document hierarchical encoding. Rather than merely concatenating text spans and presenting them as a singular flat sequence to the model, we employ an attention mechanism [19] to model inter-document interactions, therefore enhancing information sharing among many documents. The model assimilates current knowledge from previous studies by autonomously acquiring more complex structural links among textual elements.

1.1 Motivation

Legal practitioners, encompassing attorneys, paralegals, and legal researchers, manage a significant quantity of papers, including contracts, judicial decisions, statutes, and legal opinions. Due to the

complicated hierarchical structures, specialized language, and detailed contexts present in these legal texts, there is an urgent necessity for creative techniques to enhance understanding and accelerate information retrieval. BERT-based hierarchical transformers present a possible solution to address this disparity. These models efficiently capture the intricate semantics and structural intricacies of legal texts by utilizing contextualized embeddings and hierarchical attention mechanisms. The further study inquiries prompted the authors to investigate automatic legal text summaries.

- What is the impact of BERT-based hierarchical transformers on the efficacy and precision of legal document summarizing relative to traditional methods?
- To what degree do BERT-based hierarchical transformers proficiently handle the intrinsic hierarchical structure of legal documents in the summarizing task?
- How does the hierarchical attention mechanism enhance the model's capacity to produce summaries that correspond to the original document's structure?
- What is the scalability of the BERT-based hierarchical transformer method for summarizing various forms of legal documents?
- What computational resources are required to get an efficient, potentially real-time, summary of legal documents with this methodology?

1.2 Contribution

The research being conducted into legal writing summarizing with BERT-based hierarchical transformers is driven by the necessity to improve efficiency, reduce risks, and promote the integration of novel technologies in the legal field. The objective is to enhance the provision of legal services and provide more efficient judicial study and evaluation. The main contributions of this work are mentioned below.

- This study enabled the development of an automated summarizing system that swiftly extracts essential legal information, thereby markedly improving the efficiency of legal case assessment.
- This research generated automated summaries of legal papers with low turnaround time, intending to enable legal practitioners to focus on higher-value tasks such as developing legal strategy and interacting with clients.
- This study emphasizes the advantages of sophisticated artificial intelligence techniques in specialized fields, thereby offering prospects for multidisciplinary cooperation.
- This study aligns with and exhibits competitive performance relative to cutting-edge natural language processing approaches, specifically designed to meet the distinct needs of the legal sector.

1.3 Organization

This article delineates the hierarchical organization of the research. Section 2 provides a comprehensive assessment of pertinent literature, examining established approaches and previous research at the convergence of legal issues and artificial intelligence. Section 4 delineates the suggested model, including critical phases such as feature identification, data dimensionality reduction, preprocessing methods, and the foundational transformer-based architecture. Section 5 delineates the study's findings and offers an exhaustive analysis of the data. This section outlines the evaluation metrics used to analyze the proposed model's performance relative to competing models. Ultimately, Section 6 delineates the constraints of the proposed paradigm and proposes prospective directions for future inquiry.

2 Background

The increasing number of legal paperwork has markedly heightened attention on the developing field of legal document summarization [20]. Legal rulings are frequently protracted and complex, exacerbated by the prevalent use of specialized abbreviations. Extensive academic endeavors have focused on aggregating legal writings from many nations. The structure of legal documents varies throughout nations. A considerable amount of research has investigated the use of deep learning models, typically utilizing supervised or semi-supervised training, to summarize legal texts. Rule-based algorithms constitute a traditional method for legal text summarizing, depending on established rules to extract essential information [21]. Nonetheless, despite its ubiquity, this technique encounters difficulties in adeptly processing intricate or ambiguous language. The extensive textual content of legal documents might provide difficulties for automated machine learning algorithms regarding processing and categorization [22]. The challenges are exacerbated in the legal field due to complications like convoluted grammatical structures, layered phrases, specialist terminology, and the prevalent usage of abbreviations [23]. Extractive summarization is a technique that entails picking the most essential phrases or segments from a document to produce a succinct summary [24].

Graph-based methods rank among the most used extractive summarization techniques, typically being domain- and language-agnostic, hence eliminating the need for training data. LexRank and TextRank are notable instances of graph-based extraction techniques [25]. Although these algorithms generally generate summaries that are factually aligned with the original document [26], they frequently incorporate extraneous information [27]. Managing legal papers can be arduous due to their potential length and the possibility of including several interrelated documents. Studies demonstrate that the typical legal document comprises more than 4,500 tokens [28]. Currently, the maximum length for processing documents is generally restricted to approximately 3,000 tokens [29]. The ITACaseHold dataset comprises documents averaging 800 tokens in the holdings portion and 4,700 tokens in the intermediate permits section. The main aim of legal research is to improve efficiency by utilizing methods that swiftly distill lengthy documents to their fundamental elements. Consequently, various domain-specific legal summarizing algorithms have been created. The LetSum system, developed by [30], assigns rhetorical roles to phrases and subsequently employs TF-IDF to score them. A set percentage of phrases for each rhetorical role is chosen for the final summary according to their allocated rank [31].

A study [32] introduced a methodology that amalgamated TF-IDF with elements relevant to the legal area, including the enumeration of legal entities in each phrase. Conversely, [33] utilized machine learning techniques to evaluate the probability of individual sentences being incorporated into a summary. [34] employed an iterative selection methodology, integrating a Convolutional Neural Network (CNN) with a Random Forest (RF) classifier, to distinguish between phrases that contained reasoning and those that did not. Ultimately, [35] underscored the significance of adeptly amalgamating domain-specific knowledge by choosing summary sentences and integrating pertinent contextual words and sentences from legal domain information into an objective function, which could subsequently be optimized through Integer Linear Programming (ILP). Following the advent of self-attention-based transformers [36], models like BERT [37] and RoBERTa [38] have exhibited enhanced performance relative to preceding models. These methodologies frequently utilize pre-trained large models (PLMs), including BioBERT [10], SciBERT [39], and FinBERT [40], which are trained on domain-specific corpora. In the legal field, [41] presented LegalBERT, a BERT-based model pre-trained exclusively on legislative texts from the European Union and the United States, together with court records.

Another study [42] also created CaseLaw BERT, a pre-trained language model (PLM) derived from BERT, specially trained on legal documents from the United States. Additionally, [21] presented Pile of Law, a substantial BERT model trained on a comprehensive dataset of legal documents from the United States and the European Union. In the Italian legal domain, [33] introduced Italian LEGAL BERT, which entails training the XXL Italian BERT-based model and its pre-trained variation on Italian legal texts, employing the CamemBERT architecture and exploiting extensive Italian civil law corpora. Likewise, [14] utilized the Italian civil code to train a BERT model. To improve summarization via rhetorical role labeling, [17, 18] introduced a hierarchical multitask learning framework employ-

ing Sentence-BERT (SBERT), Bidirectional Gated Recurrent Units (BiGRU), and Maximal Marginal Relevance (MMR) for the selection, arrangement, and integration of sentences. This project aims to create an accurate and domain-specific legal document summarizing system utilizing BERT-based hierarchical transformers, which have demonstrated favorable results in numerous natural language processing applications. This study seeks to address the difficulties inherent in processing extensive documents and precisely summarizing legal texts.

3 Preprocessing

Preprocessing, as described by [43], involves a methodical sequence of actions designed to transform unstructured text into a more comprehensible and analyzable format. Diverse methodologies can be utilized for text preprocessing, including stemming, lemmatization, and stop word elimination. Nonetheless, due to the requirement for the model to produce coherent text and the risk that certain preprocessing techniques may undermine this coherence, only a restricted range of preprocessing methods is appropriate for text summarizing tasks. This study employed the Natural Language Toolkit (NLTK) program for text preparation. The legal documents were divided into paragraphs, phrases, subsections, and sections to enable the recognition of the document's hierarchical structure. Subsequently, text cleaning was conducted to eliminate superfluous letters, symbols, and formatting elements, including ellipses, en dashes, and em dashes. The text was transformed to lowercase for uniformity, as illustrated in Figure 1. Subsequently, each textual unit at various hierarchical levels was allocated distinct identifiers. This procedure guaranteed that each section, subsection, paragraph, and phrase was assigned a unique hierarchical identifier. These distinct identifiers allow the model to recognize hierarchical relationships among units throughout both the training and inference stages [44]. The subsequent phase entailed tokenization, wherein the text was disaggregated into discrete words and phrases, laying the foundation for further analysis. This study utilized the BERT tokenizer for this purpose. Consequently, particular legal terminology that minimally impacted the document's broader meaning, such as "plaintiff," "defendant," and "jurisdiction," was eliminated. The preparation phases are depicted in Figure 1. Algorithm 1 was employed to preprocess the data for extractive text summarization. This study utilized lemmatization to convert words to their base forms, thereby enhancing comprehension. The project aimed to improve the accuracy of legal document summaries by identifying and categorizing named entities, such as names, legal terminology, locations, and organizations. Thereafter, positional encoding was implemented on the tokenized input to convey information about each token's position in the sequence. This facilitated the model's comprehension of the sequential arrangement of tokens, essential for digesting hierarchical structures. The following preparation phase entailed the annotation of legal papers with summary labels. In this study, each tokenized statement was designated a label denoting its relevance or irrelevance to the summary. The labeled data was subsequently utilized to train the summarization component of the hierarchical transformer [45]. The datasets included in this investigation were divided into training, validation, and test subsets. Subsequently, both the legal text and their associated labels were converted into numerical representations for input as vectors into the model. Following preprocessing, a vocabulary was created, assigning a unique numerical identification to each distinct word in sequence. The vocabulary was subsequently employed to encode each sentence, and the resultant encoded list was fed into the model. The primary objective of this extensive workflow is to enhance the efficiency of collecting relevant information from legal papers and to prepare this information for subsequent tasks such as summarization and search analysis.

4 METHODOLOGY

Figure 2 depicts the framework of our proposed summarization method. To effectively manage a substantial volume of input paragraphs, we initially implement a paragraph ranking system and then utilize extractive summarization on the top-L ranked paragraphs instead of processing all paragraphs directly. Our summarizer employs the efficient encoder-decoder architecture described in [46], wherein

Algorithm 1 Extractive Approach Algorithm

procedure LEGAL TEXT PREPROCESSING

Input: $Dataset, f\_k, N$

Output: Preprocessed text

for each: $IN\_text\ in\ dataset$ **do**

$Significant\_Sent \leftarrow []$

$Token\_Sent \leftarrow nltk.s\_tokenize(IN_{text})$

$Token\_Sent \leftarrow remove\_irre_{c}haracters(IN\_text)$

$Token\_Sent \leftarrow remove\_stopwords(IN\_text)$

$Token\_Sent \leftarrow lemmatize\_tobase(IN\_text)$

$Significant\_Sent \leftarrow First\_KSent(f\_k)$

for each: $Sentenc\ in\ Token\_Sent$ **do**

$Score \leftarrow ComputScorForSentenc(trigram)$

$end\ for$

$Significant\_Sent \leftarrow SelectSignificantSent(N, Score)$

$end\ for$

end procedure

the encoder converts the input text into hidden representations, and the decoder uses these representations to produce the target summaries. This research especially examines the encoder component of the model. Our decoder implementation adheres to the Transformer architecture described in [47], successively generating the summary from the encoded source input, token by token. To enhance the production of extended and more cohesive summaries throughout the decoding process, we additionally implement beam search and a length penalty. The incorporation of the transformer architecture with hierarchical attention mechanisms is crucial for developing a hierarchical transformer designed for legal document summaries. Within this paradigm, the encoder transforms an input sequence of symbol representations $x_1, ..., x_n$ into a sequence of continuous representations $z_1, ..., z_n$. Thereafter, conditioned on z, the decoder produces an output sequence of symbols $S_1, ..., S_m$ sequentially, one element at a time. During each iteration of the decoding process, the model employs the previously generated symbols as input for the current step. This section presents legal text as a sequence labeling task. The design of our extractive, hierarchical transformer-based model is illustrated in Figure 2, consisting of three main components: the sentence encoder, the document encoder, and the document decoder.

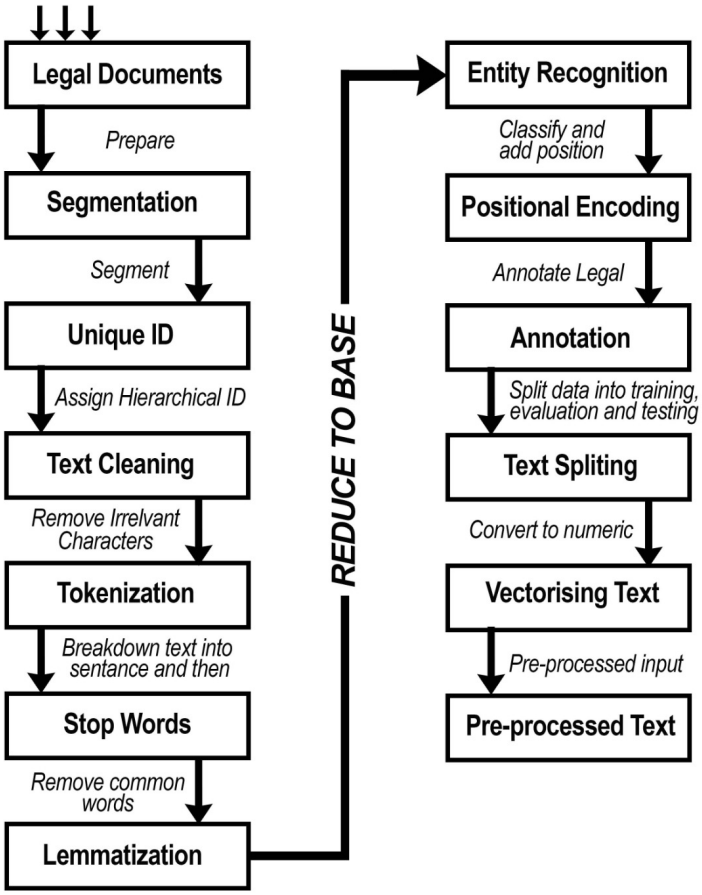

Figure 1: Linguistic Preprocessing of Legal Corpora.

4.1 Encoding Sentences

Some previous designs faced constraints in the direct processing of lengthy documents. The original Transformer model employs a global attention mechanism characterized by a quadratic time and memory complexity of $O(n^2)$, with n being the length of the input sequence. To ensure that computational complexity scales linearly with input length, we utilized the Longformer [48] as the sentence encoder. The Longformer, a transformer-based model engineered for processing lengthy text sequences, employs a sliding window attention mechanism, therefore lowering the computational complexity from $O(n^2)$ to $O(n)$. The Longformer architecture employs a hybrid attention mechanism that integrates global and local attention. The local attention mechanism is executed via a sliding window method, wherein each token interacts solely with its ω adjacent tokens. This yields a linear complexity for attention computation relative to the input sequence length n, represented as $O(\omega \times n)$, where ω signifies the size of the attention window. Conversely, global attention encompasses a comprehensive attention pattern, wherein tokens with global attention focus on every token inside the input sequence. The present study employs an encoder pre-trained on the Longformer language model [49] to convert each sentence into a representation vector. To obtain the feature representation vector for each sentence, we affix the special categorization token, [CLS], as the initial token. The feature vector of the [CLS] token serves as a prominent representation of the entire sentence. Consistent with Longformer's attention mechanism, global attention is utilized for the [CLS] tokens of the input sequence to capture extensive semantic information, whereas local windowed attention is employed for the other sentence tokens to reduce computational complexity, thus aligning Longformer's functionalities with the goals of

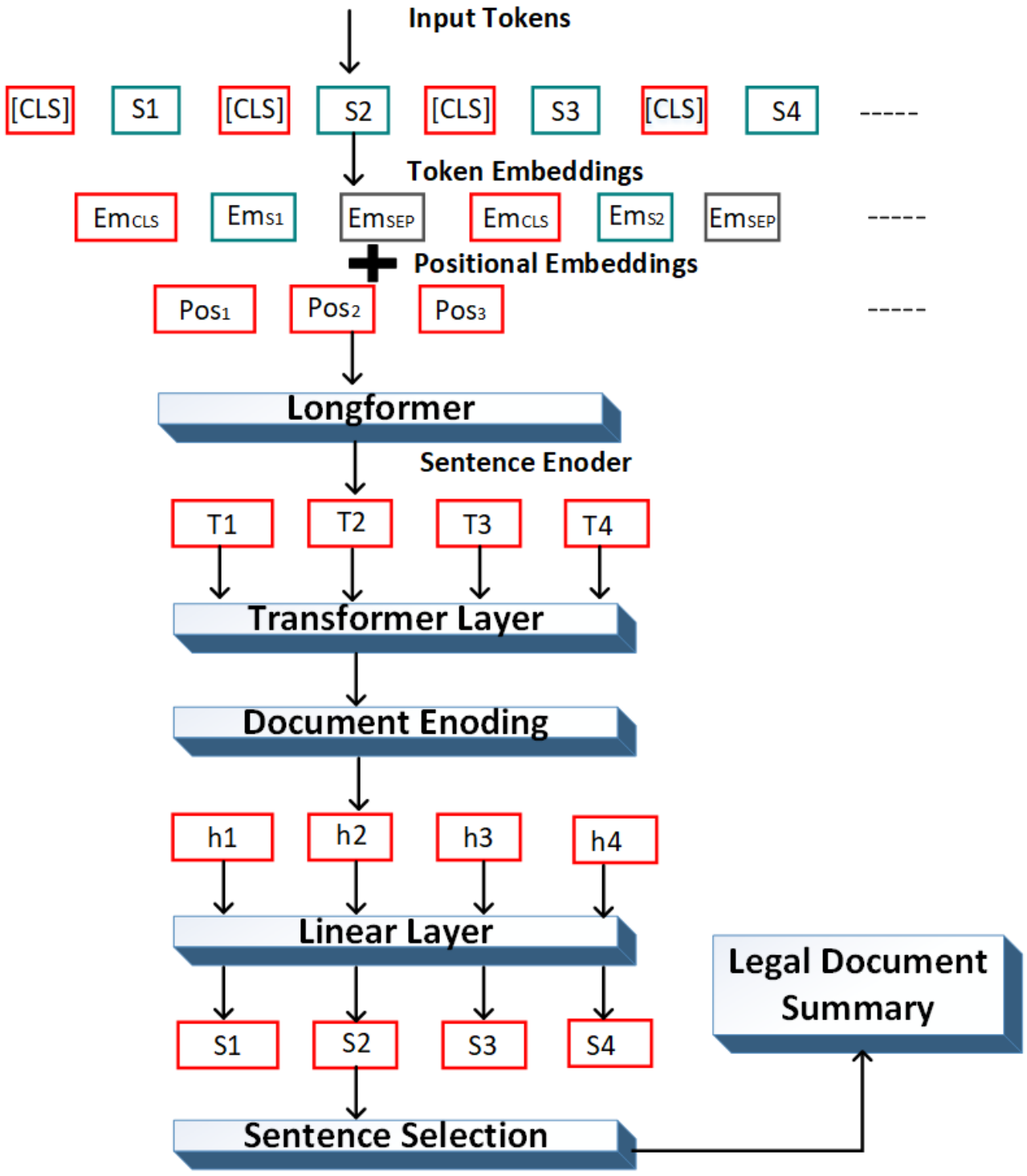

Figure 2: Longformer-Based Model for Legal Document Summarization.

extractive summarization. Figure 2 demonstrates that Equation 1 produces the representation vector for each word in the manuscript.

$$\omega_j^i = em_j^i + p_j \tag{1}$$

where ω_j^i denotes the j_{th} word of i_{th} the sentence, em_j^i is the embedding, p_j is the position embeddings [20], and ω =[ω_1^1,ω_2^1 .....,ω_2^2 ....., ..., ω_2^m,ω_2^m]. Next, Longformer is used to encode all the words given in Equation 2:

$$T = Longformer(\omega) \tag{2}$$

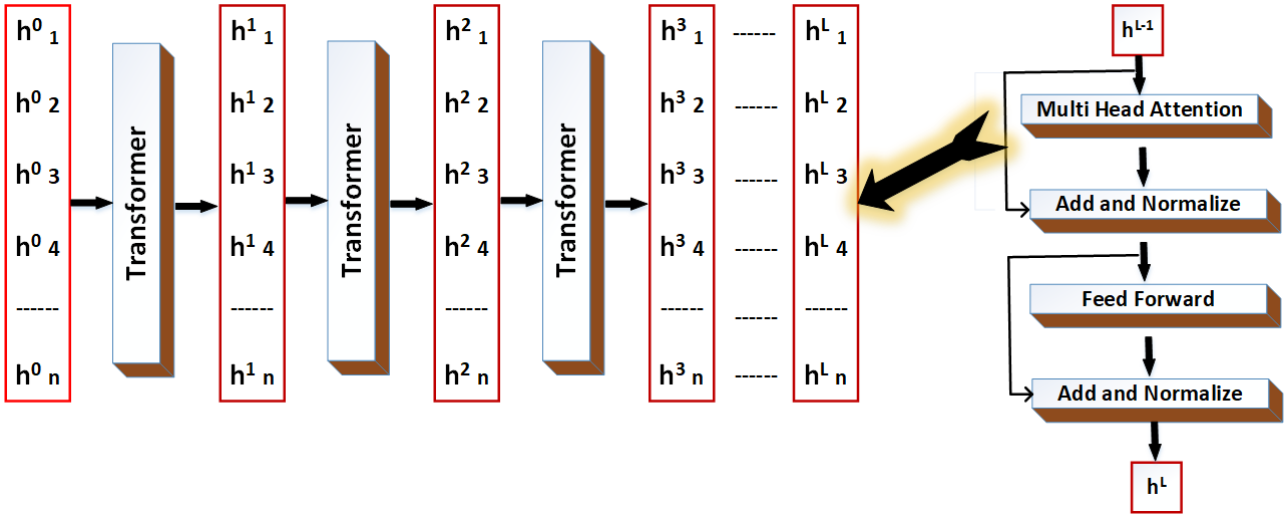

Figure 3: Detail of a Single Transformer Encoder Block within a Stacked Architecture.

where T_i, which can be seen as the representation vector of the i_{th} sentence, is the output of the corresponding position of the i_{th} [CLS] [50]. There are $T = [T_1, T_2, ...T_m]$ sentences in the document.

4.2 Encoding Document

To enhance higher-dimensional interaction among sentence representation vectors, this research deployed a transformer-based document encoder, as illustrated in Figure 3. Initially, each sentence vector T_i is amalgamated with its respective positional embedding P_i to produce the preliminary input for the document encoder, as articulated in Equation 3:

$$h_i^0 = T_i + P_i \tag{3}$$

The initial input to the document encoder is denoted as $h_0 = h_1^0, h_2^0, ..., h_m^0$. The ensuing operations of the document encoder, comprising an L-layer transformer, are delineated by Equations 4 and 5:

$$\hat{h}^l = LN(h^{l-1} + MHAtt(h^{l-1})) \tag{4}$$

$$h^l = LN(\hat{h}^l + FFN(\hat{h}^l)) \tag{5}$$

where $LN()$ indicates Layer Normalization, $MHAtt()$ means Multi-Head Attention, and FFN represents the Feed Forward Network. The contextualized sentence vectors are obtained by iteratively executing the internal transformer computations within a single layer of the document encoder, as delineated by Equations 4 and 5, for L iterations. The output of the transformer's last layer, $h^L = h_1^L, h_2^L, ..., h_m^L$, is subsequently transmitted to the decoder.

4.3 Decoding Document

In extractive summarization, a binary classifier serves as the classification head, systematically predicting the inclusion of each sentence in the document for the summary. The model predicts a binary label for each sentence, denoting its relevance. If the predicted label indicates significance, the sentence is considered essential and is selected for inclusion in the summary. This study utilizes a binary classifier to classify the final output of the document encoder, as delineated in Equation 6:

$$\hat{y}_i = \sigma(h_i^L \omega_o + b_o) \tag{6}$$

where b_o signifies the bias term, ω_o indicates the trainable weight matrix, and h_i^L is the vector representing the sentence for S_i. The loss function is thereafter calculated as specified in Equation 7:

$$Loss = BCE(\hat{y}_i, y_i) \tag{7}$$

In this context, y_i denotes the actual label, while $\hat{y}_i$ signifies the expected label. In this case, we employ the binary cross-entropy (BCE) loss function to train the classifier.

Algorithm 2 Extractive Approach Algorithm

procedure HIERARCHICAL TRANSFORMER

Input: $Epoch\_Length, batch\_size$

Output: Calculated Loss

for each: *batch in batch_size* **do**

for each: *epoch in epoch_length* **do**

$encoding \leftarrow encoder(bat.source, bat.source\_frequ)$

$Sent_t, x_t, mh\_atten\_decoder \leftarrow decoder(encoding, bat.tr\_inp, bat.tr\_mask)$

$h_t \leftarrow ContexVector(encoding, mh\_atten\_decoder)$

$Generate\_Summ\_Probab \leftarrow P\_Generator(Sent_t, x_t, h_t, encoding)$

$Loss \leftarrow CalculateLoss(Generate\_Summ\_Probab, bat.tr\_ext)$

$end\ for$

$end\ for$

end procedure

4.4 Evaluation Metrics

This research uses the ROUGE metric, a prevalent evaluation tool in the text summarizing field for assessing automatically generated summaries [51], to evaluate the effectiveness of our system. The resulting summaries are assessed using Rouge-N (with N = 1 and 2) and Rouge-L, as detailed below:

- Precision, a crucial evaluation indicator, denotes the ratio of the generated summary that is pertinent and accurate to the source content. This is articulated mathematically in Equation 8.

$$Precision_{Rouge-L} = \frac{LCS(X,Y)}{n} \tag{8}$$

- Recall, an essential metric for assessing the efficacy of a summarizing system, measures the percentage of vital information from the source that is included in the produced summary. The mathematical expression for recall is delineated in Equation 9.

$$Recall_{Rouge-L} = \frac{LCS(X,Y)}{m} \tag{9}$$

- The F1-score, a robust evaluation metric, is determined as the harmonic mean of precision and recall [52]. It offers a comprehensive assessment of the summarization's quality by evaluating both its relevance and completeness, as demonstrated in Equation 10.

$$F1score_{Rouge-L} = \frac{(1+\beta^2)R_L P_L}{R_L + \beta^2 P_L} \tag{10}$$

We do a human assessment of the computer-generated abstracts. As per [53], this study solicits evaluations from human experts on summaries based on these recognized metrics:

- Grammar (GR): GR evaluates the linguistic accuracy and fluency of the produced text.
- Synopsis Utility (SU): SU assesses the extent to which the produced summary is advantageous or valuable about the original content.
- Synopsis Consistency (SC): SC evaluates the logical and semantic consistency of a created summary.
- Non-Redundancy (NR): NR assesses the existence or lack of linguistic structures that are repetitive or redundant within the output text.
- Overall Quality (OQ): OQ offers a comprehensive assessment of the quality of the produced summary.

Expert assessors assess the quality of the summaries using a 5-point Mean Opinion Score (MOS) scale. Table 1 presents the inquiries employed to direct the human validation process.

Table 1: Questions to assess quality of generated text by Humans

Evaluation Indicators	Query
Grammar	To what extent does the produced text conform to the principles of grammar, syntax, and punctuation?
Synopsis Utility	To what extent does the summary encapsulate the most critical facts, principal arguments, and vital details contained in the primary legal document?
Synopsis Consistency	Does the summary convey facts in a coherent and logical order?
Non-Redundancy	Are there occurrences where identical material or very related language is reiterated within a single sentence or across successive words without contributing much value?
Overall Quality	To what extent does the summary accurately reflect the information contained in the primary legal document? Is the data provided in the synopsis specifically applicable to the primary points of the source text?

5 Results and Discussion

This section presents an extensive empirical examination of the proposed method. We provide comparisons with benchmark and state-of-the-art methodologies. Two legal benchmark datasets, BillSum (US and CA) and Fire, were utilized in this research's experiments.

5.1 Datasets

The inaugural release of the BillSum dataset, containing 22,218 US Congressional legislation, was presented by [54]. This dataset comprises 18,949 training samples and 3,269 testing samples of United States Congressional legislation. Table 2 indicates that both the training and testing documents for US Congressional bills contain an average of 62 sentences. The summaries for both the training and testing sets comprise an average of six sentences. The training set of the FIRE dataset comprises 500 document-summary pairings [55], derived from decisions of the Indian Supreme Court. The dataset curators further supply pre-processed, sentence-tokenized iterations of the documents alongside their descriptions. Alongside the summaries, each text is annotated with labels denoting sentence relevance (relevant or non-relevant) and rhetorical strategies. Although 50 documents were supplied for testing,

our tests were exclusively performed on the training dataset because ground truth labels for the test set were unavailable.

Table 2: BillSum & Fire Datasets Description.

Dataset	Training	Testing	Total
BillSum (US)	18,949	3,269	22,218
BillSum (CA)	-	-	1,237
FIRE	450	50	500

Table 3: Experimental Details.

Specification	Values
Processor	Intel Xeon W1370
Graphics Card	RTX3070
Hard Disk	SSD
Operating System	Windows 11
Number of Hidden Units	1-200
Number of Epochs	20-50
Platform	Jupiter Notebook
Languages	Python
Multi-Node System	RAM 16GB

5.2 Experimental Setup

All tests used to implement extractive summarization were conducted using Jupyter Notebook, as outlined in Table 3. The hardware platform consisted of a computer including an Intel Xeon W1370 processor, operating on Windows 11, with 16 GB of RAM, an SSD, and an RTX3070 GPU. The "longformer-base" pre-trained weights were utilized to initialize the Longformer model for phrase encoding. The classifier and the document encoder transformer were initialized with arbitrary weights. The vector dimensions for both words and sentences were established at 768, with a batch size of 64 utilized. In the first 1,000 training steps, just the weights of the document encoder and the classifier were modified, while the weights of the sentence encoder remained static. After this preliminary phase, comprehensive modifications were implemented to the model parameters. The model's efficacy was assessed on the validation set at intervals of 500 steps. Thereafter, the ideal model parameters were preserved, and the model's efficacy on the test set was documented as the conclusive outcome. The learning rate was established at 3.0×10^{-4}. The implementations were executed utilizing the Python programming language, accompanied by libraries including Scikit-learn, TensorFlow, and NumPy. The Rouge score Python tool was employed to assess the experimental outcomes in comparison to the reference summaries. Table 4 displays a compilation of trigram frequencies alongside their respective scores to facilitate the identification of word occurrences throughout the documents.

Table 4: Selection of Trigram Frequency

Trigram	# of Occurrences	Score
('Illegal', 'criminal', 'law')	4,337	1.8543
('Illegal', 'family', 'judge')	3,285	0.6543
('Illegal', 'protect', 'civil')	6,183	1.5032

5.3 Extractive Summarizing Techniques

Tables 5, 6, and 7 present a comparison of the BillSum and FIRE datasets utilizing six prevalent baseline techniques.

1. TextRank is a graph-based summarization technique that constructs a graph from the input document. In this graph, each sentence represents a node, and the edges between nodes are weighted based on the degree of similarity between the corresponding sentences.

2. Latent Semantic Analysis (LSA) represents an additional methodology. This technique utilizes Singular Value Decomposition (SVD) to discern the most significant semantic elements inside a document.

3. Restricted Boltzmann Machines (RBMs) exemplify an unsupervised deep learning methodology. This method initially identifies pertinent elements from the text, which are further refined to produce a summary.

4. CaseSummarizer is a legal-specific baseline technique that produces an extracted summary utilizing word frequency analysis enhanced with domain-specific data.

5. KLSum: This method's primary concept is to continuously incorporate sentences into the summary in a greedy fashion, provided that a reduction in KL divergence between the summary and document sets is evident.

6. Sumbasic is an egoistic estimation method that scores sentences based on the mean likelihood of the words within them, with a re-weighting component to reduce redundancy.

We initially performed studies utilizing the FIRE and BillSum databases. Tables 6, 6, and 7 delineate the efficacy of non-deep learning baseline methodologies, including TextRank, LSA [25], RBM, and CaseSummarizer. A notable performance disparity is apparent when contrasting these conventional methods with deep learning-based techniques. TextRank, a frequently employed and typically efficient baseline, extracts the initial three sentences of a document as the summary. Our proposed methodology exhibits enhanced efficacy relative to the TextRank baseline model. Additionally, our model surpasses the existing state-of-the-art models on both the BillSum and FIRE datasets, demonstrating significant accuracy enhancements of 0.58 and 0.4 in ROUGE-1 and ROUGE-L scores, respectively. To examine our initial research question concerning the influence of BERT-based hierarchical transformers on the accuracy and efficacy of legal document summarization relative to leading methodologies, we provide a comprehensive comparison in Tables 10, 11, and 12, highlighting precision, recall, and F1-score metrics, respectively.

5.4 Advanced methodologies for comparative analysis

- The Pegasus model is fine-tuned using the US securities lawsuit dataset to develop the Legal Pegasus model.

- BO-Textrank: The authors suggest a Bayesian Optimization (BO) strategy to enhance the Textrank algorithm for extractive summarization. The optimized TextRank is subsequently employed for the summarizing task.

- SummaRuNNer: The SummaRuNNer model utilizes Recurrent Neural Networks (RNNs) to frame the extractive summarization challenge as a binary sequence labeling task.

Table 5 provides a comparative examination of different baseline text summarizing algorithms on the BillSum dataset for US legal documents, assessed using ROUGE-1, ROUGE-2, and ROUGE-L criteria. The scores reflect the degree of similarity between the generated summaries and the human-authored reference summaries, with elevated values denoting superior performance. Significantly, CaseSummarizer and the "Proposed" model attain the highest ROUGE-1 and ROUGE-L scores, indicating their

Table 5: ROUGH based comparative results on the BillSum dataset (US).

Base Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank [32]	0.40	0.31	0.37
Latent Semantic Analysis (LSA) [44]	0.30	0.27	0.18
Restricted Boltzmann Machines (RBMs) [53]	0.29	0.17	0.36
CaseSummarizer [21]	0.49	0.29	0.40
KLSum [42]	0.41	0.32	0.42
Sumbasic [18]	0.44	0.37	0.42
Proposed	0.48	0.39	0.43

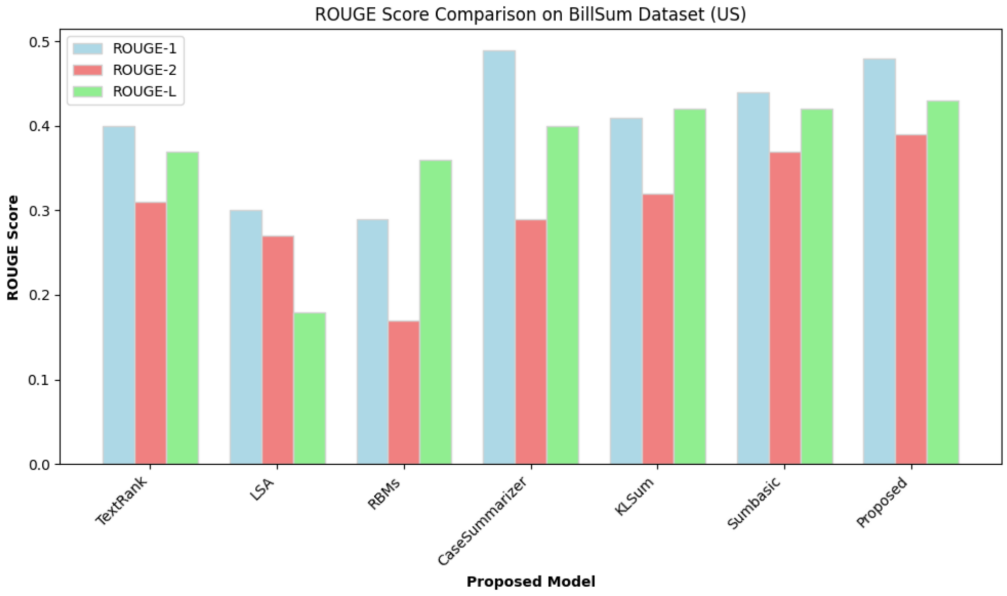

Figure 4: Human Validation using different number of shots on BillSum (US) dataset

superior ability to capture unigram and overall content similarity. Sumbasic exhibits the most robust performance in ROUGE-2, signifying a superior capacity to align bigrams. Conversely, Latent Semantic Analysis (LSA) typically demonstrates the lowest results among all three ROUGE indicators in this comparison. Table 6 contrasts the ROUGE-1, ROUGE-2, and ROUGE-L scores of several baseline

Table 6: ROUGH based comparative results on BillSum dataset (CA)

Base Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank [1]	0.40	0.31	0.37
Latent Semantic Analysis (LSA) [10]	0.30	0.27	0.18
Restricted Boltzmann Machines (RBMs) [13]	0.29	0.17	0.36
CaseSummarizer [22]	0.45	0.29	0.42
KLSum [43]	0.44	0.33	0.41
Sumbasic [55]	0.31	0.39	0.31
Proposed	0.47	0.36	0.45

summarization methods utilizing the Californian BillSum dataset. The scores indicate the degree of correspondence between the generated summaries and the reference summaries, with elevated values

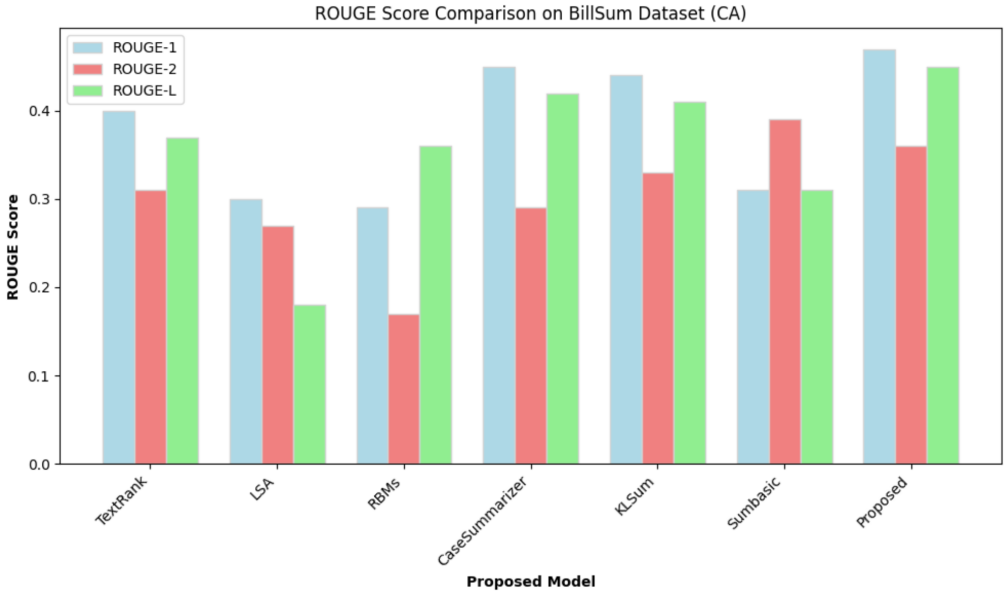

Figure 5: Human Validation using different number of shots on BillSum (CA) dataset

signifying superior performance. The "Proposed" model attains the highest ROUGE-1 and ROUGE-L scores, indicating superior performance in unigram and overall content alignment. Sumbasic [12] distinguishes itself with the highest ROUGE-2 score, signifying its efficacy in capturing bigram similarity. In contrast, Latent Semantic Analysis (LSA) [16] typically demonstrates the lowest scores across all three ROUGE measures in this assessment of Californian legal language. 1

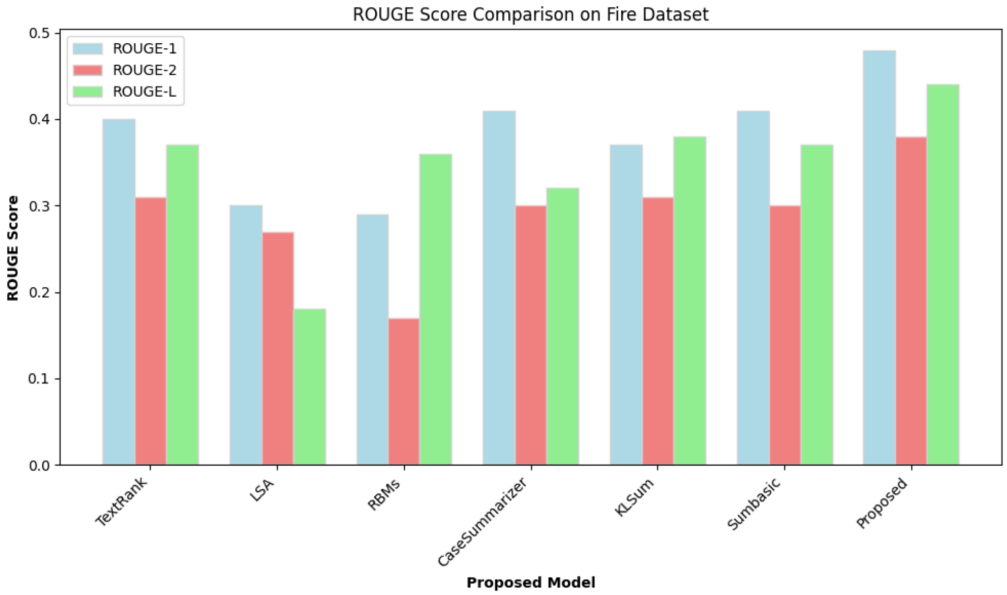

Figure 6: Human Validation using different number of shots on Fire dataset

5.5 Human evaluation of the generated summaries

We assessed a sample of 200 computer-generated summaries with the assistance of two topic experts. Tables 8 and 9 present the qualitative outcomes attained in summary generation. As the quantity of training samples utilized for LLM fine-tuning increases, there is a steady enhancement in all performance indicators, signifying an elevation in the level of accuracy of the produced material. Generated

Table 7: ROUGH based comparative results on Fire dataset

Base Model	ROUGE-1	ROUGE-2	ROUGE-L
TextRank [11]	0.40	0.31	0.37
Latent Semantic Analysis (LSA) [10]	0.30	0.27	0.18
Restricted Boltzmann Machines (RBMs) [1]	0.29	0.17	0.36
CaseSummarizer [9]	0.41	0.30	0.32
KLSum [33]	0.37	0.31	0.38
Sumbasic [25]	0.41	0.30	0.37
Proposed	0.48	0.38	0.44

summaries exhibit a commendable degree of grammatical quality and high educational value ratings throughout all of the training instances, while the non-redundancy ratings demonstrate negligible differences. In all training circumstances, the summaries consistently attain superior overall quality scores relative to abstracts. This disparity is partially attributable to the varying lengths of summaries.

Table 8: Human validation using evaluation indicators on BillSum Dataset

Number of Shots (N)	GR	SC	SI	NR	OQ
100	4.89	1.96	1.71	3.79	1.38
200	4.91	2.01	2.94	3.97	2.44
500	4.93	2.46	3.17	4.48	2.93
700	4.94	2.56	3.71	4.68	3.23
1000	4.96	2.86	3.87	4.98	3.53

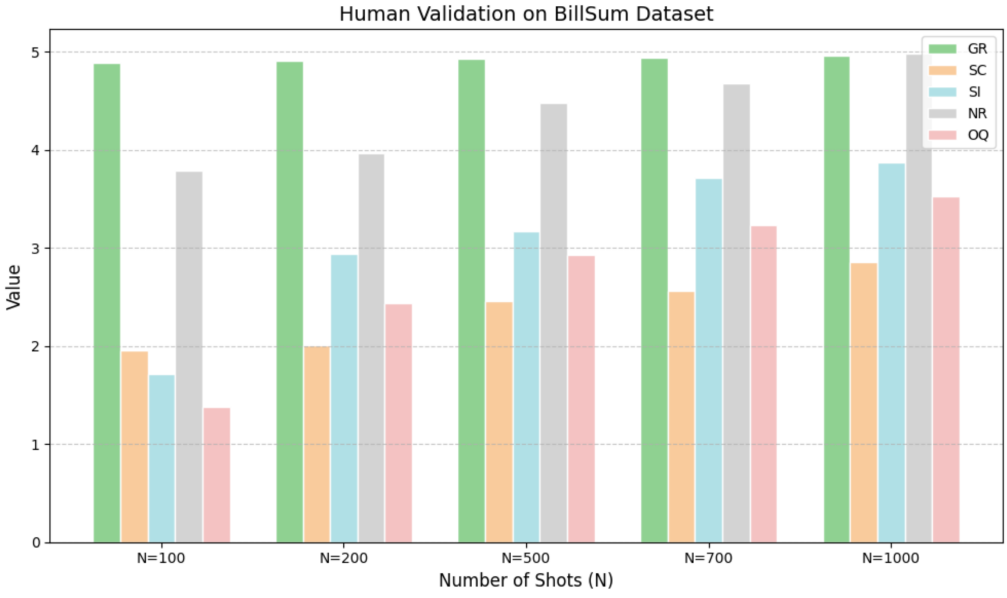

Figure 7: Human Validation using different number of shots on BillSum dataset

Table 8, entitled "Human Validation Utilizing Evaluation Indicators on the BillSum Dataset," presents the outcomes of a human evaluation procedure. It evaluates multiple assessment metrics (GR, SC, SI, NR, and OQ) across various "Number of Shots (N)" inside the BillSum Dataset. The "Number of Shots (N)" presumably denotes the quantity of training data or the number of instances

utilized in the assessment. The values associated with each assessment indication denote human ratings or scores, reflecting the degree of consensus or quality recognized by the evaluators. The number of shots (N) denotes the various levels or quantities of training data utilized in the assessment. The values are N=100, N=200, N=500, N=700, and N=1000, signifying a progressive increase in data volume. The columns (GR, SC, SI, NR, and OQ) denote various metrics or criteria employed by human evaluators to appraise quality or performance. The table does not specify the precise meanings of these indicators; however, they likely pertain to factors such as grammatical accuracy, coherence, and the importance of information. The numerical values in the table denote the scores or ratings

Table 9: Human validation using evaluation indicators on Fire Dataset

Number of Shots (N)	GR	SC	SI	NR	OQ
100	2.99	1.46	1.87	2.19	2.38
200	3.41	2.15	1.98	2.47	2.54
500	3.59	2.68	2.27	3.08	2.83
700	3.84	3.06	2.55	3.38	3.21
1000	4.26	3.36	3.01	3.78	3.43

assigned by human evaluators for each evaluation indication at each "Number of Shots (N)" level. Elevated numbers typically signify superior concordance or quality. As the "Number of Shots (N)" escalates, the values for the majority of evaluation metrics similarly tend to rise. This indicates that an increase in training data enhances the quality or performance of the assessed system. The GR indicator consistently exhibits the highest values across all levels of "Number of Shots (N)," signifying that it is most favorably assessed by human assessors. The OQ indicator typically exhibits the lowest values, indicating it is assessed the least favorably. The table illustrates the variation in human assessments of several quality indicators as the volume of training data escalates on the BillSum Dataset.

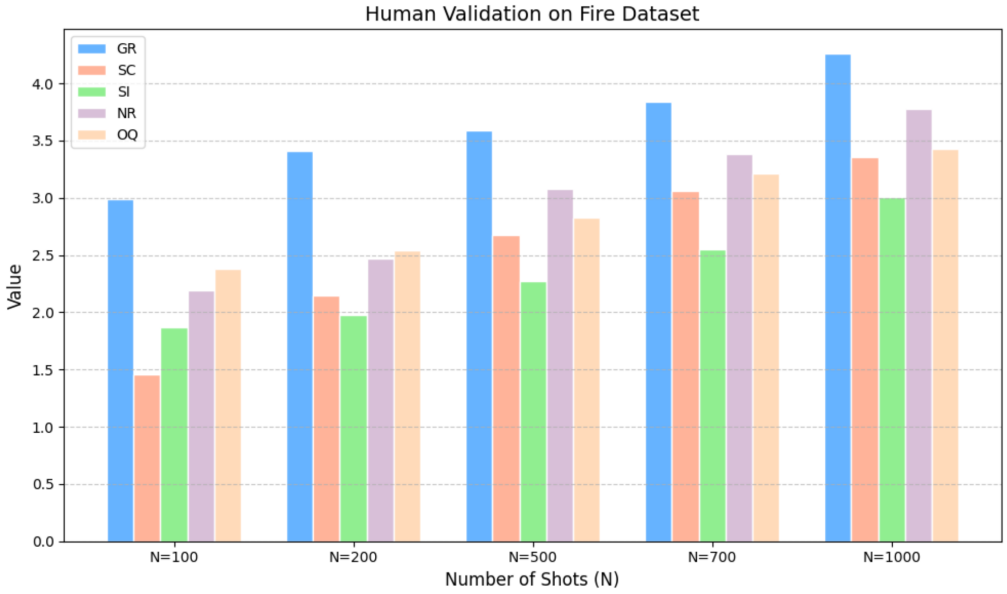

Figure 8: Human Validation using different number of shots on Fire dataset

Table 9 displays the outcomes of human validation employing several evaluation metrics on the Fire Dataset. The table analyzes the variation of these variables with an increasing "Number of Shots (N)," presumably denoting the volume of training data utilized. The assessment metrics are GR, SC, SI, NR, and OQ, with the numerical values in the table denoting human evaluations or scores for each metric. As the "Number of Shots (N)" escalates from N=100 to N=1000, a discernible pattern of enhancement is observed across the majority of evaluation metrics, indicating that an increase in

training data correlates with superior outcomes. Significantly, GR constantly attains the greatest evaluations, while SC typically exhibits the lowest. This table elucidates how human evaluators assess the quality of results as the volume of training data varies on the Fire Dataset.

5.6 Discussion

This article delineates the performance of models trained and assessed on the FIRE and BillSum datasets. The ROUGE-1, ROUGE-2, and ROUGE-L scores indicate that the TextRank model outperforms the LSA model. The ROUGE-2 value attained by TextRank is significantly superior. The two models' negligible disparity in the ROUGE-1 and ROUGE-L scores indicates similar performance in these areas. In contrast, a notable deterioration in performance occurs when the same model is enhanced with an extractive method layer, suggesting that frequency-based information exerts a restricted influence on the summarization result. Analysis of models trained with a sequence length of 300 tokens compared to those with 450 tokens indicates that the 450-token model significantly surpasses those lacking the extractive technique. Implementing the extractive methodology resulted in a notable enhancement in the models' performance, as demonstrated by the elevated ROUGE scores. We have addressed our second research question by effectively demonstrating the capabilities of BERT-based hierarchical transformers to manage the summarizing process while supporting the intrinsic hierarchical structure of legal documents. Our experiments suggest that frequency information enhances the

Table 10: Comparison of precision score with state-of-the-art techniques

Sr.#	TextRank [1]	LSA [22]	RBMs [33]	Summarizer [44]	KLSum [42]	Sumbasic [43]	Proposed Model
1	0.101	0.163	0.118	0.629	0.529	0.419	0.589
2	0.207	0.202	0.313	0.539	0.429	0.421	0.609
3	0.113	0.121	0.242	0.513	0.549	0.454	0.691
4	0.191	0.261	0.118	0.329	0.519	0.429	0.598
5	0.293	0.341	0.108	0.339	0.469	0.529	0.609
6	0.134	0.141	0.217	0.229	0.579	0.479	0.559
7	0.114	0.372	0.224	0.276	0.459	0.569	0.591
8	0.157	0.248	0.215	0.325	0.596	0.601	0.660
9	0.177	0.273	0.264	0.235	0.579	0.509	0.570
10	0.189	0.451	0.248	0.347	0.429	0.532	0.641

distribution of attention over numerous words, rather than focusing on a single word or merely replicating words. This frequency data can be utilized to create new, less common words or to reproduce rarely employed terms found in the input text. When evaluated on the BillSum and FIRE datasets, our model's performance, as indicated by ROUGE-1, ROUGE-2, and ROUGE-L scores, is comparable to that of other models. The negligible variance in these ROUGE scores among the other methods indicates a comparable degree of performance across various parameters. We noted that the inclusion of frequency information does not enhance performance relative to the same models evaluated without the extractive approach layer; in fact, the model's performance exhibits a significant decline. When comparing models trained with 300 tokens to those trained with 450 tokens, the latter demonstrated enhanced performance relative to models lacking the extractive technique. Utilizing the transformer's attention mechanism, we have shown that our approach can provide summaries that conform to the document's intrinsic structure and efficiently leverage the hierarchical attention mechanism. Tables 5, and 6 provide a comparative examination of our method against current state-of-the-art (SOTA) techniques. These tables juxtapose the extractive ground truth summaries produced by Algorithm 2. These comparisons unequivocally demonstrate that our strategy substantially surpasses all other evaluated state-of-the-art methods regarding the quality of extracting ground truth summaries.

Moreover, our suggested strategy exhibits the ability to produce enhanced summarization outcomes across nearly all iterations of the ROUGE metric. This evidence supports the capability of our methodology to improve the quality of legal document summaries. This paper assesses six distinct

Table 11: Comparison of recall score with state-of-the-art techniques

Sr.#	TextRank [11]	LSA [23]	RBMs [20]	Summarizer [53]	KLSum [42]	Sumbasic [54]	Proposed Model
1	0.121	0.113	0.108	0.829	0.419	0.313	0.749
2	0.107	0.212	0.113	0.739	0.629	0.519	0.889
3	0.313	0.131	0.212	0.813	0.459	0.389	0.611
4	0.111	0.221	0.108	0.819	0.515	0.409	0.613
5	0.223	0.121	0.138	0.639	0.565	0.423	0.629
6	0.114	0.105	0.117	0.529	0.551	0.444	0.529
7	0.144	0.142	0.124	0.776	0.534	0.458	0.651
8	0.137	0.228	0.115	0.425	0.598	0.454	0.643
9	0.157	0.173	0.264	0.335	0.533	0.389	0.667
10	0.169	0.251	0.148	0.247	0.522	0.476	0.689

models—TextRank [33], LSA [12], RBM [53], CaseSummarizer [11], KLSum [42], and Sumbasic [55], utilizing the BillSum (US and CA) and FIRE legal judgment datasets. The evaluation data have been provided with the ROUGE-1, ROUGE-2, and ROUGE-L scores, along with precision and recall measures. Consequently, we chose the model with the greatest F1 score, as an elevated F1 score signifies superior overall performance. The second research question, "What is the scalability of the BERT-based hierarchical transformer method for summarizing legal documents?" is examined using these evaluation criteria and the results produced. Table 10 displays the precision scores achieved

Table 12: Comparison of F1 score with state-of-the-art techniques

Sr.#	TextRank [13]	LSA [11]	RBMs [31]	Summarizer [15]	KLSum [18]	Sumbasic [42]	Proposed Model
1	0.221	0.213	0.208	0.729	0.512	0.576	0.619
2	0.127	0.112	0.165	0.659	0.428	0.453	0.679
3	0.213	0.231	0.265	0.543	0.746	0.465	0.778
4	0.311	0.121	0.143	0.589	0.662	0.434	0.585
5	0.123	0.321	0.145	0.639	0.469	0.423	0.591
6	0.134	0.106	0.165	0.629	0.470	0.488	0.530
7	0.124	0.192	0.176	0.676	0.529	0.576	0.659
8	0.117	0.128	0.103	0.625	0.465	0.586	0.609
9	0.357	0.133	0.243	0.635	0.463	0.460	0.611
10	0.469	0.151	0.158	0.647	0.429	0.416	0.679

for 10 unique legal documents, summarized using TextRank [11], LSA [3], RBMs [31], Summarizer [1], KLSum [2], and Sumbasic [43]. According to these results, our suggested method exhibits outstanding precision with a score of 0.691, followed by CaseSummarizer [42] at 0.629, and TextRank with the lowest precision of 0.101. Table 11 presents the recall scores for each summarization model across the identical collection of documents. As indicated in Table 11, our model demonstrates the highest average recall scores of 0.889, followed by the CaseSummarizer [43] recall score of 0.829 and the LSA [40] recall value of 0.105. The F1 scores for the three summarization models across the 16 papers are displayed in Table 12. The average F1-scores for the three models across all documents are as follows: The F1 score of the proposed model is 0.778, KLSum [11] is 0.729, and RBM is 0.103. It is essential to recognize that although a correlation exists between ROUGE scores and precision/recall, a direct comparison is not uncomplicated. Rouge metrics evaluate the intersection of n-grams, while precision and recall rely on precise word correspondence. Thus, a model may excel in one metric while demonstrating inferior performance in another. Our model exhibits a valuable enhancement in processing power and memory efficiency relative to leading methodologies for summarizing legal documents, as evidenced by assessment criteria applied to two datasets. The experimental findings are analyzed, and performance

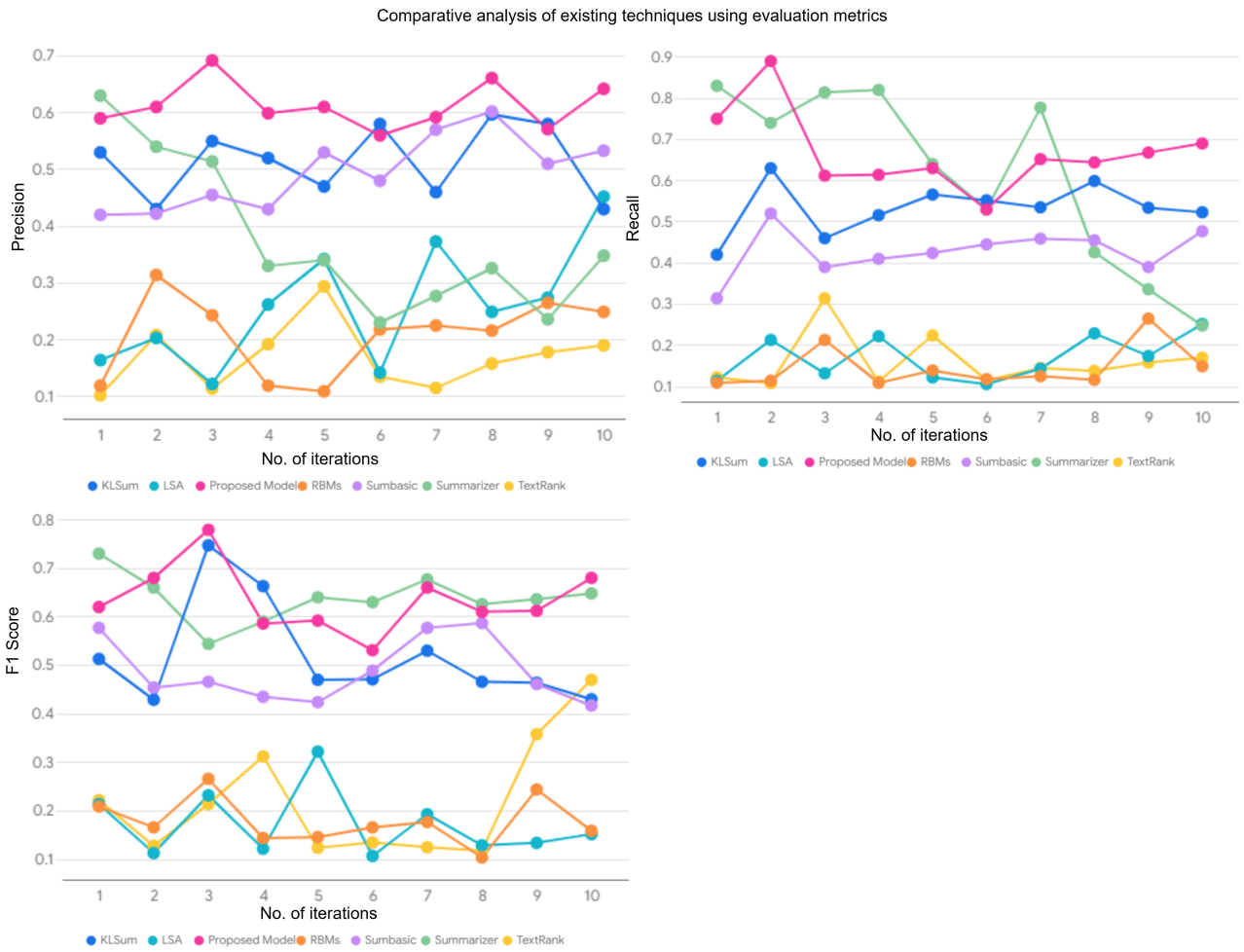

Figure 9: Experimental findings and performance comparison with existing techniques

is assessed utilizing Tables 10 - 12. Figure 9 juxtaposes the efficacy of several text-summarizing techniques across multiple criteria. The composition includes three subplots, each illustrating a distinct assessment metric: precision, recall, and F1 score. Each subplot illustrates the variation of these metrics across 10 data points, which represent distinct articles or cases of testing for each model. The x-axis denotes the data points, while the y-axis signifies the metric's score. Distinctively colored lines differentiate the efficiency of every model, thus facilitating an intuitive assessment of their efficacy in text summarization.

6 Conclusion

The computing demands for rapidly and accurately understanding extensive legal documents pose a significant obstacle. The development of efficient automated summarization techniques is crucial for addressing these difficulties. Extractive summarization, a commonly employed method, focuses on selecting salient sentences to condense lengthy documents. The inherent subjectivity of this activity and the challenges of acquiring contextual information within long legal texts make it tough, even for human experts. To address these challenges, we propose a novel approach utilizing hierarchical transformers. Our proposal is predicated on the stacked transformer encoder architecture of BERT. Due to the quadratic increase in computational costs associated with the self-attention mechanism of transformer models as sequence length expands, which limits their applicability to lengthy documents, we incorporate the Longformer. The Longformer's attention mechanism, which scales linearly with sequence length, facilitates the examination of texts including thousands of tokens or more. This mechanism replaces traditional self-attention with a hybrid approach that combines task-specific

global attention and localized windowed attention. The proposed model was evaluated utilizing two established benchmark datasets for long-sequence transformers: BillSum and FIRE. Our experimental results demonstrate that it exceeds prominent methodologies on both datasets. In the BillSum dataset, it achieves Rouge-1, Rouge-2, and Rouge-L ratings of 47.11, 33.02, and 42.19, respectively. The FIRE dataset has scores of 58.43, 44.31, and 41.54. These results underscore the superior effectiveness of our proposed method compared to existing leading strategies.

References

[1] Sharma, S., Srivastava, S., Verma, P., Verma, A., Chaurasia, S.N.: A comprehen sive analysis of indian legal documents summarization techniques. SN Computer Science 4(5), 614 (2023)

[2] Jain, D., Borah, M.D., Biswas, A.: Fine-tuning textrank for legal document sum marization: A bayesian optimization based approach. In: Proceedings of the 12th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 41–48 (2020)

[3] Parikh, V., Mathur, V., Mehta, P., Mittal, N., Majumder, P.: Lawsum: A weakly supervised approach for indian legal document summarization. arXiv preprint arXiv:2110.01188 (2021)

[4] Zhang, J., Zhao, Y., Saleh, M., Liu, P.: Pegasus: Pre-training with extracted gap sentences for abstractive summarization. In: International Conference on Machine Learning, pp. 11328–11339 (2020). PMLR

[5]] Wang, X., Wu, Y.C.: Empowering legal justice with ai: A reinforcement learning sac-vae framework for advanced legal text summarization. PloS one 19(10), 0312623 (2024)

[6] Jain, D., Borah, M.D., Biswas, A.: A sentence is known by the company it keeps: Improving legal document summarization using deep clustering. Artificial Intelligence and Law 32(1), 165–200 (2024)

[7] Yang, S., Zhang, S., Fang, M., Yang, F., Liu, S.: A hierarchical representation model based on longformer and transformer for extractive summarization. Electronics 11(11), 1706 (2022)

[8] Deroy, A., Ghosh, K., Ghosh, S.: Applicability of large language models and generative models for legal case judgement summarization. Artificial Intelligence and Law, 1–44 (2024)

[9] Ragazzi, L., Moro, G., Guidi, S., Frisoni, G.: Lawsuit: a large expert-written summarization dataset of italian constitutional court verdicts. Artificial Intelligence and Law, 1–37 (2024)

[10] Jain, D., Borah, M.D., Biswas, A.: Automatic summarization of legal bills: A comparative analysis of classical extractive approaches. In: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), pp. 394–400 (2021). IEEE

[11] Jain, D., Borah, M.D., Biswas, A.: Summarization of lengthy legal documents via abstractive dataset building: An extract-then-assign approach. Expert Systems with Applications 237, 121571 (2024)

[12] Prasad, A., Noushad, A., Gopi, N.K., Raju, R., KS, M.P.: An overview of legal document summarization techniques. irjmets (2023)

[13] Yadav, A.K., Ranvijay, Yadav, R.S., Maurya, A.K.: Graph-based extractive text summarization based on single document. Multimedia Tools and Applications 83(7), 18987–19013 (2024)

[14] Shukla, A., Bhattacharya, P., Poddar, S., Mukherjee, R., Ghosh, K., Goyal, P., Ghosh, S.: Legal case document summarization: Extractive and abstractive methods and their evaluation. arXiv preprint arXiv:2210.07544 (2022)

[15] Deroy, A., Bhattacharya, P., Ghosh, K., Ghosh, S.: An analytical study of algo rithmic and expert summaries of legal cases. In: Legal Knowledge and Information Systems, pp. 90–99. IOS Press, ??? (2021)

[16] Elaraby, M., Xu, H., Gray, M., Ashley, K.D., Litman, D.: Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In: Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval)@ LREC-COLING 2024, pp. 28–35 (2024)

[17] Sadafale, K.B., Thorat, S.A.: Review on text summarization using clustering and machine learning-deep learning models. In: 2024 International Conference on Cognitive Robotics and Intelligent Systems (ICC-ROBINS), pp. 44–51 (2024). IEEE

[18] Fabbri, A.R., Kry´sci´nski, W., McCann, B., Xiong, C., Socher, R., Radev, D.: Sum meval: Re-evaluating summarization evaluation. Transactions of the Association for Computational Linguistics 9, 391–409 (2021)

[19] Mohammadi, M., Rezaei, J.: Ensemble ranking: Aggregation of rankings produced by different multi-criteria decision-making methods. Omega 96, 102254 (2020)

[20] Alami, N., Mallahi, M.E., Amakdouf, H., Qjidaa, H.: Hybrid method for text summarization based on statistical and semantic treatment. Multimedia Tools and Applications 80, 19567–19600 (2021)

[21] Bauer, E., Stammbach, D., Gu, N., Ash, E.: Legal extractive summarization of us court opinions. arXiv preprint arXiv:2305.08428 (2023)

[22] Alkaissi, H., McFarlane, S.I.: Artificial hallucinations in chatgpt: implications in scientific writing. Cureus 15(2) (2023)

[23] Koh, H.Y., Ju, J., Liu, M., Pan, S.: An empirical survey on long document sum marization: Datasets, models, and metrics. ACM computing surveys 55(8), 1–35 (2022)

[24] Chalkidis, I., Jana, A., Hartung, D., Bommarito, M., Androutsopoulos, I., Katz, D.M., Aletras, N.: Lexglue: A benchmark dataset for legal language understanding in english. arXiv preprint arXiv:2110.00976 (2021)

[25] Deroy, A., Ghosh, K., Ghosh, S.: Ensemble methods for improving extractive summarization of legal case judgements. Artificial Intelligence and Law 32(1), 231–289 (2024)

[26] Mamakas, D., Tsotsi, P., Androutsopoulos, I., Chalkidis, I.: Processing long legal documents with pre-trained transformers: Modding legalbert and longformer. arXiv preprint arXiv:2211.00974 (2022)

[27] Kanapala, A., Pal, S., Pamula, R.: Text summarization from legal documents: a survey. Artificial Intelligence Review 51, 371–402 (2019)

[28] Filippova, K.: Controlled hallucinations: Learning to generate faithfully from noisy data. arXiv preprint arXiv:2010.05873 (2020)

[29] Umer, M., Ashraf, I., Mehmood, A., Kumari, S., Ullah, S., Sang Choi, G.: Sentiment analysis of tweets using a unified convolutional neural network-long short-term memory network model. Computational Intelligence 37(1), 409–434 (2021)

[30] Deroy, A., Ghosh, K., Ghosh, S.: How ready are pre-trained abstractive models and llms for legal case judgement summarization? arXiv preprint arXiv:2306.01248 (2023)

[31] Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., Hashimoto, T.B.: Benchmarking large language models for news summarization. Transactions of the Association for Computational Linguistics 12, 39–57 (2024)

[32] Parikh, V., Bhattacharya, U., Mehta, P., Bandyopadhyay, A., Bhattacharya, P., Ghosh, K., Ghosh, S., Pal, A., Bhattacharya, A., Majumder, P.: Aila 2021: Shared task on artificial intelligence for legal assistance. In: Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation, pp. 12–15 (2021)

[33] Laban, P., Schnabel, T., Bennett, P.N., Hearst, M.A.: Summac: Re-visiting nli based models for inconsistency detection in summarization. Transactions of the Association for Computational Linguistics 10, 163–177 (2022)

[34] Moro, G., Ragazzi, L.: Semantic self-segmentation for abstractive summariza tion of long documents in low-resource regimes. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11085–11093 (2022)

[35] Trautmann, D.: Large language model prompt chaining for long legal document classification. arXiv preprint arXiv:2308.04138 (2023)

[36] Kalamkar, P., Agarwal, A., Tiwari, A., Gupta, S., Karn, S., Raghavan, V.: Named entity recognition in indian court judgments. arXiv preprint arXiv:2211.03442 (2022)

[37] Stanczak, K., Augenstein, I.: A survey on gender bias in natural language processing. arXiv preprint arXiv:2112.14168 (2021)

[38] Chalkidis, I., Dai, X., Fergadiotis, M., Malakasiotis, P., Elliott, D.: An exploration of hierarchical attention transformers for efficient long document classification. arXiv preprint arXiv:2210.05529 (2022)

[39] Trautmann, D., Petrova, A., Schilder, F.: Legal prompt engineering for multilin gual legal judgement prediction. arXiv preprint arXiv:2212.02199 (2022)

[40] Grail, Q., Perez, J., Gaussier, E.: Globalizing bert-based transformer architectures for long document summarization. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 1792–1810 (2021)

[41] Clark, K.: Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555 (2020)

[42] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems 30 (2017)

[43] Marino, G., Licari, D., Bushipaka, P., Comand´e, G., Cucinotta, T.: Automatic rhetorical roles classification for legal documents using legal-transformeroverbert. CEUR Workshop Proceedings (2023)

[44] Ghosh, S., Dutta, M., Das, T.: Indian legal text summarization: A text normalization-based approach. In: 2022 IEEE 19th India Council International Conference (INDICON), pp. 1–4 (2022). IEEE

[45] Gu, N., Ash, E., Hahnloser, R.H.: Memsum: Extractive summarization of long documents using multi-step episodic markov decision processes. arXiv preprint arXiv:2107.08929 (2021)

[46] Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does nlp benefit legal system: A summary of legal artificial intelligence. arXiv preprint arXiv:2004.12158 (2020)

[47] Iskender, N., Polzehl, T., M¨oller, S.: Reliability of human evaluation for text summarization: Lessons learned and challenges ahead. In: Proceedings of the Workshop on Human Evaluation of NLP Systems (HumEval), pp. 86–96 (2021)

[48] Verma, P., Om, H.: Fuzzy evolutionary self-rule generation and text summariza tion. In: 15th International Conference on Natural Language Processing, p. 115 (2018)

[49] Ba, J.L., Kiros, J.R., Hinton, G.E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)

[50] Wagh, V., Khandve, S., Joshi, I., Wani, A., Kale, G., Joshi, R.: Comparative study of long document classification. In: TENCON 2021-2021 IEEE Region 10 Conference (TENCON), pp. 732–737 (2021). IEEE

[51] Park, H.H., Vyas, Y., Shah, K.: Efficient classification of long documents using transformers. arXiv preprint arXiv:2203.11258 (2022)

[52] Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., Kambhatla, N.: Text simplification for legal domain:I nsights and challenges. In: Proceedings of the Natural Legal Language Processing Workshop 2022, pp. 296–304 (2022)

[53] Agarwal, A., Xu, S., Grabmair, M.: Extractive summarization of legal deci sions using multi-task learning and maximal marginal relevance. arXiv preprint arXiv:2210.12437 (2022)

[54] Cui, P., Hu, L.: Sliding selector network with dynamic memory for extractive summarization of long documents. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5881–5891 (2021)

[55] Henderson, P., Krass, M., Zheng, L., Guha, N., Manning, C.D., Jurafsky, D., Ho, D.: Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. Advances in Neural Information Processing Systems 35, 29217–29234 (2022)