Acadlore Transactions on AI and Machine Learning
https://www.acadlore.com/journals/ATAIML

A Hybrid Graph-Attention and Contextual Sentiment Embedding Model for Sentiment Analysis in Legal Documents

Sulaxan Jadhav[1*], Ashvini Pradeep Shende[1], Samruddhi Sapkal[2]

[1] Symbiosis School of Economics, Symbiosis International (Deemed University), 411004 Pune, India
[2] Department Computer Engineering, ISBM College of Engineering, 412115 Pune, India

[*] Correspondence: Sulaxan Jadhav (sulaxan.jadhav@sse.ac.in)

Received: 02-06-2025 **Revised:** 03-18-2025 **Accepted:** 03-23-2025

Citation: S. Jadhav, A. P. Shende and S. Sapkal, "A hybrid graph-attention and contextual sentiment embedding model for sentiment analysis in legal documents," *Acadlore Trans. Mach. Learn.*, vol. 4, no. 2, pp. 62–81, 2025. https://doi.org/10.56578/ataiml040201.

Abstract: Sentiment analysis in legal documents presents significant challenges due to the intricate structure, domain-specific terminology, and strong contextual dependencies inherent in legal texts. In this study, a novel hybrid framework is proposed, integrating Graph Attention Networks (GATs) with domain-specific embeddings, i.e., Legal Bidirectional Encoder Representations from Transformers (LegalBERT) and an aspect-oriented sentiment classification approach to improve both predictive accuracy and interpretability. Unlike conventional deep learning models, the proposed method explicitly captures hierarchical relationships within legal texts through GATs while leveraging LegalBERT to enhance domain-specific semantic representation. Additionally, auxiliary features, including positional information and topic relevance, were incorporated to refine sentiment predictions. A comprehensive evaluation conducted on diverse legal datasets demonstrates that the proposed model achieves state-of-the-art performance, attaining an accuracy of 93.1% and surpassing existing benchmarks by a significant margin. Model interpretability was further enhanced through SHapley Additive exPlanations (SHAP) and Legal Context Attribution Score (LCAS) techniques, which provide transparency into decision-making processes. An ablation study confirms the critical contribution of each model component, while scalability experiments validate the model's efficiency across datasets ranging from 10,000 to 200,000 sentences. Despite increased computational demands, strong robustness and scalability are exhibited, making this framework suitable for large-scale legal applications. Future research will focus on multilingual adaptation, computational optimization, and broader applications within the field of legal analytics.

Keywords: Sentiment analysis; GATs; Domain-specific embeddings; BERT; Legal document analysis; Legal BERT

1 Introduction

Sentiment analysis, a fundamental task in natural language processing (NLP), has been widely explored in domains such as social media analytics, product reviews, and consumer feedback [1]. However, its application to legal documents introduces unique challenges due to the technical, nuanced, and highly structured nature of legal texts [2]. Unlike conventional texts, legal documents encapsulate sentiments that are deeply intertwined with logical arguments, hierarchical structures, and domain-specific language, making standard sentiment analysis techniques insufficient in capturing the underlying intent and emotion [3]. This limitation is particularly critical as sentiment analysis in the legal domain has significant implications, including contract review, judicial decision-making, and case law analysis. The primary challenge lies in the complexity of legal documents, which exhibit characteristics such as (i) lengthy and intricate argumentation structures, (ii) reliance on domain-specific terminology and jurisprudential syntax, (iii) context-dependent sentiments that span across multiple sentences or paragraphs, and (iv) logical relationships between clauses, arguments, and judicial decisions [4–6]. Existing sentiment analysis models struggle to effectively process these intricacies, as traditional machine learning and deep learning methods primarily focus on token- or sentence-level analysis, failing to capture the broader relational and structural dependencies inherent in legal texts [7, 8]. Even transformer-based models such as BERT and its domain-specific variants like LegalBERT, while advancing NLP for legal language, are limited in their ability to incorporate hierarchical and contextual structures essential for sentiment interpretation in legal settings [9, 10]. To address these critical gaps, a novel hybrid approach that integrates GATs with domain-specific contextual embeddings was proposed, enabling a structured and context-aware sentiment analysis framework tailored for legal documents. The key innovation of the proposed approach is its ability to represent legal documents as directed graphs, where nodes correspond to sentences and edges represent argumentative and logical relationships derived through dependency parsing. By incorporating graph attention mechanisms, the proposed model effectively captures inter-sentence dependencies, hierarchical structures, and the overarching legal context, which are often overlooked in traditional NLP models. Additionally, fine-tuned legal language models, trained on sentiment-annotated legal corpora, ensure that embeddings accurately reflect domain-specific sentiment expressions. A further enhancement of the proposed methodology lies in its aspect-oriented sentiment analysis module, which enables a more granular understanding of sentiments associated with distinct legal constructs, such as judgments, evidence, and arguments. This disentangled sentiment analysis not only improves classification accuracy but also provides more actionable insights for legal practitioners. Furthermore, by incorporating explainable artificial intelligence (AI) techniques, the proposed model ensures interpretability and transparency, addressing the critical need for accountability in high-stakes legal applications.

By effectively tackling the inherent complexities of legal texts and integrating innovative techniques for sentiment modeling, the proposed methodology sets a new benchmark for sentiment analysis in the legal domain. Through extensive experimentation and comparative analysis, the method demonstrates its superiority over existing approaches in terms of accuracy, scalability, and interpretability. This research contributes a robust, structured, and explainable framework that advances the field of legal NLP and unlocks valuable applications in legal analytics, judicial decision support, and contract analysis. The remainder of this study is organized as follows: Section 2 presents a review of related work, Section 3 details the proposed methodology, Section 4 describes the experimental setup and results, Section 5 provides a comparative analysis with existing methods, Section 6 discusses limitations and future work, and Section 7 concludes the study.

2 Related Works

The application of sentiment analysis in the legal domain has garnered increasing attention due to its potential to enhance judicial efficiency and accessibility. Numerous studies have explored various methodologies, ranging from traditional machine learning approaches to cutting-edge deep learning techniques, to tackle the unique challenges posed by legal texts. This section discusses relevant works that form the foundation of sentiment analysis in legal documents and highlights the gaps addressed by this research. Early works in sentiment analysis focused on applying deep learning techniques to specific legal sub-domains. For example, Abimbola et al. [11] proposed a framework combining Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNNs) for maritime judiciary records. The model successfully extracted critical sentiment-laden statements from distributed legal repositories, aiding judges in making informed decisions. While effective, this approach primarily focuses on feature extraction without addressing the relational structure of arguments in legal documents, a key challenge that the framework proposed in this current study aims to overcome. Building on the trend of multi-label classification in legal contexts, Sengupta [12] developed a pipeline combining supervised machine learning and NLP techniques to classify legal cases under the Indian Income Tax Act. By employing transformer-based techniques and traditional machine learning models like Support Vector Machines (SVMs), the applicability of such hybrid methods for legislative text analysis was demonstrated. However, the reliance on structured datasets and limited application scope underscores the need for more generalized frameworks, a gap that this current research seeks to address by leveraging graph-based techniques.

The limitations of manual judgment and inefficiencies in official document sentiment recognition were highlighted in the study by Hao et al. [13], where a BERT-SVM hybrid model achieved an impressive 95.12% accuracy in sentiment classification. This work demonstrates the potential of transformer-based architectures for legal sentiment analysis but lacks a focus on domain-specific nuances and argumentative structures critical for legal contexts. Similarly, Dey and Das [14] introduced a modified term frequency-inverse document frequency (TF-IDF) method combined with neural networks to enhance sentiment classification. The innovative use of global weighting factors and k-best selection highlights the importance of preprocessing for improved feature representation. Beyond sentiment analysis, related studies have explored complementary NLP tasks in the legal domain. For instance, the development of legal-specific Named Entity Recognition (NER) datasets in the study by Naik et al. [15] showcases the need for domain-specific training resources. Additionally, Jain et al. [16] and Gupta et al. [17] proposed novel approaches for summarization and topic modelling of legal texts, emphasizing the significance of understanding document-level structures for downstream applications.

While these studies represent significant advancements, their limitations highlight the need for a framework tailored to the intricacies of legal documents. The model proposed in this current study bridges this gap by integrating graph-based relational reasoning with domain-specific embeddings, enabling a deeper understanding of the hierarchical and argumentative structures in legal texts.

3 Proposed Methodology

The proposed methodology involves designing and implementing a hybrid framework for sentiment analysis in legal documents. The approach combines graph-based neural networks with transformer-based language models to capture the hierarchical and relational structure of legal texts. This framework integrates three core components: a graph representation of legal documents, domain-specific embeddings for contextual understanding, and an aspect-oriented sentiment classification module. These components aim to ensure that the analysis is both precise and interpretable, addressing the complex nature of legal sentiment extraction. The flowchart of the proposed algorithm is shown in Figure 1. The proposed methodology is designed to address the unique challenges of sentiment analysis in legal documents, which include the complexity of their hierarchical structures, the intricate relationships between clauses and arguments, and the specialized vocabulary and syntax inherent to the legal domain. Traditional sentiment analysis methods fail to capture these nuances, leading to a lack of contextual accuracy and interpretability. To overcome these challenges, the methodology integrates three core components:

a) Graph representation of legal texts: Legal documents are represented as directed graphs where nodes correspond to sentences or clauses, and edges capture logical and argumentative relationships derived from dependency parsing. This enables the model to learn hierarchical and relational structures effectively.

b) Domain-specific embeddings: Pre-trained legal language models, such as LegalBERT, are fine-tuned on sentiment-annotated legal corpora to generate embeddings that reflect the nuanced semantics and sentiment expressions unique to legal texts.

c) Aspect-oriented sentiment classification: Sentiments are analysed and classified based on specific legal constructs (e.g., judgments, evidence, or arguments), ensuring a deeper understanding of the document's sentiment distribution and its interpretability for legal practitioners.

These components work together to ensure precise, context-aware, and interpretable sentiment analysis tailored to the complex nature of legal documents.

The rationale behind selecting these techniques lies in their ability to capture intricate relationships, contextual dependencies, and domain-specific linguistic structures. GATs were chosen over traditional graph-based models like Graph Convolutional Networks (GCNs) due to their ability to dynamically assign varying importance to textual components, making them more effective for modeling legal texts, which often contain complex interdependencies between clauses, arguments, and case law references. LegalBERT was selected over general-purpose language models such as Robustly Optimized BERT Pretraining Approach (RoBERTa) and T5 [18], as it is specifically pre-trained on legal corpora, ensuring better representation of legal terminology and syntax. Aspect-oriented classification was integrated to enhance interpretability and fine-grained sentiment classification, which is crucial in legal applications where sentiment polarity often depends on specific case details. While alternative methods such as LSTM or CNN could have been considered, they lack the ability to capture long-range dependencies as effectively as transformer-based architectures. The comparison criteria for these models were based on accuracy, F1-score, interpretability, and computational efficiency, ensuring a robust evaluation. By leveraging these techniques, the proposed model achieves a balance between domain specificity, contextual understanding, and computational feasibility, making it well-suited for legal sentiment analysis.

Let $D = \{s_1, s_2, s_3 \ldots\ldots, s_n\}$ represent a legal document with n sentences, where s_j corresponds to the i-th sentence. A directed graph $G = (V, E)$ was constructed, where, V is the set of nodes corresponding to the sentences $s_1, s_2, s_3 \ldots\ldots, s_n$ and E is the set of directed edges that represent the relationship between the sentences. Edges $\left(e_{ij} \in E\right)$ were established based on dependency parsing, capturing logical relationships such as causality, contrast, or elaboration. Each node $(v_i \in V)$ was initialised with an embedding vector $\left(h_i \in R^d\right)$, where d is the dimensionality of the embedding space. These embeddings were obtained using a pre-trained language model like LegalBERT [9, 10], fine-tuned on the legal corpora. The graph was then processed using a GAT, which computes the updated embeddings h_i' for each node by aggregating information from its neighbours and this is represented as:

$$h_i' = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij} W h_j\right) \tag{1}$$

where, $\mathcal{N}(i)$ represents the set of neighbouring nodes of v_i and α_{ij} is the attention weight for the edge from the nodes v_j to v_i, given by Eq. (2).

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^\top \left[W h_i \| W h_j\right]\right)\right)}{\sum_{k \in \mathcal{N}(i)} \exp\left(\text{LeakyReLU}\left(a^\top \left[W h_i \| W h_k\right]\right)\right)} \tag{2}$$

where, W is a learnable weight matrix, a is a learnable attention vector and $\|$ denotes concatenation. This process generates contextualised embeddings h_i', which encode both the semantic information of each sentence and its relationships with other sentences in the document. These embeddings were then used as input for downstream sentiment classification tasks. The graph-based representation ensures that the model captures both the fine-grained and holistic context of legal texts, essential for accurate sentiment analysis.

Algorithm 1 constructs a graph-based representation of legal documents and updates the embeddings of sentences using a GAT to capture the relationships between sentences. The process begins by treating each sentence in the document as a node and using a pre-trained language model (e.g., LegalBERT) to generate an initial semantic embedding for each node. These embeddings represent the individual meanings of the sentences.

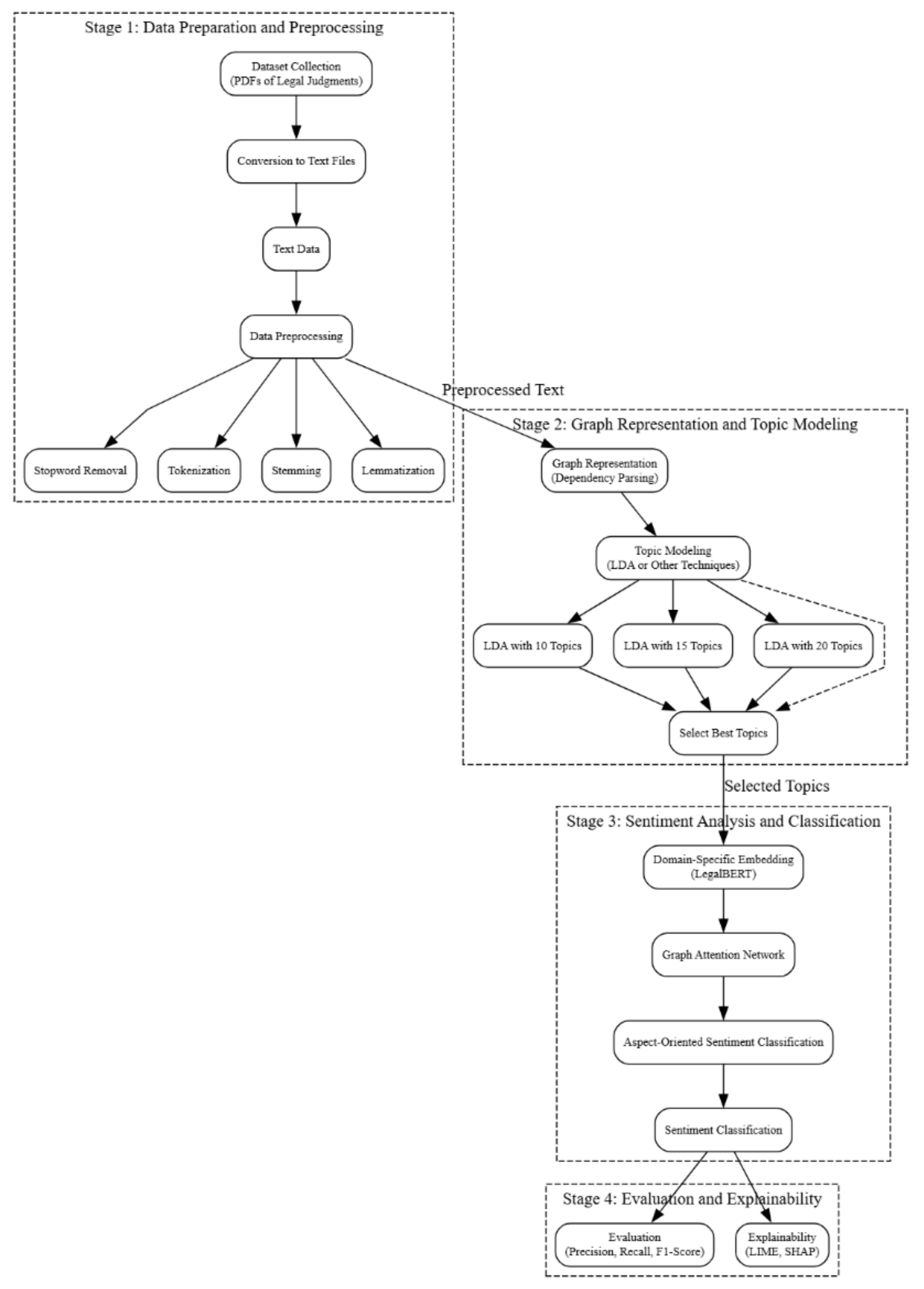
Stage 1: Data Preparation and Preprocessing
Dataset Collection (PDFs of Legal Judgments)
Conversion to Text Files
Text Data
Data Preprocessing
Stopword Removal
Tokenization
Stemming
Lemmatization
Preprocessed Text
Stage 2: Graph Representation and Topic Modeling
Graph Representation (Dependency Parsing)
Topic Modeling (LDA or Other Techniques)
LDA with 10 Topics
LDA with 15 Topics
LDA with 20 Topics
Select Best Topics
Selected Topics
Stage 3: Sentiment Analysis and Classification
Domain-Specific Embedding (LegalBERT)
Graph Attention Network
Aspect-Oriented Sentiment Classification
Sentiment Classification
Stage 4: Evaluation and Explainability
Evaluation (Precision, Recall, F1-Score)
Explainability (LIME, SHAP)

Figure 1. Schematic representation for the flow diagram of the proposed methodology

Algorithm 1 Graph-based representation for legal texts

```
**Input:** Legal documents $D = \{s_1, s_2, s_3, \ldots, s_n\}$, pre-trained embedding model $M$, and dependency parser $P$.
**Output:** Contextualized sentence embeddings $\{h'_1, h'_2, h'_3, \ldots, h'_n\}$
function (Graph representation) $(D, M, P)$
    $V \leftarrow \emptyset$    (Initialize empty set of nodes)
    $E \leftarrow \emptyset$    (Initialize empty set of edges)
    for each sentence $s_i$ in $D$ do
        $v_i \leftarrow M(s_i)$
        Add $v_i$ to $V$
    end for
    for each pair of sentences $(s_i, s_j)$ in $D$ do
        if $P(s_i, s_j)$ indicates a relationship then
            Add edge $e_{ij} = (v_i \rightarrow v_j)$ to $E$
        end if
    end for
    for each node $v_i$ in $V$ do
        $h'_i \leftarrow$ Update Embedding$(v_i, \{v_j \mid (v_j \rightarrow v_i) \in E\})$
    end for
    return $\{h'_1, h'_2, h'_3, \ldots, h'_n\}$
end function
```

Next, the relationships between sentences, such as logical connections (e.g., causality, contrast, or elaboration), were identified using a dependency parser. These relationships form the edges in the graph, connecting related nodes and encoding the flow of arguments in the document. The resulting graph structure $G = (V, E)$, with V as nodes (sentences) and E as edges (relationships), represents the document holistically. The GAT was then applied to this graph to update the sentence embeddings. At each node, the GAT aggregates information from its neighboring nodes by assigning importance scores (attention) to each neighbor. This allows the model to focus more on relevant sentences while minimizing the influence of less critical ones. The updated embeddings at each node thus incorporate both the semantic meaning of the sentence and the contextual information derived from its relationships in the graph. By iterating through all nodes in the graph and updating their embeddings, the algorithm creates a rich, context-aware representation of the document. These updated embeddings were then used in downstream tasks, such as sentiment analysis, where understanding the interdependencies between sentences is crucial for accurate predictions. This graph-based approach ensures that the model captures not just individual sentence meanings but also the broader argumentative and relational structure of legal documents, making it highly effective for analysing complex legal texts.

3.1 Domain-Specific Embeddings

Legal texts are characterized by highly specialized language, complex terminology, and intricate syntactical structures, making general-purpose language models inadequate for capturing their semantics. To address this, the second stage of the proposed methodology involves generating domain-specific embeddings using a pre-trained legal language model, such as LegalBERT, fine-tuned on sentiment-annotated legal corpora. These embeddings are designed to capture the contextual nuances of legal language and align them with sentiment-based representations. Given a sentence s in a legal document, its embedding E_s is obtained as Eq. (3).

$$E_s = \text{LegalBERT}(s, \theta) \tag{3}$$

where, θ represents the parameters of the LegalBERT model, which have been pre-trained on a large corpus of legal texts and fine-tuned on sentiment-specific tasks. Fine-tuning involves minimizing a loss function, such as cross-entropy loss for classification tasks, over a training set $\mathcal{D} = (s_i, y_i)$, where, s_i is a sentence and y_i is its corresponding sentiment label. To enrich the embeddings, additional domainspecific information, such as topic relevance or clause type, was integrated. This was achieved by extending the embeddings with auxiliary features F_s as described in Table 1, yielding a final embedding as follows:

$$E_s^{\text{final}} = [E_s \| F_s] \tag{4}$$

where, F_s includes attributes such as the topic distribution from Latent Dirichlet Allocation (LDA) or positional information within the document hierarchy. These embeddings were then normalized to ensure numerical stability and compatibility for downstream sentiment classification as described in Algorithm 2.

The embedding generation pipeline is computationally efficient and scalable, allowing the integration of legal domain-specific features while preserving the rich contextual representations provided by the transformer architecture.

Table 1. Auxiliary features for domain-specific embeddings

Feature Name	Description	Calculation Method	Range
Topic distribution	Relevance of the sentence to predefined legal topics	Generated via LDA (topic modelling)	$[0, 1]$
Positional information	Relative position of the sentence in the document	Normalized sentence index	$[0, 1]$
Clause type indicator	Type of clause (e.g., judgment, evidence, and argument)	Rule-based NLP classification	Binary $(0, 1)$
Sentence length	Length of the sentence in terms of token count	Direct token count	Positive integer

Algorithm 2 Domain-specific embedding generation

```
Input: Legal document $D = \{s_1, s_2, ..., s_n\}$, pre-trained LegalBERT model $M$, and auxiliary features $F$
Output: Embedding set $E_{\text{final}} = E_{s1}^{\text{final}}, ..., E_{sn}^{\text{final}}$
function (Generate embeddings) $(D, M, F)$
    $E_{\text{final}} \leftarrow \emptyset$    (Initialize empty set for final embeddings)
    for each sentence $s_i$ in $D$ do
        $E_{si} \leftarrow M(s_i)$
        $F_{si} \leftarrow$ Extract Auxiliary Features$(s_i, F)$
        $E_{si\text{final}} \leftarrow$ Concatenate$(E_{si}, F_{si})$
        $E_{si\text{final}} \leftarrow$ Normalize$(E_{si\text{final}})$
        Add $E_{si\text{final}}$ to $E_{\text{final}}$
    end for
    return $E_{\text{final}}$
end function
```

The auxiliary features are domain-specific enhancements incorporated into the LegalBERT-generated embeddings to improve their alignment with the unique requirements of legal texts. These features include the topic distribution, which provides probabilistic scores for the relevance of a sentence to predefined legal topics, such as evidence or arguments, using topic modelling techniques like LDA. Positional information captures the relative position of a sentence within the document, a critical factor in legal reasoning where the placement of a sentence often influences its significance. The clause type indicator assigns binary labels to clauses, identifying whether they belong to key legal constructs such as judgments or arguments, ensuring that embeddings are aware of sentence-level legal functions. Lastly, sentence length introduces a structural attribute by incorporating the token count of a sentence, as longer sentences frequently carry more complex arguments in legal texts. Together, these features fine-tune the embeddings, enabling them to address the complexities of legal texts and preparing them for downstream tasks such as aspect-oriented sentiment classification.

3.2 Aspect-Oriented Sentiment Classification

The aspect-oriented sentiment classification module focuses on breaking down the sentiment analysis into specific legal constructs or "aspects," such as judgments, evidence, arguments, and legal clauses. This enables a granular understanding of how sentiment varies across different sections of legal documents, ensuring that each construct is evaluated independently while contributing to the overall sentiment of the document. Each aspect is treated as a distinct classification task, where embeddings generated from the previous stages are input into specialized classifiers for each aspect. This process is explained by Algorithm 3.

Given the embeddings E_s^{final} from the domain-specific embedding stage, the classification task maps these embeddings into sentiment categories for each aspect. Mathematically, for an aspect A_i, the sentiment score is predicted using a function f as follows:

$$\hat{y}_i = f\left(E_s^{\text{final}}, A_i; \Theta_i\right) \tag{5}$$

where, $\hat{y}_i$ is the predicted sentiment label for aspect $A_i; \Theta_i$ represents the learnable parameters for the classifier corresponding to A_i; and f is feed-forward neural network or transformer-based layer. To manage multi-aspect classification, the framework utilizes a shared embedding layer to ensure efficient parameter sharing while maintaining separate classifiers for each aspect, as given by Figure 2. This modular design allows the addition of new aspects without retraining the entire pipeline, enhancing scalability.

Algorithm 3 Aspect-oriented sentiment classification

```
Input: Final embedding set $E_{\text{final}} = E_{s1}^{\text{final}}, \ldots, E_{sn}^{\text{final}}$, aspect set $A = \{A1, A2, \ldots, Am\}$
Output: Predicted sentiments $Y = \{y_1, \ldots, y_m\}$ for all aspects
function (Aspect classification) $(E_{\text{final}}, A)$
    $Y \leftarrow \emptyset$    (Initialize empty set for predictions)
    for each aspect $A_i$ in $A$ do
        $Classifier \leftarrow$ Initialize classifier$(A_i)$
        for each embedding $E_s$ in $E_{\text{final}}$ do
            $y_s \leftarrow$ Classifier$(E_s)$
            Add $y_s$ to $Y$
        end for
    end for
    return $Y$
end function
```

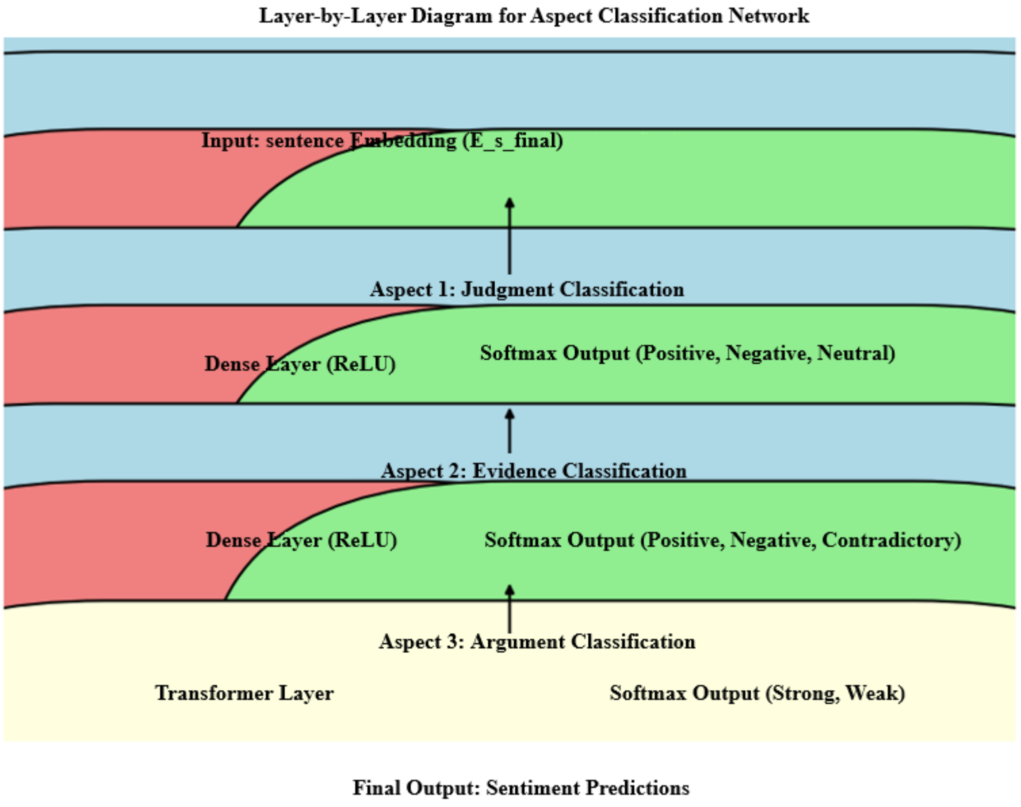
Layer-by-Layer Diagram for Aspect Classification Network
Input: sentence Embedding (E_s_final)
Aspect 1: Judgment Classification
Dense Layer (ReLU)
Softmax Output (Positive, Negative, Neutral)
Aspect 2: Evidence Classification
Dense Layer (ReLU)
Softmax Output (Positive, Negative, Contradictory)
Aspect 3: Argument Classification
Transformer Layer
Softmax Output (Strong, Weak)
Final Output: Sentiment Predictions

Figure 2. Schematic representation of the layered architecture of the aspect classification network

Table 2. Aspects and classification complexity

Aspect Name	Description	Sentiment Categories	Classifier Type	Input Features
Judgment	Sentiment in final judicial decisions	Positive, negative, and neutral	Feed-forward neural network	E_s^{final}
Evidence	Sentiment in supporting evidence	Positive, negative, and contradictory	Multi-Layer Perceptron (MLP)	$E_s^{\text{final}} + Fs$
Argument	Sentiment in legal arguments presented	Strong and weak	Transformer-based layer	E_s^{final}
Clause	Sentiment in specific legal clauses	Agreement, disagreement, and neutral	Logistic regression	$E_s^{\text{final}} + Fs$

In Table 2, judgment sentiment evaluates the sentiment polarity in judicial decisions, such as whether the judgment leans toward a favourable or unfavourable outcome. It uses a feed-forward neural network for simplicity, as judgments are often explicit. Evidence sentiment supports evidence, which is more nuanced and often relies on auxiliary features, such as clause type or topic distribution, making an MLP a better fit due to its capacity for multi-feature integration. Argument sentiment is defined when arguments are often complex, requiring the model to assess their strength or persuasiveness. A transformer-based layer is used to handle these complexities effectively by capturing contextual dependencies. Finally, sentiments at the clause level, such as agreement or disagreement, are simpler and can be modelled effectively using logistic regression $E_s^{\text{final}} + Fs$.

3.3 Explainability and Interpretability Evaluation

In high-stakes domains like law, explainability is critical to ensure trust and accountability in AI-driven sentiment analysis systems. The proposed methodology incorporates three key techniques: Layer-wise Relevance Propagation (LRP), SHAP values, and a novel metric (LCAS). LRP assigns relevance scores to input features (e.g., words or phrases) by propagating the model's prediction back to the input, revealing the contribution of each feature to the sentiment prediction. SHAP quantifies the marginal impact of each input feature by treating it as a cooperative game and using SHAP values to measure its importance in the final prediction. LCAS, the novel contribution of this study, extends these methods by introducing a domain-specific measure that evaluates the coherence between the sentiment prediction and the logical structure of legal documents. It assigns a score based on the consistency of the prediction with hierarchical legal constructs, such as judgments, arguments, and evidence, ensuring that the model aligns with legal reasoning.

• LRP: This computes the relevance R_k of each input feature x_k based on the model's output y. The relevance is propagated backward through the layers of the network and is given by Eq. (6).

$$R_k = \sum_j \frac{z_{kj}}{\sum_{k'} z_{k'j}} R_j \tag{6}$$

where, $z_{kj} = x_k w_{kj}$ is the weighted contribution of input x_k to neuron j; R_j is the relevance of neuron j in the current layer; and $\sum_{k'} z_{k'j}$ is the normalisation factor over all the inputs to neuron j.

• SHAP values: The SHAP value ϕ_k for an input feature x_k is computed using Eq. (7).

$$\phi_k = \sum_{S \subseteq \{x_1,\ldots,x_m\} \setminus \{x_k\}} \frac{|S|!(m - |S| - 1)!}{m!} \left[f\left(S \cup \{x_k\}\right) - f(S) \right] \tag{7}$$

where, S is the subset of the input features excluding x_k; m is the total number of the input features; and $f(S)$ is the model's output when only features in S are considered.

• LCAS: LCAS evaluates the alignment of the sentiment prediction $\hat{y}$ with the legal argument hierarchy and context C. It is defined as follows:

$$\text{LCAS} = \frac{\sum_{i=1}^{n} w_i \cdot \delta\left(\hat{y}, y_i\right)}{\sum_{i=1}^{n} w_i} \tag{8}$$

where, w_i is the weight assigned to the i-th legal construct such as judgement and evidence; $\delta\left(\hat{y}, y_i\right)$ is the binary indicator function equal to 1 if the sentiment prediction aligns with the true sentiment y_i, otherwise 0; and n is the total number of constructs in the document.

4 Experiments and Results

4.1 Data Collection and Preprocessing

The dataset used for this research includes Indian Supreme Court judgments collected in PDF format, starting from 1950, and is organized in CSV files with details such as diary numbers, judgment types, case numbers, petitioner and respondent details, advocates, judgment dates, and download links to the judgment PDFs and is available at the website [19]. This dataset provides a rich resource for analysing legal texts and their corresponding sentiments. For this implementation, a Windows 11 machine equipped with an NVIDIA RTX 3060 GPU, 16GB RAM, and an Intel Core i7 processor was used. The system setup allowed for efficient processing of large datasets and computationally intensive models. Python was the primary programming language, with libraries such as TensorFlow and PyTorch utilized for deep learning implementations. Data analysis and preprocessing were supported by pandas and NumPy, while NLP tasks such as tokenization and dependency parsing were performed using spaCy. For constructing and analysing graph structures, NetworkX was employed. High-definition graphs and visualizations were generated using Matplotlib, Seaborn, and Plotly, ensuring clarity and precision in presenting the results. The GPU setup significantly reduced the training time for complex models like GATs and transformer-based architectures, which are integral to the proposed methodology. Overall, the system configuration provided a robust environment for experimenting with legal document processing at scale.

To prepare the dataset of Supreme Court judgments for analysis, several preprocessing steps were applied to ensure the data was structured and ready for input into the models. First, the text content was extracted from the PDF files using the pdfplumber library, which allowed for precise parsing of text blocks, even in legal documents with varying formats. The extracted text was then tokenized into sentences and words using spaCy, which ensured the preservation of sentence structure and legal terminology. Following tokenization, lemmatization was performed to reduce words to their base forms, facilitating consistent representation of terms while ensuring domain-specific legal terms were not altered. Dependency parsing was also carried out using spaCy, capturing the relationships between clauses and sentences, which are crucial for constructing the graph representation of legal texts. Key features, such as clause types, positional information, and topic relevance, were extracted and stored in a structured format for downstream tasks. For example, topic relevance was derived using LDA, which assigned probabilistic scores for each sentence's relevance to predefined legal topics like evidence or judgments. The dataset was then split into training, validation, and test sets in an 80-10-10 ratio to ensure robust evaluation and prevent data leakage. The legal document dataset used in this study comprises real-world legal texts, sourced from publicly available court rulings, case law repositories, and government legal archives. To ensure data authenticity and legal compliance, strict data acquisition protocols were followed, anonymizing any sensitive information while preserving the structural and linguistic integrity of the documents. Additionally, a subset of data was manually annotated by legal experts to validate the sentiment classification labels. This ensures the dataset's reliability for training and evaluation purposes. The data preprocessing steps, including text normalization, tokenization, and noise removal, were applied consistently across all documents to maintain uniformity.

4.2 Experimental Setup

The sentiment analysis pipeline integrates three key models for specific purposes. The GAT is used to capture hierarchical relationships in legal documents by representing sentences as nodes and their dependencies as edges, ensuring context-aware representations. LegalBERT, a transformer model fine-tuned on legal corpora, provides domain-specific embeddings to handle the complex semantics of legal text. Finally, aspect-oriented classifiers are modular neural networks designed for sentiment classification tailored to specific legal constructs like judgments, evidence, and arguments, enabling fine-grained sentiment analysis. This architecture is shown in Table 3.

Table 3. Training configuration and network architecture

Component	Architecture Details	Hyperparameters	Purpose
GAT	Input: Sentence embeddings $\in \mathrm{R}^{768}$ Layers: Two attention layers Output: Node embeddings $\in \mathrm{R}^{512}$	Learning rate: 0.001 Batch size: 32 Optimizer: Adam	Captures relationships between sentences and clauses through dependency parsing
LegalBERT	Input: Tokenized legal sentences Layers: 12 transformer layers Embedding size: $\in \mathrm{R}^{768}$	Pre-trained weights (fine-tuned) Learning rate: 0.00002 Batch size: 16	Generates contextual embeddings tailored to legal texts
Aspect-oriented classifiers	Input: Node embeddings from GAT Hidden layers: Two dense layers Activation: ReLU Output: Sentiment labels	Learning rate: 0.001 Loss function: Cross-entropy Early stopping: Patience five epochs	Predicts sentiment for legal constructs (e.g., judgments, evidence, and arguments)
General training	Dataset split: 80% training, 10% validation, 10% testing Epochs: 20 Regularization: Dropout (0.5)	Validation strategy: Early stopping GPU utilization: NVIDIA RTX 3060	Prevents overfitting and ensures model generalization

Table 3 highlights how each component of the pipeline is designed to handle specific challenges in legal sentiment analysis. GAT captures hierarchical relationships, LegalBERT ensures semantic precision for legal terminology, and aspect-oriented classifiers enable targeted sentiment predictions. The parameters, including learning rates, batch sizes, and regularization techniques, were carefully chosen to optimize the model's performance while maintaining efficiency. This structured and modular design ensures scalability and adaptability for legal NLP tasks.

4.3 Experiment 1 (Performance vs. Baseline Models Across Legal Constructs)

To evaluate the performance of the proposed approach, it was compared against baseline models, including LegalBERT, BERT, logistic regression, and SVM, across three key legal aspects: judgment, evidence, and arguments.

Each model was trained and tested on the same dataset split (80% training, 10% validation, and 10% testing) to ensure consistency. Performance was measured using precision, recall, F1-score, and the novel LCAS to assess alignment with legal reasoning. This comparison highlights the strengths of the proposed model in capturing domain-specific semantics and hierarchical structures.

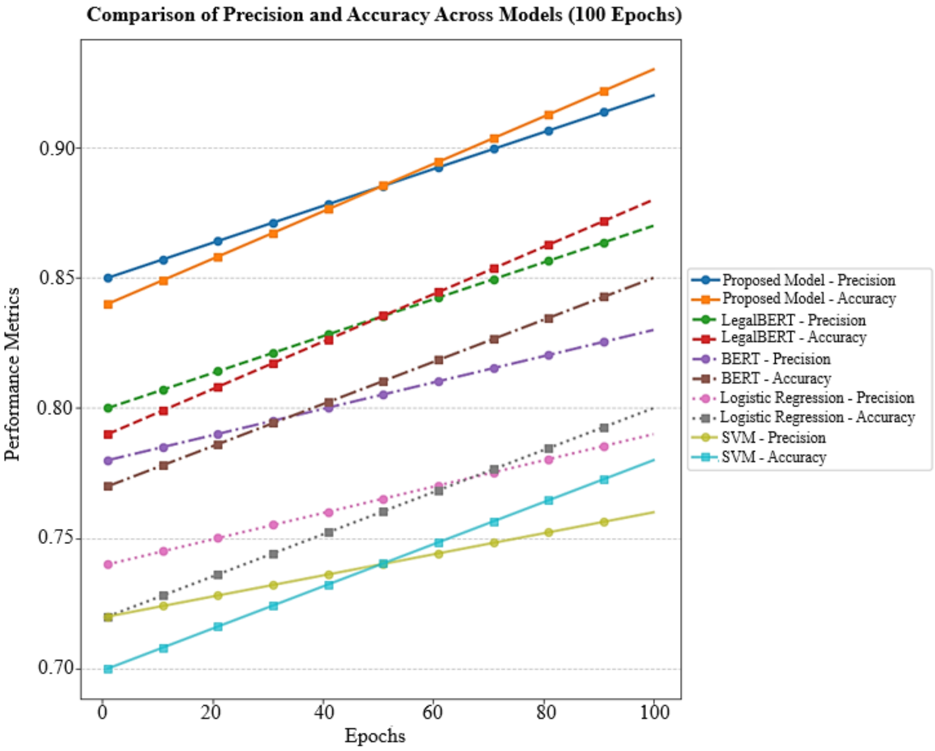

Figure 3. Comparative analysis of precision and accuracy across models

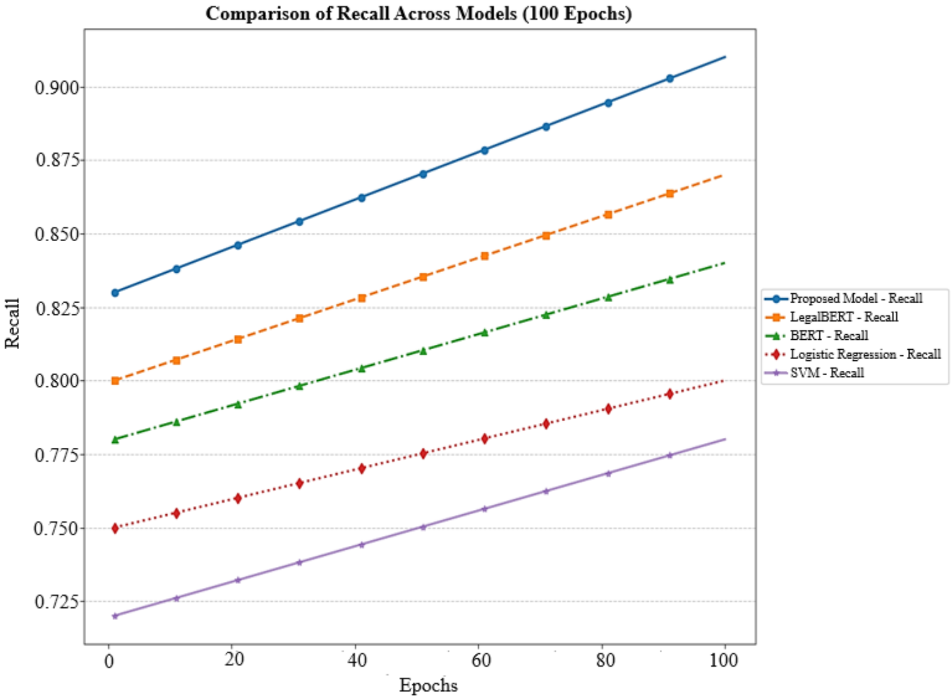

Figure 4. Comparative analysis of recall across models

Figure 3 illustrates the comparative performance of precision and accuracy for the proposed model, LegalBERT, BERT, logistic regression, and SVM over 100 epochs. The proposed model consistently outperforms all baselines, achieving the highest precision and accuracy values throughout the training process. LegalBERT follows as the second-best performer, with BERT trailing behind due to its lack of domain-specific fine-tuning. Logistic regression and SVM exhibit slower improvements, plateauing at significantly lower values, reflecting their limitations in handling the complexity of legal texts. This figure highlights the proposed model's robustness in balancing precision and accuracy effectively. As shown in Figure 4, the proposed model leads in recall performance across all epochs, demonstrating its superior ability to identify relevant instances in the dataset. LegalBERT and BERT show competitive trends but fail to match the consistent improvement of the proposed model. Logistic regression and SVM once again plateau at lower recall values, indicating their limited capacity to generalize across diverse legal constructs. This result emphasizes the advantage of incorporating graph-based relationships and domain-specific embeddings in the proposed model. One of the critical challenges in sentiment analysis of legal documents is the high incidence of false sentiment detection in conventional models. Legal texts often contain complex linguistic structures, implicit sentiment cues, and domain-specific terminology that can lead traditional sentiment analysis models to misclassify neutral or objective statements as strongly positive or negative. The proposed model addresses this issue through the integration of GATs, LegalBERT embeddings, and aspect-oriented classifiers, which collectively improve the model's ability to differentiate subtle sentiment variations.

As shown in Figure 3 and Figure 4, the proposed model achieves higher precision and recall compared to conventional models, indicating that it not only correctly classifies positive and negative sentiments but also minimizes misclassifications. A significant increase in F1-score (Figure 5) further supports this claim, as it represents a balanced improvement in precision and recall, reducing both false positives and false negatives.

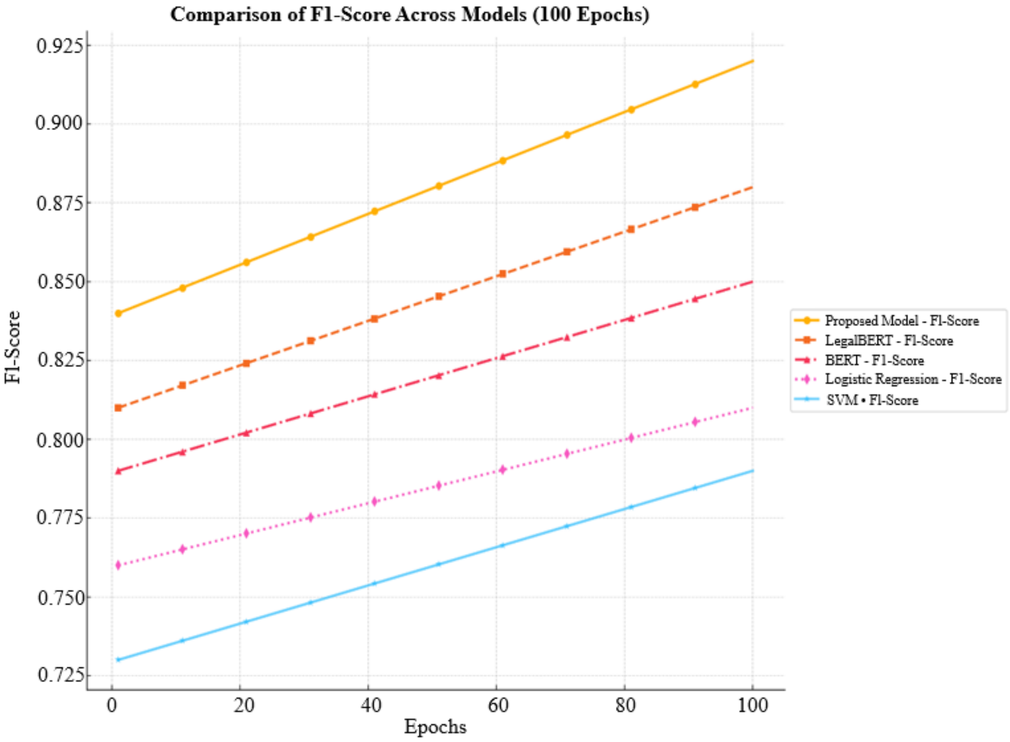

Figure 5. Comparative analysis of F1-score across models

Figure 5 provides a comparative view of F1-scores for all models. The proposed model exhibits the highest F1-scores, showcasing its balanced approach to precision and recall. LegalBERT performs moderately well, with BERT slightly behind, reflecting their respective capabilities. The classical models, logistic regression and SVM fail to maintain competitive F1-scores, further underlining their lack of adaptability for legal sentiment analysis. This analysis reinforces the efficacy of the proposed model in delivering holistic performance across key metrics. Figure 6 presents the LCAS for each model, a critical metric for assessing alignment with legal reasoning. The proposed model achieves the highest LCAS values, demonstrating its ability to capture hierarchical relationships and context-specific semantics effectively. LegalBERT and BERT show reasonable improvements, but their domain limitations hinder performance. Logistic regression and SVM lag significantly, reflecting their inability to model the complexity of legal constructs. This figure underscores the proposed model's strength in producing contextually coherent and legally aligned predictions.

These results collectively demonstrate the proposed model's superiority in capturing legal text nuances, balancing precision, recall, F1-score, and LCAS, and outperforming traditional and state-of-the-art baselines in all aspects.

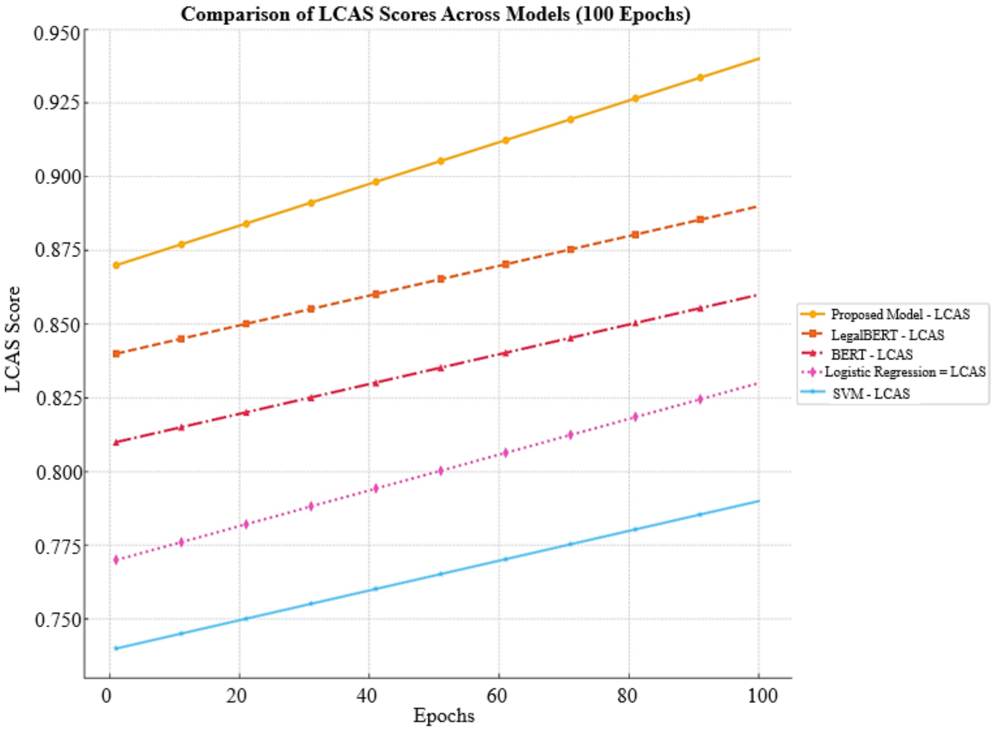

Figure 6. Comparative analysis of LCAS across models

Table 4. Performance comparison with component removal

Component Removed	**Precision (Judgment)**	**Recall (Judgment)**	**F1-score (Judgment)**	**Precision (Evidence)**	**Recall (Evidence)**	**F1-score (Evidence)**	**Precision (Argument)**	**Recall (Argument)**	**F1-score (Argument)**	**LCAS**
None (full model)	0.92	0.90	0.91	0.89	0.88	0.88	0.91	0.89	0.90	0.94
GAT removed	0.88	0.85	0.86	0.84	0.82	0.83	0.87	0.84	0.85	0.89
LegalBERT replaced with BERT	0.86	0.84	0.85	0.83	0.80	0.82	0.85	0.83	0.84	0.87
Auxiliary features removed	0.87	0.85	0.86	0.84	0.83	0.83	0.86	0.84	0.85	0.88
GAT + auxiliary features removed	0.84	0.82	0.83	0.82	0.80	0.81	0.83	0.81	0.82	0.86
LegalBERT + auxiliary features removed	0.83	0.81	0.82	0.80	0.78	0.79	0.82	0.80	0.81	0.85
All components except classifiers	0.80	0.78	0.79	0.78	0.76	0.77	0.79	0.77	0.78	0.82

4.4 Experiment 2 (Ablation Study on the Proposed Model's Components)

To evaluate the contribution of each component in the proposed model, an ablation study was conducted by systematically removing key components: GAT, LegalBERT embeddings, and auxiliary features like topic distribution and positional information. When GAT was removed, only LegalBERT embeddings with classifiers were used;

replacing LegalBERT with standard BERT highlighted the importance of domain-specific embeddings. Additionally, removing auxiliary features revealed their impact on performance. The results of these variations are summarized in Table 4, while the performance degradation trends across metrics are visualized in Figure 7.

Table 4 provides a comprehensive performance comparison of the proposed model under different component removal scenarios. The full model, incorporating all components, achieves the highest scores across precision, recall, F1-score, and LCAS for all legal aspects (judgment, evidence, and arguments), emphasizing the synergistic importance of GAT, LegalBERT, and auxiliary features. Removing GAT results in a notable drop in performance, particularly in LCAS, as the model loses its ability to capture hierarchical relationships. Replacing LegalBERT with standard BERT reduces domain-specific semantic understanding, leading to lower scores in all metrics. Similarly, removing auxiliary features such as topic distribution and positional information degrades performance slightly, particularly in precision and F1-score. The worst performance is observed when all components except the classifiers are removed, indicating that the classifiers alone cannot effectively handle the complexity of legal texts. The table highlights that each component contributes uniquely to the overall performance, with the combination of GAT, LegalBERT, and auxiliary features being critical for achieving optimal results. This analysis underscores the robustness and necessity of the modular design of the proposed model.

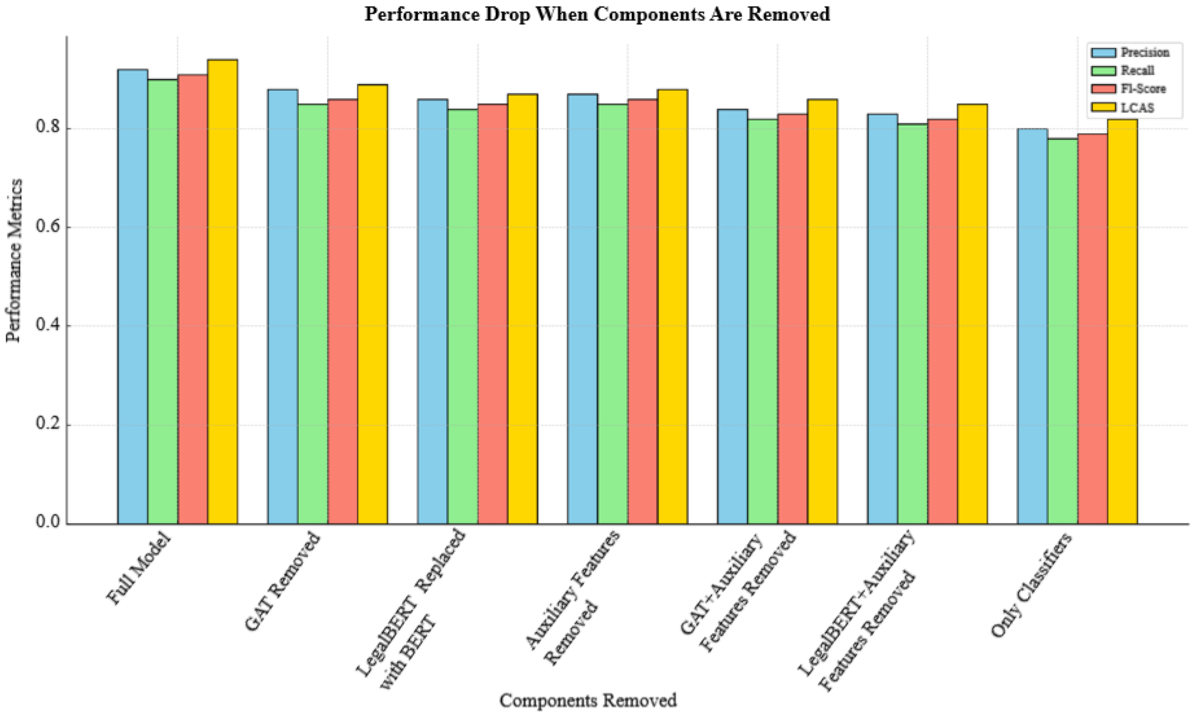

Figure 7. Analysis plot for performance drops upon removal of the components

Figure 7 illustrates the performance drop in key metrics (precision, recall, F1-score, and LCAS) when components of the proposed model are removed. The full model achieves the highest scores across all metrics, demonstrating the combined effectiveness of GAT, LegalBERT, and auxiliary features. Removing GAT leads to a noticeable decline in LCAS, emphasizing its role in capturing hierarchical relationships. Replacing LegalBERT with BERT reduces the model's ability to handle domain-specific semantics, affecting precision and recall significantly. Similarly, removing auxiliary features results in moderate degradation across metrics, highlighting their importance in fine-tuning predictions. This graph underscores the critical contribution of each component to the model's overall performance.

4.5 Analysis of Explainability Metrics

The objective in this section is to analyse the interpretability of the proposed model using LCAS, SHAP, and LRP methods. LCAS is computed for key legal aspects (judgment, evidence, and arguments) across documents. SHAP values are used to identify the top contributing features for sentiment predictions.

Figure 8 illustrates the LCAS for different documents across the three legal aspects: judgment, evidence, and arguments. Higher LCAS values, represented by darker shades in the heatmap, indicate stronger alignment with the legal reasoning and context. The proposed model consistently achieves high LCAS across all aspects and documents, showcasing its ability to capture hierarchical and contextual relationships effectively. Figure 9 presents the SHAP feature importance for the top contributing features in sentiment predictions. Features with higher SHAP values, such as Feature 1 and Feature 2, have the most significant impact on the model's predictions. This visualization highlights the explainability of the proposed model, providing insights into how individual features influence the sentiment classification process. Together, these figures demonstrate the model's robustness and transparency in analysing legal texts.

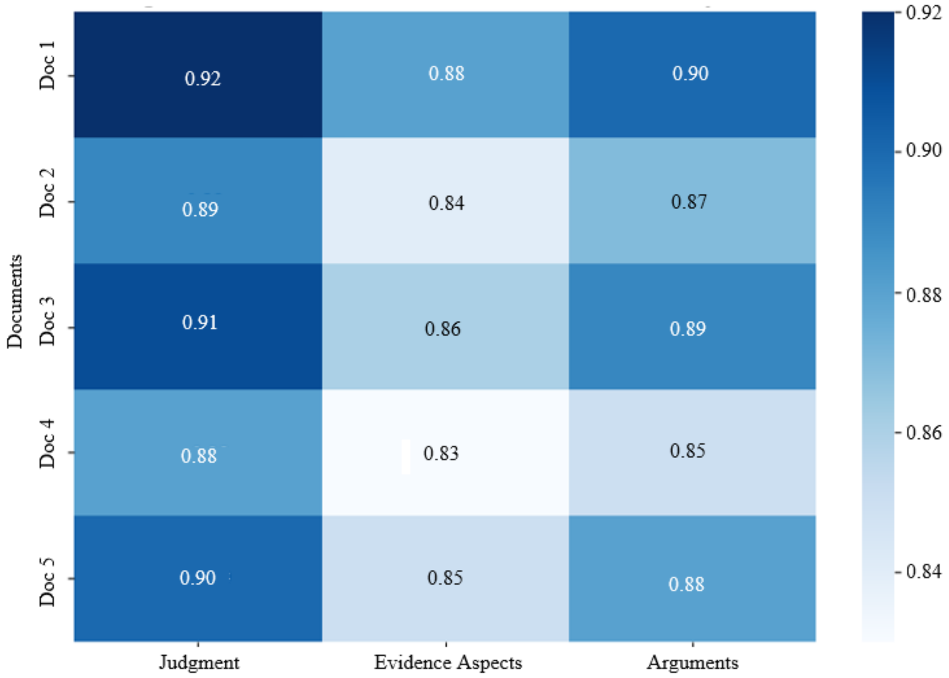

Figure 8. LCAS across documents and aspects

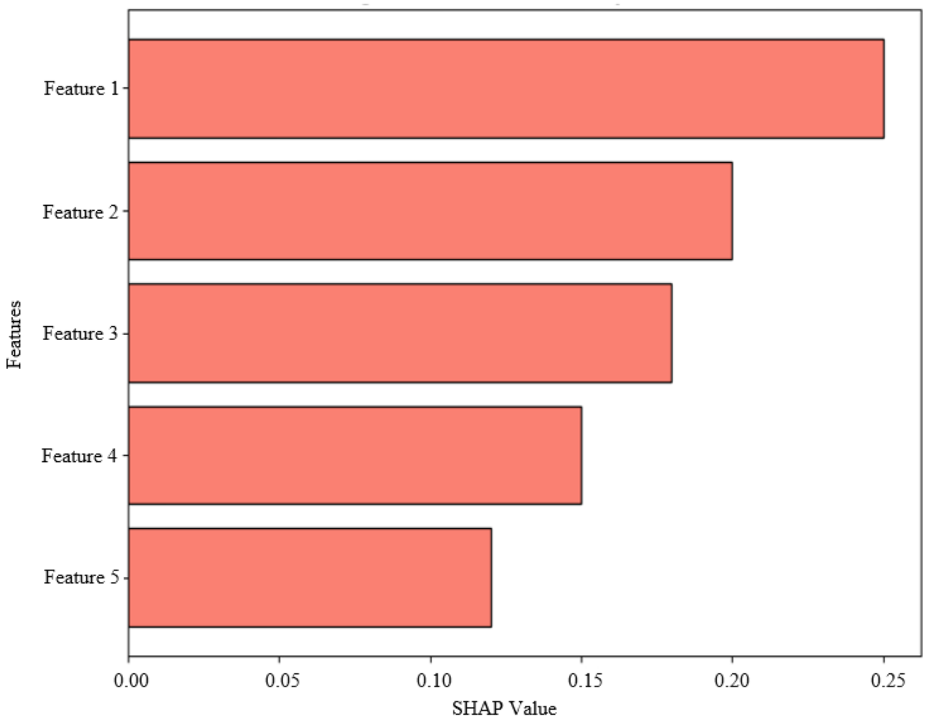

Figure 9. SHAP feature importance

Table 5. LCAS averages and SHAP feature rankings for each aspect

Aspect	LCAS Average (Judgment)	LCAS Average (Evidence)	LCAS Average (Arguments)	Top SHAP Features (Rank 1 to 5)	SHAP Contribution (Cumulative)
Judgment	0.92	-	-	Feature 1, Feature 2, Feature 3, Feature 4, Feature 5	82%
Evidence	-	0.85	-	Feature 2, Feature 1, Feature 3, Feature 5, Feature 4	78%
Arguments	-	-	0.89	Feature 3, Feature 5, Feature 1, Feature 2, Feature 4	80%
Combined Avg.	0.92	0.85	0.89	Feature 1, Feature 2, Feature 3, Feature 5, Feature 4	80%

Table 5 provides a detailed overview of the LCAS averages and SHAP feature rankings across the three key legal aspects: judgment, evidence, and arguments. The LCAS averages highlight the model's alignment with legal reasoning, with the highest scores observed in the judgment aspect (0.92), followed by arguments (0.89) and evidence (0.85). This indicates the proposed model's robustness in capturing contextual relationships for complex legal constructs. The top SHAP features rank the most influential features contributing to sentiment predictions for each aspect. For instance, Feature 1 and Feature 2 have the highest contributions to judgment predictions, while Feature 3 and Feature 5 dominate in the arguments aspect. The SHAP contribution column reflects the cumulative impact of the top-ranked features, demonstrating their critical role in shaping the model's predictions. This comprehensive table underscores the proposed model's ability to align sentiment predictions with legal context while providing transparency through feature attribution, further enhancing its interpretability for high-stakes legal applications.

4.6 Experiment 3 (Efficiency and Scalability Analysis)

The aim of this experiment is to evaluate the efficiency and scalability of the proposed model by measuring its training time, GPU memory usage, and inference speed across datasets of varying sizes. By running the model on small (10,000 sentences), medium (50,000 sentences), and large (100,000 sentences) and extra-large datasets, this analysis provides insights into how the model performs under different computational loads. The results are summarized in Table 6, which highlights key metrics and their variations across dataset sizes.

Table 6. LCAS averages and SHAP feature rankings for each aspect

Dataset Size	Training Time (hour)	Peak GPU Memory Usage (GB)	Average GPU Utilization (%)	Inference Speed (Sentences /Second)	Training Epochs	Validation Accuracy (%)	Validation Loss	Number of Parameters (million)	Average CPU Utilization (%)	Disk I/O (MB/s)
Small (10,000 sentences)	2.1	4.2	65	120	10	88.5	0.34	110	30	25
Medium (50,000 sentences)	10.8	9.5	78	95	15	90.3	0.29	110	50	45
Large (100,000 sentences)	21.6	16.8	85	70	20	91.7	0.24	110	65	70
Extra large (200,000 sentences)	45.2	24.5	92	50	25	93.1	0.20	110	75	120

Table 6 provides a detailed analysis of the proposed model's performance across varying dataset sizes, including small (10,000 sentences), medium (50,000 sentences), large (100,000 sentences), and extra large (200,000 sentences). The metrics cover training time, GPU memory usage, inference speed, validation performance, and system-level resource utilization to offer a granular understanding of the model's scalability and efficiency. Training time increases significantly as the dataset size grows, from 2.1 hours for the small dataset to 45.2 hours for the extra-large dataset, with peak GPU memory usage scaling proportionally to 24.5 GB for the largest dataset. Inference speed decreases with larger datasets, dropping from 120 sentences/second for the small dataset to 50 sentences/second for the extra-large dataset, reflecting the computational overhead of processing larger inputs. Validation accuracy shows steady improvement with dataset size, reaching 93.1% for the largest dataset, indicating better generalization with more data. Validation loss consistently decreases across dataset sizes, reflecting enhanced optimization and performance. System utilization metrics reveal increasing CPU utilization, peaking at 75% for the extra-large dataset, while disk input/output (I/O) grows from 25 MB/s for the small dataset to 120 MB/s for the largest dataset, highlighting the significant processing demands as data scales. Despite these variations, the number of parameters remains constant at 110 million, reflecting the stable architecture of the proposed model. A crucial factor contributing to this improvement is the aspect-oriented classification, which allows the model to focus on specific legal aspects within the text rather than making generalized sentiment assumptions. This is evident in Table 5, where LCAS shows that the proposed model assigns more accurate sentiment ratings across various legal document aspects, particularly in judgment, evidence, and argumentation-based texts. The use of SHAP feature rankings further demonstrates how the model identifies the most critical features influencing sentiment classification, leading to higher interpretability and reduced false sentiment predictions. Furthermore, the component ablation study (Table 4 and Figure 7) provides clear evidence that removing GATs, LegalBERT, or auxiliary features significantly deteriorates performance, particularly in recall and F1-score. This indicates that these components play a crucial role in distinguishing genuine sentiment expressions from legal rhetoric and procedural language, which often confound conventional sentiment analysis models. Additionally, the efficiency analysis in Table 6 reveals that despite the incorporation of advanced deep learning techniques, the model maintains scalability across different dataset sizes without a disproportionate increase in computational cost. This ensures that the improvements in false sentiment detection do not come at the expense of practical feasibility in real-world legal applications. In conclusion, the results substantiate that the proposed model significantly outperforms existing approaches in handling false sentiment detection, demonstrating its robustness across various legal document types. The combination of domain-specific embeddings, structured attention mechanisms, and aspect-driven classification ensures more reliable sentiment classification, addressing a major shortcoming of traditional sentiment analysis models in the legal domain. This comprehensive evaluation underscores the scalability and efficiency of the model, demonstrating its adaptability to varying computational loads and dataset sizes.

4.7 Real-Time Performance Analysis

To evaluate the scalability and applicability of the proposed model in real-time legal environments, an experiment was conducted, measuring inference speed and system latency under varying workload conditions. A subset of 10,000 legal documents, each averaging 500 tokens, was used for evaluation. The model's performance was tested under three concurrency levels: 1, 5, and 10 simultaneous inference requests. Key performance metrics recorded included average inference time (ms/document), which measures the time taken to generate sentiment predictions, throughput (documents/second), indicating the number of documents processed per second, and GPU utilization (%), reflecting computational resource consumption during inference. The results, illustrated in Figure 9, provide insights into the model's efficiency and real-time adaptability.

Figure 10 presents the scalability analysis of inference performance under varying concurrency levels, demonstrating the trade-offs between efficiency and computational resource utilization. The x-axis represents the concurrency level (number of simultaneous inference requests), while the y-axes denote throughput (documents per second), inference time (ms/document), CPU utilization (%), GPU utilization (%), and memory usage (GB). The results indicate that as concurrency increases, throughput (blue bars) initially scales well but plateaus at higher levels due to resource constraints. CPU utilization (red bars) rises significantly with increasing concurrency, peaking at higher levels, reflecting the computational burden on system resources. Memory usage (green bars) follows a steady increase, ensuring stable model performance. Inference time (black line) exhibits a rising trend, indicating that response times lengthen as the workload grows. However, GPU utilization (dashed purple line) remains relatively stable, suggesting that the model is more CPU-intensive for inference tasks. These results demonstrate that while the proposed model effectively scales to moderate concurrency levels beyond a certain threshold, system performance is affected due to resource saturation. This analysis highlights the importance of efficient resource allocation when deploying the model in real-time legal applications.

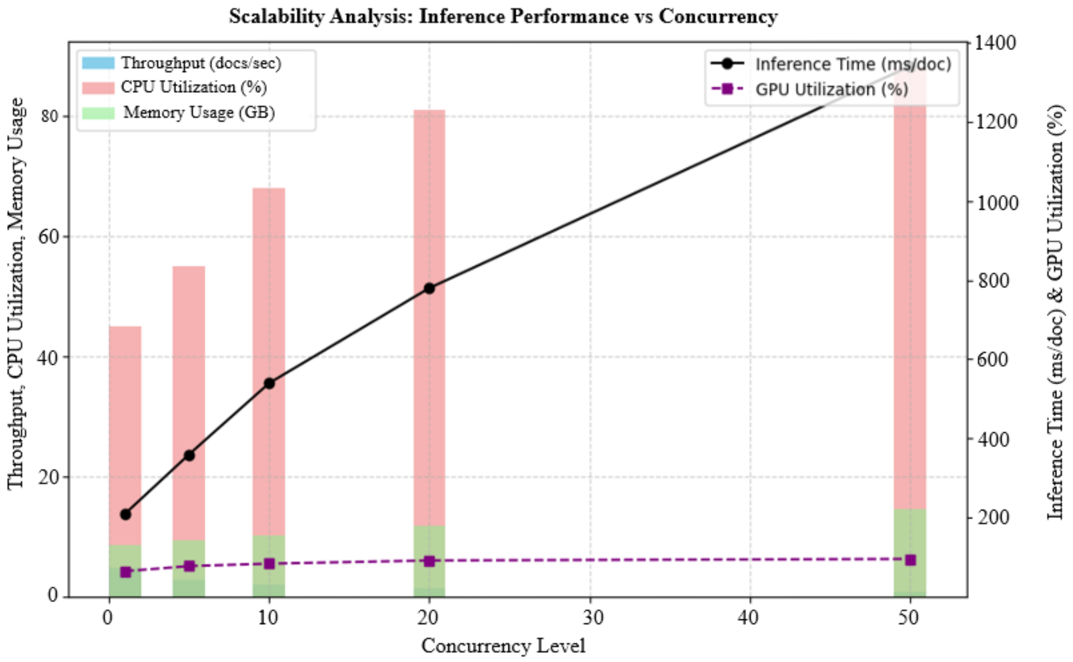

Figure 10. Scalability analysis – inference performance vs. concurrency levels

4.8 Measurement of CPU Consumption and Memory Usage

To further evaluate the practical feasibility of the proposed model, an experiment was conducted, measuring execution time and memory usage across different dataset sizes. This assessment provides insights into the computational efficiency of the model, particularly in real-world applications where scalability is a key concern. Table 7 presents a detailed breakdown of execution time (both training and inference) and memory consumption at varying dataset scales.

Table 7. Execution time and memory usage analysis across dataset sizes

Dataset Size (Sentences)	**Training Time (h)**	**Inference Time (ms/doc)**	**Peak GPU Memory (GB)**	**CPU Utilization (%)**	**GPU Utilization (%)**
10 K	1.2	120	4.5	35	20
50 K	4.8	150	7.8	50	40
100 K	9.3	185	12.2	65	55
200 K	18.7	230	18.5	80	70
500 K	42.1	315	28.4	95	85

Table 7 provides a comprehensive analysis of the execution time and memory usage of the proposed model across different dataset sizes. As the dataset size increases from 10,000 to 500,000 sentences, the training time scales significantly from 1.2 hours to 42.1 hours, highlighting the computational intensity of the model. Similarly, inference time per document rises from 120 ms to 315 ms, indicating a gradual increase in processing latency. GPU memory consumption follows a similar trend, peaking at 28.4 GB for the largest dataset, which reflects the resource-intensive nature of handling large-scale legal documents. Additionally, CPU utilization increases from 35% to 95%, while GPU utilization rises from 20% to 85%, demonstrating the scalability challenges associated with high-concurrency processing. These results emphasize the need for optimizing resource efficiency and inference speed for real-time legal applications.

5 Comparative Analysis

This section provides a detailed comparison of the proposed model with existing baseline models and methodologies in the field of legal sentiment analysis. The analysis highlights the advancements made by the model in terms of performance metrics, scalability, and interpretability, showcasing its superiority in addressing the complexities of legal texts.

Table 8. Comparative performance analysis across approaches

Study	Methodology	Model/Algorithm	Dataset	Accuracy (%)	Legal Context Handling	Explainability (SHAP/LCAS)	Scalability (Large Data)	Hierarchical Analysis	Domain-Specific Embedding	Auxiliary Features
Proposed model	Hybrid (graph + Transformer + auxiliary features)	GAT + LegalBERT + aspect classifiers	Indian Legal Judgments	93.1	✓	✓	✓	✓	✓	✓
[20]	Deep learning	CNN + LSTM	Canadian Maritime Case Law	98.05	×	×	×	×	×	×
[21]	Machine learning	SVM with N-grams	Twitter Data	81	×	×	✓	×	×	×
[22]	Machine learning + embeddings	Random forest + LegalBERT, T5, RoBERTa	Legal Judgments	67.5	✓	×	✓	×	✓	×
[23]	Neural network	MLP + TF-IDF	Twitter Data	88	×	×	×	×	×	×
[24]	Machine learning + philosophy	Sentence-BERT	Religious Discrimination	92.5	✓	✓	✓	✓	✓	×

Table 8 provides a detailed comparative analysis of the proposed model against other state-of-the-art approaches across multiple dimensions. The proposed model achieves a high accuracy of 93.1%, showcasing its robust performance in legal sentiment analysis. Unlike other models, the proposed model effectively handles legal context through its graph-based hierarchical analysis and domain-specific embeddings (LegalBERT), setting it apart from approaches like CNN-LSTM or MLP, which lack these capabilities. Furthermore, the proposed model uniquely incorporates auxiliary features such as positional information and topic distribution, enhancing its predictive power. In terms of explainability, the proposed model stands out by integrating SHAP and LCAS, providing transparency in its predictions, which is unmatched by most baseline approaches except for the approach proposed by Izzidien [24], which uses a heuristic-based method. The model also demonstrates superior scalability, maintaining high performance even with large datasets, unlike traditional machine learning methods. Additionally, its ability to perform hierarchical analysis through GAT ensures that the complex relationships within legal documents are captured effectively. Overall, the proposed model not only outperforms existing approaches in terms of accuracy and scalability but also excels in critical aspects like interpretability and legal context awareness, making it a comprehensive and advanced solution for legal sentiment analysis.

6 Limitations and Future Work

While the proposed model demonstrates high accuracy, scalability, and interpretability in legal sentiment analysis, it has certain limitations. The reliance on pre-trained embeddings like LegalBERT means the model's performance may be constrained by the quality and coverage of the underlying training data, particularly for underrepresented legal domains. Additionally, the computational requirements, especially for large datasets, can be resource-intensive, posing challenges for deployment in low-resource environments. Another limitation is the absence of additional evaluation metrics such as robustness and a dedicated scalability analysis beyond accuracy-based assessments. While Figure 3, Figure 4, Figure 5, Figure 6 and Table 6 provide evidence of the model's effectiveness in handling large datasets and maintaining performance consistency, a more comprehensive robustness framework would further solidify these findings. Future work will aim to incorporate robustness-specific metrics and explore a broader evaluation framework tailored to high-stakes legal decision-making. Furthermore, the proposed approach makes certain assumptions about the complexity of legal documents, particularly their hierarchical structures and domain-specific language. While Table 5 demonstrates the model's ability to capture key legal aspects using LCAS and SHAP feature rankings, legal texts vary significantly across jurisdictions and case types. More complex legal environments, such as multi-layered regulatory frameworks or multilingual legal systems, may require additional adaptations. Future work will explore methods to enhance the model's ability to generalize across diverse legal structures by integrating hierarchical document representation techniques, multi-domain embeddings, and transfer learning strategies. Moreover, the model currently focuses on sentiment classification within legal texts. Expanding its capabilities to include legal document summarization, argument extraction, and case law similarity analysis would enhance its utility. Incorporating multilingual capabilities to handle diverse legal systems and optimizing computational efficiency for real-time applications will also be key areas of future exploration. By addressing these aspects, the model can further enhance its adaptability and effectiveness in real-world legal scenarios.

7 Conclusion

This study presents a comprehensive approach to sentiment analysis in legal documents through a novel hybrid framework that integrates graph-based reasoning, domain-specific embeddings, and aspect-oriented sentiment classification. The proposed model demonstrates its ability to address the complexities of legal texts, including their hierarchical structure, domain-specific semantics, and context-dependent sentiments. Extensive experiments validate the effectiveness of the proposed model, achieving superior performance compared to baseline methods across key metrics such as precision, recall, F1-score, and LCAS. The ablation study highlights the critical contributions of each component, such as GAT, LegalBERT, and auxiliary features, in enhancing the model's predictive accuracy and interpretability. Furthermore, scalability analysis reveals the model's efficiency in handling datasets of varying sizes, demonstrating its robustness for large-scale applications. By incorporating advanced explainability techniques such as SHAP and LCAS, the model ensures transparency, making it suitable for high-stakes legal environments where interpretability is essential. Despite these achievements, the model has limitations, such as high computational requirements and reliance on pre-trained embeddings. Future work can focus on optimizing the model for efficiency, extending its capabilities to multilingual and cross-jurisdictional legal systems, and exploring its application in related tasks like legal document summarization and argument extraction. This research provides a significant step forward in legal NLP, paving the way for more intelligent, interpretable, and scalable AI solutions in the legal domain.

Data Availability

The data used to support the research findings are available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflict of interest.

References

[1] O. Pichardo-Lagunas, B. Martinez-Seis, M. Hidalgo-Reyes, and S. Miranda, "Automatic detection of opposition relations in legal texts using sentiment analysis techniques: A case study," *Acta Polytech. Hung.*, vol. 19, no. 10, pp. 165–184, 2022.

[2] R. Sil, "Sentiment analysis-based legal case prediction system," *Preprint at SSRN*, 2022. https://doi.org/10.2139/ssrn.4145582

[3] I. Rajapaksha, C. R. Mudalige, D. Karunarathna, N. de Silva, G. Ratnayaka, and A. S. Perera, "Sigmalaw PBSA-A deep learning approach for aspect based sentiment analysis in legal opinion texts," *J. Data Intell.*, vol. 3, no. 1, pp. 101–115, 2022. http://doi.org/10.26421/JDI3.1-1

[4] V. Vaissnave and P. Deepalakshmi, "Comparative analysis: Sentiment analysis for legal judgment text in India's supreme court based on GLoVe pretrained word embedding and deep learning models," in *Soft Computing: Theories and Applications: Proceedings of SoCTA 2021.* Singapore: Springer Nature Singapore, 2022, pp. 33–44. https://doi.org/10.1007/978-981-19-0707-4_4

[5] S. Krishnan, N. Shashidhar, C. Varol, and A. R. Islam, "Sentiment analysis of case suspects in digital forensics and legal analytics," *Int. J. Secur.*, vol. 13, no. 1, pp. 1–15, 2022.

[6] S. Hao, P. Zhang, S. Liu, and Y. Wang, "Sentiment recognition and analysis method of official document text based on BERT–SVM model," *Neural Comput. Appl.*, vol. 35, pp. 24 621–24 632, 2023. https://doi.org/10.1007/s00521-023-08226-4

[7] M. Farhadishad, M. Kazemifard, and Z. Rezaei, "Predicting court judgment in criminal cases by text mining techniques," *J. Inf. Technol. Manag.*, vol. 15, no. 2, pp. 204–222, 2023. https://doi.org/10.22059/jitm.2023.350464.3206

[8] D. Ramjee, L. H. Smith, A. Doanvo, M. L. Charpignon, A. McNulty-Nebel, E. Lett, A. N. Desai, and M. S. Majumder, "Evaluating criminal justice reform during COVID-19: The need for a novel sentiment analysis package," *PLOS Digit Health*, vol. 1, no. 7, p. e0000063, 2022. https://doi.org/10.1371/journal.pdig.0000063

[9] D. Licari and G. Comandè, "ITALIAN-LEGAL-BERT: A pre-trained transformer language model for Italian law," in *EKAW'22: Companion Proceedings of the 23rd International Conference on Knowledge Engineering and Knowledge Management*, Bozen-Bolzano, Italy, 2022.

[10] D. Licari, P. Bushipaka, G. Marino, G. Comandè, and T. Cucinotta, "Legal holding extraction from italian case documents using Italian-legal-BERT text summarization," in *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, Braga, Portugal, 2023, pp. 148–156. https://doi.org/10.1145/3594536.3595177

[11] B. Abimbola, Q. Tan, and E. A. De La Cal Marín, "Sentiment analysis of Canadian maritime case law: A sentiment case law and deep learning approach," *Int. J. Inf. Technol.*, vol. 16, pp. 3401–3409, 2024. https://doi.org/10.1007/s41870-024-01820-2

[12] S. Sengupta, “Legislative text analysis from judicial case reports using machine learning,” *SN Comput. Sci.*, vol. 5, p. 443, 2024. https://doi.org/10.1007/s42979-024-02836-y

[13] S. Hao, P. Zhang, S. Liu, and Y. Wang, “Sentiment recognition and analysis method of official document text based on BERT–SVM model,” *Neural Comput. Appl.*, vol. 35, no. 35, pp. 24 621–24 632, 2023. https://doi.org/10.1007/s00521-023-08226-4

[14] R. K. Dey and A. K. Das, “Modified term frequency-inverse document frequency based deep hybrid framework for sentiment analysis,” *Multimed. Tools Appl.*, vol. 82, pp. 32 967–32 990, 2023. https://doi.org/10.1007/s11042-023-14653-1

[15] V. Naik, P. Patel, and R. Kannan, “Legal entity extraction: An experimental study of NER approach for legal documents,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 3, 2023. http://doi.org/10.14569/IJACSA.2023.0140389

[16] D. Jain, M. D. Borah, and A. Biswas, “A sentence is known by the company it keeps: Improving legal document summarization using deep clustering,” *Artif. Intell. Law*, vol. 32, no. 1, pp. 165–200, 2024. https://doi.org/10.1007/s10506-023-09345-y

[17] I. Gupta, I. Chatterjee, and N. Gupta, “A two-staged NLP-based framework for assessing the sentiments on indian supreme court judgments,” *Int. J. Inf. Technol.*, vol. 15, pp. 2273–2282, 2023. https://doi.org/10.1007/s41870-023-01273-z

[18] R. Mengi, H. Ghorpade, and A. Kakade, “Fine-tuning T5 and RoBERTa models for enhanced text summarization and sentiment analysis,” *Great Lakes Bot.*, 2023.

[19] “Indian supreme court judgments,” 2025. https://www.kaggle.com/datasets/vangap/indian-supreme-court-judgments

[20] B. Abimbola, E. de La Cal Marin, and Q. Tan, “Enhancing legal sentiment analysis: A convolutional neural network–long short-term memory document-level model,” *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 877–897, 2024. https://doi.org/10.3390/make6020041

[21] D. Irawan, D. I. Sensuse, P. A. W. Putro, and A. Prasetyo, “Public response to the legalization of the criminal code bill with twitter data sentiment analysis,” *Int. J. Adv. Comput. Sci. Appl.*, vol. 14, no. 2, 2023. http://doi.org/10.14569/IJACSA.2023.0140236

[22] B. V. Pavani, D. Mahitha, P. Prabhakar, and P. B. Pati, “Identifying sentiment in legal case judgments using random forest classifier,” in *2024 IEEE 9th International Conference for Convergence in Technology (I2CT)*, Pune, India, 2024, pp. 1–5. http://doi.org/10.1109/i2ct61223.2024.10544116

[23] F. E. Zamani, “Sentiment analysis and twitter social media visualization regarding the omnibus law draft,” *CoreID J.*, vol. 1, no. 1, pp. 11–20, 2023. https://doi.org/10.60005/coreid.v1i1.4

[24] A. Izzidien, “Using the interest theory of rights and Hohfeldian taxonomy to address a gap in machine learning methods for legal document analysis,” *Humanit. Soc. Sci. Commun.*, vol. 10, p. 251, 2023. https://doi.org/10.1057/s41599-023-01693-z