Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

SPS-LIME: Role-Aware Segment Perturbation for Explaining Long Legal Documents

Meng-Luen Wu

158769@mail.tku.edu.tw

Tamkang University

Bo-Xun Huang

Tamkang University

Research Article

Keywords: Explainable Artificial Intelligence, Legal Text Classification, Perturbation-based Explanation, Rhetorical Roles, Judicial Decision Support

Posted Date: August 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10591506/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

SPS-LIME: Role-Aware Segment Perturbation for Explaining Long Legal Documents

Abstract. Legal text classifiers may achieve strong classification performance, but their practical value depends on whether their decisions can be explained through discourse units that correspond to recognizable forms of legal reasoning. Existing perturbation-based explanation methods commonly operate on individual tokens or sentences, which can fragment factual narratives, party arguments, and judicial reasoning that extend across multiple sentences. We introduce Semantic-Preserved Segment LIME, or SPS-LIME, a model-agnostic explanation framework that merges consecutive sentences assigned the same rhetorical role into variable-length semantic segments. These role-continuous segments serve as the units of perturbation and attribution, allowing explanations to better retain natural-language continuity and legal-semantic structure. SPS-LIME further localizes influential sentences within highly ranked segments through a coarse-to-fine explanation mechanism.

We evaluate SPS-LIME on LexGLUE-SCOTUS and a document-level binary violation-classification task derived from the ECtHR corpus. Compared with LIME variants that perturb individual tokens or sentences, SPS-LIME achieves the highest mean Consistency and Comprehensiveness scores on both datasets and the lowest Sufficiency score on ECtHR. These results indicate that rhetorical-role-based segments provide a more stable and semantically continuous explanation space for long legal documents.

The experiments also reveal that faithfulness to classifier behavior and legal relevance do not necessarily coincide. Influential segments identified by SPS-LIME may concentrate on factual or contextual passages rather than on rhetorical roles that explicitly express judicial reasoning or disposition. SPS-LIME therefore contributes both a discourse-aware explanation mechanism and a diagnostic framework for examining potential mismatches between classifier-sensitive content and role-based legal salience.

Keywords: Explainable Artificial Intelligence, Legal Text Classification, Perturbation-based Explanation, Rhetorical Roles, Judicial Decision Support.

1 Introduction

Legal text classification has become an important research topic in Legal Natural Language Processing (Legal NLP), with applications such as issue tagging, judgment classification, and case outcome analysis. This line of work has led to benchmark datasets and evaluation suites for legal text classification, including tasks on Supreme Court opinions and European Court of Human Rights cases (Chalkidis et al., 2019, 2022). Recent advances in transformer-based and long-document models have improved classification performance on benchmark datasets such as LexGLUE-SCOTUS and ECtHR Task A. Longformer and BigBird introduced sparse-attention mechanisms that enable

the processing of substantially longer sequences than standard BERT (Beltagy et al., 2020; Zaheer et al., 2020). These architectures have subsequently been evaluated on long legal-text benchmarks such as LexGLUE (Chalkidis et al., 2022). However, classification accuracy alone does not address the transparency requirements of legal decision-support settings, where users often expect not only an assigned class label but also an explanation of which parts of a judgment support that classification.

Post-hoc explanation methods are widely used to interpret text classifiers by identifying input features that are important to model outputs. Among them, perturbation-based methods such as LIME estimate importance scores by masking portions of the input and observing changes in classification probabilities. Ribeiro et al. (2016) introduced LIME as a model-agnostic local explanation method, while Lundberg and Lee (2017) proposed SHAP as a unified framework for additive feature attribution. DeYoung et al. (2020) subsequently evaluated textual rationales at token, span, and sentence levels. In practice, many such methods operate at the token level, assigning importance to words or subword units one at a time. While this provides fine-grained explanations, it also creates an important limitation in legal documents: highlighted words can be scattered across a judgment, detached from their surrounding context, and difficult to interpret as complete legal reasoning. Similar concerns have been identified in research on explanation granularity. Rychener et al. (2023) showed that word-level perturbation can generate out-of-distribution samples and create a high-dimensional local search space, motivating the use of larger explanatory units.

This limitation is illustrated conceptually in Fig. 1. At the word level, important clues may appear as isolated lexical fragments that do not form a coherent description of the decision. Sentence-level explanations improve readability, but a single sentence may still provide only a partial statement of an argument, a precedent, or a judicial holding. By contrast, grouping consecutive sentences that share the same rhetorical role makes it possible to explain classifications using semantically continuous segments rather than fragmented words or incomplete sentences. This observation motivates explanation units that preserve semantic continuity and reduce lexical fragmentation.

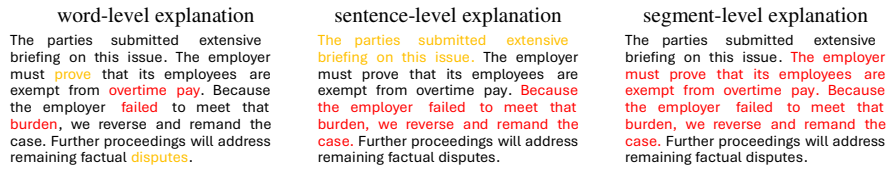


Fig. 1 Comparison of explanations at the word level, sentence level, and rhetorical segment level. Red intensity indicates the importance assigned to each explanatory unit. Explanations based on individual words often distribute salient evidence across isolated tokens, whereas explanations based on sentences may still capture only partial statements. By grouping consecutive sentences with the same rhetorical role into semantic segments, explanations based on rhetorical segments preserve discourse continuity and provide a more coherent basis for legal interpretation.

This concern is particularly relevant to judicial writing, where legal meaning is commonly organized through discourse structure rather than isolated lexical items. A judgment usually contains multiple functional components, including factual background, procedural history, party arguments, statutory references, precedents, lower-court

rulings, judicial analysis, and final holdings. Prior work in legal NLP has modelled these components as rhetorical or argumentative roles, annotating court decisions with discourse-level labels such as facts, arguments, and holdings (Saravanan et al., 2008). From a legal perspective, these components are not interchangeable. Some parts primarily describe background facts, while others articulate the reasoning that justifies the court's conclusion. As a result, an explanation that merely highlights model-sensitive words does not necessarily reveal whether the model relies on the portions of the document that are legally central.

To address this issue, this paper proposes Semantic-Preserved Segment LIME (SPS-LIME), a perturbation-based explanation framework designed for legal document classification. The method first assigns each sentence a rhetorical-role label, such as Fact, Argument, Statute, Precedent, Ratio of the Decision, Ruling by Lower Court, or Ruling by Present Court. Consecutive sentences with the same rhetorical role are then merged into variable-length semantic segments, which are used as the units of perturbation and explanation instead of isolated words or independent sentences. SPS-LIME further adopts a dual-layer mechanism, combining segment-level attribution with sentence-level localization, so that explanations can identify important discourse units while still pinpointing salient sentences within them. In this way, the method is designed to produce explanations with greater semantic continuity than conventional token-level heat maps.

In this study, we use a seven-role inventory adapted from later rhetorical-role annotation schemes: Ratio of the Decision, Ruling by Present Court, Argument, Precedent, Statute, Ruling by Lower Court, and Fact (Malik et al., 2022; Bambroo et al., 2025). For every sentence, our method assigns a single label from this set, and consecutive sentences sharing the same rhetorical role are eligible to be merged into semantic segments.

The proposed framework is evaluated on LexGLUE-SCOTUS and a document-level binary violation-classification task derived from the ECtHR corpus using three commonly used explanation criteria: Consistency, Comprehensiveness, and Sufficiency. Consistency measures whether repeated runs of an explainer highlight similar content. Comprehensiveness evaluates how much the model's classification changes when the selected important units are removed, while Sufficiency measures whether the selected units alone can preserve the original classification. Comprehensiveness and Sufficiency follow the deletion- and retention-based evaluation framework of DeYoung et al. (2020). Consistency is separately used in this study to measure explanation stability across repeated runs.

Motivated by legal XAI research emphasizing legally grounded and evidentially meaningful explanations (Collenette et al., 2023; Richmond et al., 2024), this paper further employs Legal Role Alignment at k (LRA@k) to examine whether highly ranked explanatory units occur within rhetorically central legal roles.

The experimental results show that segment-level perturbation achieves higher consistency on both datasets and competitive deletion-based faithfulness. SPS-LIME performs best across all three explanation metrics on ECtHR, whereas Sentence LIME obtains the lowest Sufficiency score on SCOTUS. These results support the idea that semantic segments formed by rhetorical-role continuity provide a more stable and

semantically continuous basis for perturbation-based explanation. At the same time, the experiments reveal two non-trivial phenomena that motivate a more careful discussion of legal explainability.

The first phenomenon concerns the use of rhetorical roles themselves. When discourse-role distributions are strongly associated with class labels, especially under class imbalance, classifiers may partially rely on the global distribution of rhetorical roles rather than on the semantic details of the underlying text. In such cases, role-based information can yield relatively high micro-F1 while macro-F1 remains substantially lower, suggesting that majority classes are being classified more effectively than minority classes. The ECtHR results are consistent with the possibility that rhetorical-role distributions function as classification shortcuts under class imbalance. Thus, while rhetorical roles help construct semantically coherent explanation units, they may also introduce a potential dataset-dependent shortcut or majority-class bias.

The second phenomenon concerns the distinction between classification importance and legal importance. In many cases, deleting the top-k segments identified by SPS-LIME substantially reduces classification confidence, indicating that those segments are indeed important for the classifier. However, qualitative analysis shows that the most influential segments are often dominated by factual narratives rather than by doctrinally central units such as Ratio of the Decision or Ruling by Present Court. Moreover, token-level deletion sometimes achieves higher Legal Role Alignment scores than sentence- or segment-level explanations, despite being less coherent semantically. This suggests that text most useful for classification is not always the same as text assigned higher legal-role weights under the proposed LRA@k formulation of the judgment.

These observations suggest that deletion-based metrics may not fully capture the distinction between model sensitivity and role-based legal relevance. This motivates an evaluation framework that considers faithfulness to classifier behavior, explanation stability, semantic continuity, and rhetorical-role alignment separately. SPS-LIME is therefore intended not only to generate segment-based explanations but also to help diagnose potential mismatches between classifier-sensitive content and role-based legal salience.

Accordingly, this paper makes three main contributions. First, we introduce a semantic-preserved segment-based explanation framework for legal documents, in which consecutive sentences with the same rhetorical role are merged into coherent perturbation units. Second, we combine Consistency, Comprehensiveness, Sufficiency, and Legal Role Alignment at k into a unified evaluation framework that assesses explanations from both classification and legal-role perspectives. Third, through experiments on SCOTUS and ECtHR, we examine how segment-level explanations compare with word- and sentence-level baselines and analyse trade-offs involving stability, faithfulness to classifier behavior, and legal-role alignment.

The remainder of this paper is organized as follows. Section 2 reviews perturbation-based explanation methods, explanation granularity, rhetorical-role analysis, and legal rationale alignment in legal NLP. Section 3 introduces SPS-LIME, including rhetorical-role tagging, semantic segment construction, segment-level perturbation, and dual-layer explanation. Section 4 presents experiments on SCOTUS and ECtHR, evaluating classification performance and explanation quality in terms of Consistency,

Comprehensiveness, Sufficiency, and Legal Role Alignment at k. Section 5 concludes the paper by summarizing the main findings, limitations, and directions for future research. Appendix A presents case studies of four judgments, including their top-k explanatory segments and corresponding heatmap visualizations.

2 Related Work

This section reviews four research areas relevant to SPS-LIME: perturbation-based explanations for legal text classification, explanation granularity, rhetorical-role modelling in legal documents, and the distinction between classification importance and legal relevance. Together, these studies motivate the use of variable-length rhetorical segments as perturbation units and the evaluation of explanations from both model-faithfulness and legal-structure perspectives.

2.1 Perturbation-Based Explanations for Legal Text Classification

Perturbation-based explanation methods were initially developed as general model-agnostic techniques. Ribeiro et al. (2016) introduced LIME, which explains an individual classification by fitting an interpretable surrogate model to perturbed samples located near the target instance. Lundberg and Lee (2017) proposed SHAP, which provides additive feature attributions based on a unified formulation of Shapley-value-based explanation methods. DeYoung et al. (2020) subsequently introduced the ERASER benchmark and formalized rationale evaluation through comprehensiveness and sufficiency. These metrics assess whether removing selected text reduces classification confidence and whether retaining only that text preserves the original classification.

The reliability of perturbation-based explanations depends on how the local neighborhood is constructed. Alvarez-Melis and Jaakkola (2018) demonstrated that local explanations may change substantially under small variations in the input. Atanasova et al. (2020) further showed that text explanation methods differ considerably in faithfulness, stability, and sensitivity to evaluation conditions. Rychener et al. (2023) showed that word-level perturbation can generate out-of-distribution text and a high-dimensional local search space, whereas sentence-level units can reduce these problems and improve local surrogate fidelity. Taken together, these studies indicate that token-level perturbation may produce unstable explanations or distributionally implausible samples.

This limitation is particularly relevant to long legal documents. Legal meaning is often distributed across multiple sentences and discourse components rather than concentrated in individual lexical features. Chalkidis et al. (2021) addressed this issue through paragraph-level rationale extraction for European Court of Human Rights cases. Their results showed that sparsity, continuity, comprehensiveness, and classification performance interact in complex ways when rationales are extracted from long judgments. In a related line of research, Mamakas (2022) investigated bag-of-words and BERT-based approaches to large legal document classification. This work is relevant to the representation of long legal texts, although it does not directly examine rationale extraction or explanation quality. Together, these studies illustrate distinct challenges in paragraph-level rationale extraction and the representation of large legal documents.

Several studies have examined explainability in legal classification and reasoning. Waltl et al. (2019) used local linear explanations to analyse the classification of semantic types in German legal norms. de Arriba-Pérez et al. (2022) developed an explainable multi-label classifier for Spanish legal judgments. Collenette et al. (2023) designed and evaluated an argumentation-based decision-support tool for reasoning about European Court of Human Rights cases, while Richmond et al. (2024) surveyed explainability approaches in AI and law. These studies demonstrate the broader relevance of explainability to legal applications, but they do not directly examine how perturbation units should be defined in structurally complex judgments.

Human-centred evaluation introduces an additional requirement. Rong et al. (2024) surveyed human-centred evaluations of XAI and organized existing user studies around trust, understanding, usability, and human-AI collaboration performance. Their findings motivate complementing formal fidelity metrics with evaluations of how explanations are understood and used by their intended users. In legal applications, this means that an explanation should not merely identify model-sensitive features. It should also present those features in units that legal readers can interpret as coherent factual, argumentative, or doctrinal content.

2.2 Explanation Granularity in Legal Text Models

The granularity of an explanation determines both the dimensionality of the perturbation space and the semantic completeness of the selected units. DeYoung et al. (2020) introduced the ERASER benchmark, which includes human-annotated rationales across multiple NLP tasks and evaluates both rationale alignment and faithfulness to classifier behavior. Rychener et al. (2023) more directly examined the effect of explanation granularity and showed that sentence-level perturbation can reduce search-space dimensionality and improve local fidelity compared with word-level sampling. Their findings do not imply that sentence-level perturbation is universally superior, but they show that finer units do not automatically produce more useful explanations.

As illustrated in Fig. 2, granularity also affects the size of the local perturbation space. For a document containing N_W words and N_S sentences, word-level masking produces up to 2^{N_W} possible configurations, whereas sentence-level masking produces 2^{N_S} configurations, where N_S is generally much smaller than N_W. Reducing the number of explanatory dimensions makes local sampling more tractable and decreases the probability of generating highly fragmented text. However, sentence-level perturbation may still divide an argument, factual narrative, or judicial holding across several independent units.

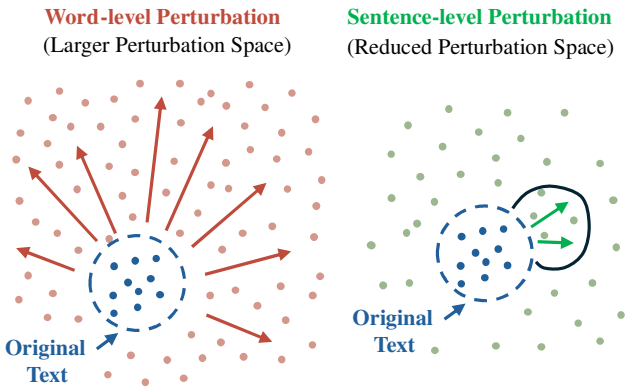


Fig. 2 Conceptual comparison of word-level and sentence-level perturbation spaces. Word-level masking produces a larger combinatorial neighborhood, whereas sentence-level masking reduces the search space and preserves larger semantic units.

Zafar and Wu (2026) examined related methodological issues in a systematic review of explainable fraud detection. Although their study does not concern legal text or textual granularity directly, it identifies preprocessing, class imbalance, local-neighborhood construction, and explanation-level evaluation as factors that may affect the reliability of post-hoc explanations. Their analysis supports the narrower methodological point that perturbation-based explanation should be evaluated in relation to the structure of the data and the conditions under which local samples are generated.

Legal NLP provides several alternatives to fixed token-level units. Malik et al. (2022) and Kalamkar et al. (2022) annotated legal sentences with rhetorical roles and used sentence-level labels to support automatic document structuring and semantic segmentation. Chalkidis et al. (2021), by contrast, used paragraphs as rationale units for ECtHR outcome classification to preserve broader evidential context. Aumiller et al. (2020) showed that paragraph and document structure provide useful signals for legal text segmentation. Wu and Liu (2026) further demonstrated that evidence relevant to long legal document classification may be distributed across multiple segments and therefore requires explicit cross-segment integration.

These studies reveal a limitation of fixed granularity. Tokens provide high resolution but often fragment legal meaning. Sentences improve readability but may capture only part of a multi-sentence legal function. Paragraphs preserve more context but may combine facts, arguments, authorities, and rulings within a single unit. The appropriate explanatory unit should therefore reflect the functional structure of the legal document rather than being determined only by fixed textual boundaries.

This position is also consistent with legal-informatics scholarship. Ashley (2017) treats computational legal analysis as dependent on representations that preserve legally meaningful concepts, relationships, and argument structures. Goodrich (1984) describes legal language as inherently rhetorical and emphasizes that legal meaning emerges from the organization of facts, authorities, and arguments within discourse. These perspectives support explanation units that correspond to legal-discourse functions rather than arbitrary token counts or paragraph boundaries.

2.3 Rhetorical Roles in Legal Document Classification

Computational analysis of rhetorical roles predates contemporary neural legal NLP. Hachey and Grover (2004) classified legal sentences using lexical, syntactic, positional, and structural features for legal summarization. Hachey and Grover (2005) subsequently incorporated sequential dependencies between adjacent sentences, recognizing that the rhetorical function of a sentence depends partly on its surrounding context. Saravanan and Ravindran (2010) formulated rhetorical-role identification as a sequence-labelling task and connected rhetorical structure to legal-document segmentation and summarization.

Later studies developed larger annotated corpora and neural architectures. Bhattacharya et al. (2019) investigated contextual and sequential representations for rhetorical-role labelling in Indian Supreme Court judgments. Kalamkar et al. (2022) constructed a corpus for automatic structuring of legal documents, while Malik et al. (2022) proposed semantic segmentation based on rhetorical roles and explicitly modelled transitions between adjacent legal sentences. Santosh et al. (2024) introduced hierarchical curriculum learning to account for both document-level difficulty and rhetorical-role-level difficulty. Nigam et al. (2025) expanded available resources through the LegalSeg corpus for rhetorical-role classification.

Journal studies have further developed this research direction. Bhattacharya et al. (2023) introduced DeepRhole and evaluated deep sequential models across multiple legal domains and courts. Bambroo et al. (2025) proposed MARRO, which uses multi-headed attention and a seven-role inventory consisting of Facts, Ruling by Lower Court, Argument, Ratio of the Decision, Statute, Precedent, and Ruling by the Present Court. This inventory closely corresponds to the rhetorical roles adopted in SPS-LIME.

Rhetorical roles have also been modelled as interacting components rather than independent sentence labels. Ahmed and Ahmad (2025) proposed the Rhetorical Role Relatedness framework to examine relationships among rhetorical roles in legal case-brief generation. Belfathi et al. (2026) combined local sentence context with global semantic prototypes for rhetorical-role labelling. These studies support modelling rhetorical roles with contextual and cross-role information rather than treating sentence labels as independent classifications.

Existing rhetorical-role research primarily uses role labels for summarization, segmentation, document structuring, or classification. SPS-LIME uses the same structural information for a different purpose. Consecutive sentences assigned to the same rhetorical role are merged into variable-length perturbation units. The rhetorical labels therefore define the boundaries of the explanation space rather than serving only as classification features or output annotations.

2.4 Shortcut Learning and Legal Rationale Misalignment

A faithful explanation identifies the features that actually influence a model, but those features are not necessarily substantively appropriate. Geirhos et al. (2020) described shortcut learning as the use of easily available statistical regularities that perform well on a benchmark without representing the intended reasoning process. Gururangan et al. (2018) showed that annotation artefacts can create such shortcuts in NLP datasets.

Jacovi and Goldberg (2020) distinguished faithful explanations from plausible explanations, noting that an explanation may accurately represent model behaviour while failing to provide a rationale that humans consider meaningful.

This distinction is important in legal NLP because model importance and legal importance may diverge. Staliūnaitė et al. (2024) compared explanation methods for legal outcome classification and found substantial variation in the rationales produced by different explainers. Valvoda and Cotterell (2024) introduced an influence-function-based method for identifying the precedents used by legal outcome classification models. Their comparison showed that the types of precedent relied on by the models differed from those used by human judges. Namazov et al. (2026) directly compared extracted rationales with legal-expert assessments and reported disagreement over which parts of ECtHR cases were relevant to outcome classification.

Valvoda et al. (2021) investigated a related question through an information-theoretic analysis of common-law precedents. They examined the outcome-related information contained in factual and argumentative components of precedent cases and showed that these components contribute differently across legal contexts. Their study does not evaluate perturbation-based explanations directly, but it supports the conclusion that textual components should not be treated as interchangeable sources of legal or classification information.

Legal-AI scholarship places this problem within a broader justificatory context. Atkinson et al. (2020) argue that explanations in AI and law should be connected to legal argumentation and justification rather than reduced to numerical feature attribution. Brożek et al. (2024) distinguish model opacity from unpredictability, unfamiliarity, and the inability to provide adequate justification. Kolt et al. (2026) propose legal alignment as a framework for examining whether AI systems conform to legal norms and legally relevant reasoning patterns. Richmond et al. (2024) similarly emphasize that legal explanation involves evidential and normative assumptions that are not captured by model fidelity alone.

These studies indicate that deletion-based metrics answer only one question: whether the selected text affects the classifier. They do not determine whether the selected text represents doctrinally central reasoning. Conversely, overlap with legally important rhetorical roles does not establish that those roles causally influence the classification. Legal explanation therefore requires separate evaluation of faithfulness to classifier behavior and legal-role relevance.

2.5 Research Gap

The reviewed literature establishes three relevant findings. First, perturbation-based methods provide model-agnostic explanations, but their stability and faithfulness depend on neighborhood construction and feature granularity. Second, legal judgments contain structured rhetorical components that can be identified through sentence-level sequence labelling. Third, model-sensitive text may differ from the text that legal experts regard as doctrinally important.

The explanation methods reviewed above commonly operate on fixed units such as tokens, sentences, or paragraphs (DeYoung et al., 2020; Chalkidis et al., 2021; Rychener et al., 2023). Token-level methods preserve attribution resolution but often fragment

discourse. Sentence-level methods improve readability but may divide multi-sentence arguments or rulings. Paragraph-level methods retain broader context but may combine several rhetorical functions. To the best of our knowledge, existing rhetorical-role models have primarily used role labels for segmentation, summarization, document structuring, or classification, rather than for defining role-continuous perturbation units in model-agnostic explanation.

SPS-LIME addresses this gap by merging consecutive sentences with the same rhetorical role into variable-length semantic segments. These segments preserve discourse continuity while reducing the dimensionality of the perturbation space. The framework then performs segment-level attribution followed by sentence-level localization within highly ranked segments. This design separates two analytical questions: which coherent discourse segments influence the classifier, and which sentences within those segments contribute most strongly.

The evaluation framework also distinguishes faithfulness to classifier behavior from legal-role alignment. Consistency, comprehensiveness, and sufficiency assess the behaviour of the classifier and explainer, whereas Legal Role Alignment at k examines whether highly ranked units occur in rhetorically central parts of the judgment. This distinction allows SPS-LIME to identify not only coherent model explanations but also potential mismatches between classifier-sensitive content and role-based legal salience.

3 SPS-LIME Methodology

This section presents SPS-LIME, a model-agnostic local explanation framework designed to clarify why a legal document is assigned to a particular classification category. Standard LIME represents text using word-level interpretable features and estimates their local contributions through randomized perturbation (Ribeiro et al. 2016). However, independently perturbing words can produce fragmented or out-of-distribution inputs and substantially increase the dimensionality of the local search space (Rychener et al. 2023). In legal judgments, such fragmentation may also obscure information expressed across multiple sentences. This limitation is particularly relevant to legal judgments, in which facts, arguments, precedents, and judicial reasoning frequently extend across multiple sentences. Explanations constructed from isolated lexical features may therefore provide limited information about the discourse structures on which a legal classifier relies.

SPS-LIME addresses this limitation by defining explanations at the semantic-segment level. Each segment consists of a contiguous sequence of sentences assigned the same rhetorical role. Rhetorical-role modeling represents the functional structure of legal judgments through categories such as Fact, Argument, Precedent, Statute, and Ratio of the Decision (Malik et al. 2022). Accordingly, consecutive Fact sentences are merged into a factual segment, whereas consecutive Ratio of the Decision sentences form a reasoning segment. These role-homogeneous segments serve as the atomic units of perturbation and attribution, thereby preserving more of the semantic and discourse coherence of the source document.

Based on these segments, SPS-LIME provides a dual-layer explanation mechanism. The first layer assigns a contribution score to each semantic segment, allowing users to identify the discourse units that most strongly influence the classification result. The second layer performs sentence-level localization within highly ranked segments to identify the specific statements that contribute most strongly to the classification. This coarse-to-fine design combines segment-level coherence with sentence-level precision and provides explanations that are more closely aligned with the structural organization of legal judgments.

3.1 System Architecture

The overall operational workflow of SPS-LIME is designed as a hierarchical pipeline that transforms long legal texts into structurally coherent explanations without disrupting the underlying argumentative semantics. Conceptually, the framework comprises three main stages: text structuring, segment-level neighborhood generation, and local surrogate optimization.

The architecture is decomposed into two interconnected parts. The first part performs rhetorical role analysis and semantic unit extraction, converting raw judgments into semantic segments, as illustrated in Fig. 3. The second part carries out segment-based perturbation and explanation, constructing a dual-layer interpretive output on top of these segments.

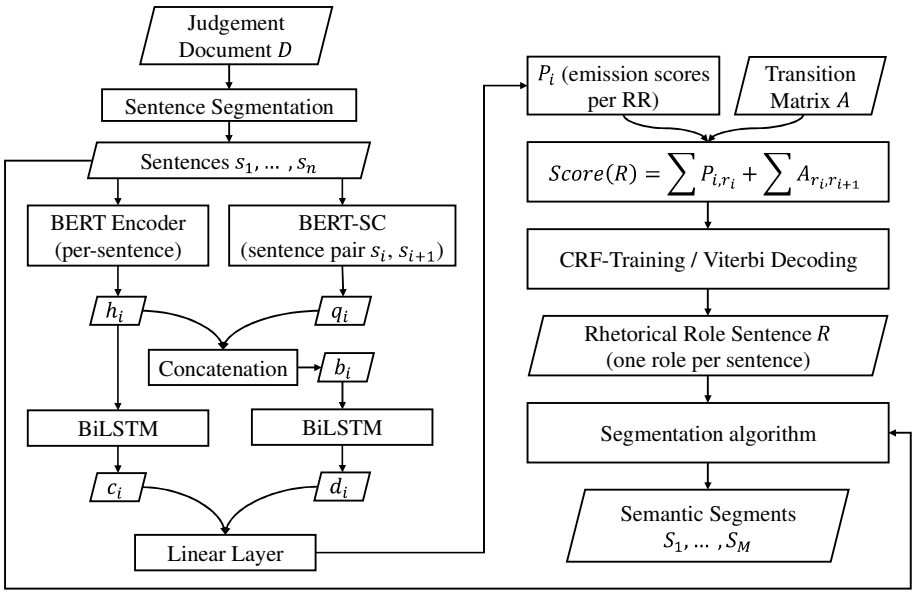


Fig. 3 Architecture of the first part: sentence-level rhetorical role tagging and semantic segmentation. Raw judgments are split into sentences, encoded with BERT and BERT-SC, passed through BiLSTM layers and a CRF to assign one rhetorical role per sentence, and finally grouped by a segmentation algorithm into semantic segments.

3.1.1 Rhetorical Role-Guided Document Segmentation

Let a lengthy legal document, such as a court judgment, be denoted by $D = \{s_1, s_2, \ldots, s_n\}$, where s_i represents the i-th sentence. Rather than perturbing the raw token stream directly, SPS-LIME first applies a Rhetorical Role (RR) classifier to the document. Each sentence s_i is assigned a legal discourse function $R(s_i) \in \mathcal{Y}$, where $\mathcal{Y}$ denotes the predefined set of seven rhetorical roles: Fact, Argument, Statute, Precedent, Ruling by Lower Court, Ratio of the Decision, and Ruling by Present Court.

Following this role assignment, an aggregation operator groups together contiguous sentences that share the same rhetorical role into a single semantic segment. Formally, the document is reorganized into a sequence of M segments:

$$S = \{S_1, S_2, \ldots, S_M\} \tag{1}$$

where each segment $S_j = \{s_{st_j}, s_{st_j+1}, \ldots, s_{ed_j}\}$ satisfies $R(s_{st_j}) = R(s_{st_j+1}) = \cdots = R(s_{ed_j})$, with $1 \le st_j \le ed_j \le n$. The indices st_j and ed_j mark the start and end of segment S_j within the original sentence sequence.

This segmentation step lifts the representation space from individual sentences to macro-level argumentative units, aligning the explanation granularity with the functional structure of legal reasoning.

Building on these semantic segments, the second part of the architecture treats the downstream classifier as a black box and generates local explanations by perturbing and re-weighting combinations of segments, producing the dual-layer SPS-LIME output, as illustrated in Fig. 4.

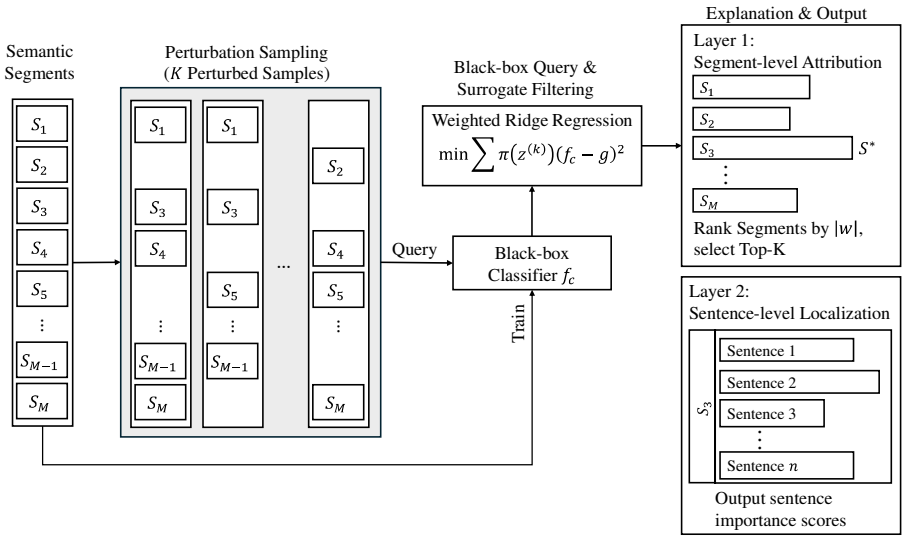


Fig. 4 Architecture of the second part: segment-based perturbation and dual-layer explanation. Semantic segments are sampled to create K perturbed inputs for a black-box classifier, and a weighted ridge regression surrogate is fitted to estimate segment-level importance. The top-k segments are then selected and, for each selected segment, sentence-level localization yields fine-grained importance scores within the segment.

3.1.2 Segment-Level Perturbation and Neighborhood Generation

SPS-LIME operates in an interpretable binary vector space $\{0,1\}^M$, where each dimension corresponds to one semantic segment. Let x denote the original legal document D to be explained, expressed in its raw textual form. The activation state of each segment in x is represented as a binary vector $z \in \{0,1\}^M$, whose components indicate whether each segment is retained or masked: a value of 1 denotes that the segment is retained, and 0 denotes that it is masked. The fully observable mask is denoted by $z_{\mathrm{o}} = \mathbf{1}_M = (1,1,\ldots,1)$, indicating that all semantic segments are retained.

A neighborhood dataset Z is constructed by randomly sampling perturbed binary vectors around z_{o}. Each sample $z = (z_1, z_2, \ldots, z_M) \in \{0,1\}^M$ represents a possible masking configuration:

$$z_j = \begin{cases} 1, & \text{if segment } S_j \text{ is retained} \\ 0, & \text{if segment } S_j \text{ is masked} \end{cases} \tag{2}$$

To evaluate each perturbed sample with the black-box classifier, the binary vector z is mapped back into the text domain as a perturbed document $\tilde{x}$: if $z_j = 0$, the entire content of segment S_j is removed; if $z_j = 1$, the segment is retained in its original form. This design compresses the original token-level feature space from $O(2^N)$ to $O(2^M)$, with $M \ll N$, while preserving the semantic continuity of the surrounding legal discourse blocks. The segment-level perturbation process is illustrated in Fig. 5.

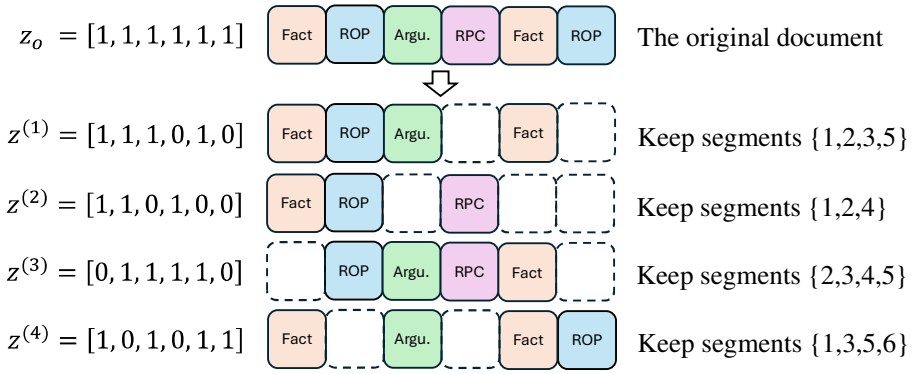


Fig. 5 Illustration of the segment-level perturbation scheme in SPS-LIME. Each semantic segment is randomly masked as a unit during perturbation, each segment is associated with a single rhetorical role, and adjacent segments correspond to different roles.

3.1.3 Local Surrogate Optimization and Attribution Calculation

Building on the previous subsections, the original judgment has been mapped into the rhetorical-role-based segment space, and a neighborhood of perturbed segment masks has been generated. The next task is to translate the resulting changes in black-box classification into segment-level importance scores meaningful for legal interpretation. To emphasize neighborhood samples that are close to the original document, SPS-LIME assigns each perturbed segment mask $z^{(k)}$a proximity weight relative to the original fully observable mask $\mathbf{1}_M$. In the binary segment-mask space, the L_1distance is

defined as $d\left(z^{(k)}, \mathbf{1}_M\right) = \sum_{j=1}^{M}\left(1 - z_j^{(k)}\right)$. Because $z_j^{(k)} \in \{0,1\}$, this distance is equivalent to the Hamming distance and counts the number of masked semantic segments. The proximity weight is then computed using the normalized exponential kernel:

$$\pi\left(z^{(k)}\right) = \exp\left(-\frac{d\left(z^{(k)}, \mathbf{1}_M\right)}{\sigma M}\right) \tag{3}$$

where σ is the kernel width controlling the decay rate of proximity and is fixed to 1 in the experiments. Consequently, samples containing fewer masked segments receive larger proximity weights, whereas more extensively perturbed samples receive smaller weights. The complete formulation is provided in Section 3.3.4.

For each document x, SPS-LIME fits a local linear surrogate model g_x over the segment mask space:

$$g_x(z) = b + w^{\top} z \tag{3}$$

where $w = [w_1, \ldots, w_M]$ is the coefficient vector and b is the bias term. Each coefficient w_j corresponds to semantic segment S_j, representing its linear contribution to the target-class probability.

Let f denote the black-box classifier that SPS-LIME is trying to explain at the document level. Let g_x denote an interpretable local surrogate model defined over the segment mask space. The surrogate g_x is selected by solving a regularized least-squares problem over a hypothesis class G, where G denotes the candidate set of linear surrogate models:

$$g_x = \arg\min_{g \in G} \mathcal{L}(f_c, g, \pi) + \Omega(g) \tag{4}$$

The proximity-weighted loss is defined as

$$\mathcal{L}(f_c, g, \pi) = \sum_{z \in Z} \pi(z)\,(f_c(\tilde{x}) - g(z))^2 \tag{5}$$

A regularization term $\Omega(g)$ is used to stabilize the local surrogate model. The specific weighted ridge regression objective is presented in Section 3.3.4, while explanation conciseness is introduced separately through top-k selection in Section 3.4.1.

Once converged, the coefficient vector serves as the segment-level attribution weights.

3.2 Rhetorical Role Tagging and Semantic Segmentation

This section describes how SPS-LIME converts raw judicial sentences into functional semantic units that support subsequent perturbation-based explanation. The pipeline proceeds in two stages: a sequence labeling model first assigns a rhetorical role to each sentence, and a deterministic segmentation procedure then groups adjacent sentences with the same role into coherent segments.

3.2.1 Definition of Rhetorical Roles (RR)

To make the discourse structure of legal judgments amenable to computational analysis, this study adopts the seven coarse-grained rhetorical roles used in the experiments

of Malik et al. (2022): Fact, Argument, Statute, Precedent, Ruling by Lower Court, Ratio of the Decision, and Ruling by Present Court. Let a judgment document D consist of a sequence of sentences $S = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes the i-th sentence. For each sentence s_i, the tagging model predicts a functional label $r_i \in \mathcal{Y}$, where $\mathcal{Y}$ contains seven rhetorical roles associated with legal discourse:

- Fact: objective factual background and chronological events relevant to the case.
- Argument: claims, submissions, or contentions advanced by the parties.
- Statute: explicit statutory provisions or codified rules cited in the judgment.
- Precedent: prior judicial decisions invoked as sources of legal authority.
- Ruling by Lower Court: outcomes and decisions rendered in earlier instances.
- Ratio of the Decision (ROD): the core legal reasoning and application of law articulated by the judge.
- Ruling by Present Court (RPC): the final verdict or order issued in the present instance.

This scheme maps the raw sentence sequence into a corresponding label sequence $R = \{r_1, r_2, \ldots, r_n\}$, providing a structural backbone for downstream discourse analysis and attribution.

3.2.2 Sequence Labeling Model Architecture: Hierarchical BiLSTM-CRF

Each sentence s_i is separately encoded using a shared BERT encoder (Devlin et al. 2019), producing a contextual sentence representation $h_i \in \mathbb{R}^{d_h}$, where d_h denotes the dimensionality of the sentence representation. In parallel, each pair of adjacent sentences (s_i, s_{i+1}) is processed by a sentence-pair encoder, denoted BERT-SC, to obtain a transition representation $q_i \in \mathbb{R}^{d_q}$. This representation captures the semantic relationship between adjacent sentences and provides information about potential rhetorical-role transitions. Following Malik et al. (2022), the sentence representations and transition representations are processed by two separate BiLSTM branches. Their outputs are concatenated and passed to a CRF layer for rhetorical-role sequence decoding, which predicts the rhetorical role of each sentence while modeling dependencies between adjacent role labels.

RR Module (rhetorical role branch): The sentence vector h_i is fed into an independent BiLSTM, producing a contextual sentence representation c_i that captures the sentence's role within the global document context.

Shift Module (transition branch): This module captures transitions in the argumentative structure between adjacent sentences. For each pair (s_i, s_{i+1}), the module computes an intermediate representation q_i encoding the semantic and argumentative relation between the two sentences. It then concatenates the previous pair representation q_{i-1}, the current sentence vector h_i, and the current pair representation q_i to form the Shift Module input at position i:

$$b_i = [q_{i-1}; h_i; q_i] \tag{6}$$

where $[\cdot;\cdot]$ denotes vector concatenation, and missing pair representations at document boundaries are padded with zero vectors. The sequence $B = \{b_1, b_2, \ldots, b_n\}$ is fed into

an independent BiLSTM, producing hidden states d_i that summarize local structural shifts in the discourse.

To compute the final emission scores for rhetorical role tagging, the model concatenates the RR Module's contextual representation c_i with the Shift Module's transition representation d_i, and applies a linear transformation:

$$P_i = W_{\mathrm{rr}}[c_i; d_i] + b_{\mathrm{rr}} \tag{7}$$

where W_{rr} and b_{rr} are the trainable weight matrix and bias term. The scalar P_{i,r_i} denotes the emission score of assigning label r_i to sentence s_i. Each role classification is thus jointly informed by sentence-level semantics and macroscopic argumentative dynamics.

Following Malik et al. (2022), a Conditional Random Field layer is applied to model dependencies among adjacent rhetorical roles and decode the globally optimal label sequence (Lafferty et al. 2001). Given a candidate label sequence $R = \{r_1, r_2, \ldots, r_n\}$, its sequence score is defined as:

$$\mathrm{Score}(R) = \sum_{i=1}^{n} P_{i,r_i} + \sum_{i=1}^{n-1} A_{r_i, r_{i+1}} \tag{8}$$

where n is the number of sentences and $A_{r_i, r_{i+1}}$ is the learned transition score from role r_i to role r_{i+1}. All model parameters, including the BERT and BERT-SC encoders, both BiLSTM branches, the linear projection layer, and the CRF transition matrix A, are jointly optimized by maximizing the conditional log-likelihood of the gold rhetorical role sequence over the training corpus. At inference time, the optimal role sequence is decoded via the Viterbi algorithm.

3.2.3 Semantic Segment Construction Algorithm

Once the global rhetorical role sequence has been decoded, SPS-LIME applies a deterministic segmentation algorithm to partition the document into contiguous and functionally homogeneous text blocks called semantic segments.

Formally, given the sentence sequence $s = \{s_1, s_2, \ldots, s_n\}$ and its role sequence $R = \{r_1, r_2, \ldots, r_n\}$ with $r_i \in \mathcal{Y}$, the algorithm scans the document to detect role boundaries. A boundary is recorded at index i whenever $r_i \neq r_{i-1}$. Consecutive sentences sharing the same role are merged into a single segment. Fig. 6 illustrates this merging process using an example judgment.

ID	Sentence	Rhetorical Role	Segment
s_1	Holywell Corporation filed a federal income tax return for the fiscal year ending July 31, 1985.	Fact	S_1
s_2	The return included capital gains from the sale of the Washington Properties.	Fact	
s_3	Neither the corporate debtors nor trustee Smith filed returns for later fiscal years.	Fact	
s_4	Those years included gains from the sale of Miami Center and interest earned on the proceeds.	Fact	
s_5	The Bankruptcy Court declared that Smith had no duty to file tax returns or pay income tax.	RulingBy LowerCourt	S_2
s_6	The District Court affirmed that declaration.	RulingBy LowerCourt	
s_7	The Court of Appeals also affirmed.	RulingBy LowerCourt	
s_8	We hold that Smith must file returns and pay tax on the income attributable to the debtors' property.	RulingBy PresentCourt	S_3

Fig. 6 Illustration of how consecutive sentences sharing the same rhetorical role are merged into semantic segments within a legal judgment.

For the j-th segment, let $S_j = \{s_{st_j}, s_{st_j+1}, \cdots, s_{ed_j}\}$ denote the ordered set of contiguous sentences in that segment, where $1 \le st_j \le ed_j \le n$, and $\mathcal{Y}_j = r_{st_j} = r_{st_j+1} = \cdots = r_{ed_j}$ is the unified rhetorical role of the block. The first segment S_1 is initialized with $st_1 = 1$, and for any adjacent pair S_j, S_{j+1}, the continuity constraint $ed_j + 1 = st_{j+1}$ holds together with $\mathcal{Y}_j \neq \mathcal{Y}_{j+1}$.

Through this aggregation, the document is mapped from an n-dimensional sentence space to an M-dimensional semantic segment space, compressing the feature space faced by the local surrogate model while preserving rhetorical and logical coherence.

3.3 Semantic-Preserved Perturbation Strategy

Following the construction of semantic segments, the next step in SPS-LIME is to generate a representative neighborhood of perturbed documents in order to probe the local decision boundary of the black-box classifier. This section presents the semantic-preserved perturbation strategy, which shifts the atomic unit of masking from individual tokens to holistic functional segments.

3.3.1 Dimensionality Reduction and Search Space Analysis

Traditional token-level perturbation applied to lengthy legal documents suffers from the curse of dimensionality. For a document containing N tokens, the theoretical perturbation search space grows as $O(2^N)$. By elevating the perturbation unit to semantic segments, SPS-LIME compresses the feature space from N tokens to M segments ($M \ll N$), reducing the search space to $O(2^M)$. This compression improves computational efficiency, reduces within-sentence fragmentation, and preserves the internal content of retained segments, since each masked unit corresponds to a complete legal functional block rather than an arbitrary token span.

3.3.2 Mathematical Modeling of Segment-Level Sampling

The original unperturbed document is represented by a fully observable binary state vector:

$$\mathbf{1}_M = (1,1,\ldots,1) \in \{0,1\}^M \tag{9}$$

where each dimension corresponds to one semantic segment and a value of 1 indicates that the segment is present.
SPS-LIME draws K perturbed samples to construct the local neighborhood ($K = 1000$ in the experimental setup). The binary mask vector of the k-th perturbed sample is denoted:

$$z^{(k)} = [z_1^{(k)}, z_2^{(k)}, \ldots, z_M^{(k)}], z_j^{(k)} \in \{0,1\} \tag{10}$$

The sampling process employs an independent Bernoulli masking mechanism: the retention state of each segment S_j in the k-th sample is governed by

$$\Pr(z_j^{(k)} = 1) = 1 - p_{\text{mask}}, \Pr(z_j^{(k)} = 0) = p_{\text{mask}} \tag{11}$$

where p_{mask} is the predefined masking probability, empirically set to 0.4. The masking decision for each segment is independent of all others, providing sufficient local variation while maintaining a balanced proportion of retained argumentative blocks.

3.3.3 Sample Synthesis and Preservation of Semantic Coherence

Given a mask vector $z^{(k)}$, the framework synthesizes the corresponding perturbed text $\tilde{x}^{(k)}$ via the mapping:

$$\tilde{x}^{(k)} = \bigoplus_{j=1}^{M} \left(\mathbb{I}[z_j^{(k)} = 1] \cdot S_j \right) \tag{12}$$

where $\oplus$ denotes string concatenation in the original chronological order of the document, and $\mathbb{I}[\cdot]$ is the indicator function that equals 1 if the condition holds and 0 otherwise. A segment S_j is included in $\tilde{x}^{(k)}$ if and only if $z_j^{(k)} = 1$; otherwise it is omitted entirely.

Because each retained segment corresponds to an intact legal functional module, the synthesized samples preserve the original ordering of retained segments and reduce within-segment fragmentation of the judgment, rather than breaking its argumentative structure into isolated fragments.

3.3.4 Local Surrogate Model Fitting

After constructing a neighborhood of perturbed document samples and obtaining their corresponding classifications from the black-box classifier, the next step is to build an interpretable local surrogate model g that approximates the classification behavior of f around the original judgment instance. This section describes how SPS-LIME defines distances and proximity weights in the segment-masking space and fits a locally weighted ridge regression model to estimate segment-level attribution scores, following

the local surrogate framework of Ribeiro et al. (2016) and the ridge regularization formulation of Hoerl and Kennard (1970).

Distance Metric and Proximity Weighting: To ensure the local explanation remains faithful to the black-box model's behavior in the immediate vicinity of the original input, each perturbed sample is assigned a proximity weight. In the binary segment-masking space of SPS-LIME, the L_1 (Manhattan) distance measures the divergence between a perturbed sample $z^{(k)}$ and the original fully observable state vector $\mathbf{1}_M$:

$$d(z^{(k)}, \mathbf{1}_M) = \sum_{j=1}^{M} \left|1 - z_j^{(k)}\right| \tag{13}$$

Since $z_j^{(k)} \in \{0,1\}$, this L_1 distance is numerically equivalent to the Hamming distance and directly counts the number of masked semantic segments. Larger values indicate samples structurally farther from the original document.

This distance is transformed into a continuous proximity weight $\pi(z^{(k)})$ via an exponential kernel:

$$\pi(z^{(k)}) = \exp\left(-\frac{d(z^{(k)}, \mathbf{1}_M)}{\sigma \cdot M}\right) \tag{14}$$

where σ is the kernel width controlling the decay rate of proximity, empirically fixed to 1, simplifying the function to:

$$\pi(z^{(k)}) = \exp\left(-\frac{d(z^{(k)}, \mathbf{1}_M)}{M}\right) \tag{15}$$

Normalizing by M keeps the spatial scale of proximity weights consistent across documents of varying lengths.

Weighted Ridge Regression Optimization: Because the primary objective is to deliver segment-level attribution rather than a highly complex local approximation, a linear model is adopted as the interpretable surrogate:

$$g(z) = b_{\text{seg}} + w_{\text{seg}}^{\top} z \tag{16}$$

where $z \in \{0,1\}^M$ is the segment-level mask vector, $w_{\text{seg}} \in \mathbb{R}^M$ is the weight vector representing the importance of each segment, and b_{seg} is the bias term.

For a target class c, the surrogate parameters (w_{seg}, b_{seg}) are learned using the synthesized perturbed texts, their corresponding masks, and the black-box classification probabilities, by solving the weighted ridge regression problem:

$$\min_{w_{\text{seg}}, b_{\text{seg}}} \sum_{k=1}^{K} \pi(z^{(k)}) \left(f_c(\tilde{x}^{(k)}) - \left(b_{\text{seg}} + w_{\text{seg}}^{\top} z^{(k)}\right)\right)^2 + \alpha_{\text{seg}} \left\|w_{\text{seg}}\right\|_2^2 \tag{17}$$

This objective comprises three components: the weighted squared error measures local fidelity; the proximity weights $\pi(z^{(k)})$ enforce locality; and the L_2 regularization term

$\alpha_{\text{seg}} \| w_{\text{seg}} \|_2^2$ ($\alpha_{\text{seg}} = 1.0$) prevents overfitting and stabilizes weight estimates in the presence of correlated segment features.

After optimization, the weight vector $w_{\text{seg}} = [w_1, w_2, \ldots, w_M]$ provides the segment-level attribution scores. The absolute value $|w_j|$ quantifies the relative importance of segment S_j, while the sign indicates the direction of its influence: $w_j > 0$ increases confidence in class c, $w_j < 0$ weakens it.

3.4 Dual-Layer Explanation Mechanism

Building on the segment-level attribution weights, SPS-LIME introduces a dual-layer explanation mechanism intended to support coarse-to-fine inspection of legal judgments. The first layer provides a macroscopic view by ranking semantic segments according to their contribution to the classification outcome; the second layer provides a microscopic view by localizing the most decisive sentences within the highest-ranked segments.

3.4.1 Layer 1: Segment-Level Attribution

The first layer ranks the semantic segments by their absolute attribution weights. Based on $w_{\text{seg}} = [w_1, w_2, \ldots, w_M]$, segments are ordered by $|w_j|$ in descending order, and the top-K_{attr} most influential segments are selected for further sentence-level analysis in Layer 2. Unlike traditional LIME, which produces scattered token-level attributions, this segment-level ranking maps directly to cohesive legal functional blocks, including factual narratives, statutory citations, and judicial reasoning.

3.4.2 Layer 2: Sentence-Level Localization

The second layer zooms into the top-ranked segments to identify the most decisive individual sentences. The top $K_{\text{attr}} = \max(1, \lfloor M/2 \rfloor)$ segments are selected to balance explanation coverage against computational cost. For each such segment S^*, the analysis proceeds in three steps.

Step 1: Target Segment Decomposition. The selected segment $S^*$is decomposed into its constituent sentence sequence $\mathcal{T}^* = \{s_1, s_2, \ldots, s_L\}$, where L is the number of sentences.

Step 2: Localized Sentence-Level Sampling. A new binary mask space is constructed over $\mathcal{T}^*$, treating sentences as atomic perturbation units. Perturbed samples are drawn with a fixed retention probability of 0.5, with sample size capped at $K_{\text{sent}} = \min(100, K/2)$.

Step 3: Sentence-Level Surrogate Fitting. A local linear surrogate g_{sent} is fitted:

$$g_{\text{sent}}(z_{\text{sent}}) = b_{\text{sent}} + (w^{(\text{sent})})^{\top} z_{\text{sent}} \tag{18}$$

where $z_{\text{sent}} \in \{0,1\}^L$, b_{sent} is the bias term, and $w^{(\text{sent})} \in \mathbb{R}^L$ is the sentence-level weight vector, estimated by solving:

$$\min_{w^{(\text{sent})},b_{\text{sent}}} \sum_{k=1}^{K_{\text{sent}}} \pi_{S^*}\left(z_{\text{sent}}^{(k)}\right)\left[f_c\left(\tilde{x}_{\text{sent}}^{(k)}\right) - g_{\text{sent}}\left(z_{\text{sent}}^{(k)}\right)\right]^2 + \alpha_{\text{sent}}\left\|w^{(\text{sent})}\right\|_2^2 \tag{19}$$

where $\tilde{x}_{\text{sent}}^{(k)}$ is the k-th synthesized sentence-level perturbed text, $\pi_{S^*}(\cdot)$ is the proximity weight centered on S^*, and $\alpha_{\text{sent}} = 1.0$. The resulting weights $w^{(\text{sent})}$ isolate the most decisive sentences within S^*.

3.4.3 Visualization Design of Explanation Reports

SPS-LIME renders both layers of attribution as a hierarchical heatmap projected back onto the original judgment text. At the first layer, segment boundaries are highlighted with a lighter background shade. Within each highlighted block, second-layer critical sentences are accentuated with higher-saturation colors whose intensity is proportional to $\left|w_p^{(\text{sent})}\right|$, where p indexes sentences within the selected segment.

This two-level visualization allows users to inspect explanations first at the segment level and then at the sentence level.

3.5 Hybrid Black-Box Classifier Architecture

To demonstrate the applicability of SPS-LIME, a hybrid feature classifier is instantiated as the underlying black-box model, jointly leveraging high-dimensional lexical information and low-dimensional discourse-structural signals.

3.5.1 Lexical Feature Extraction: TF-IDF Representation

Each judgment is represented as a sparse lexical feature vector using term frequency-inverse document frequency weighting (Salton and Buckley 1988). A judgment is treated as a single document, and the TF-IDF representation is computed from its complete preprocessed text. Preprocessing includes tokenization, lowercasing, stop-word removal, and the removal of very short non-alphabetic tokens. The TF-IDF representations were implemented using the scikit-learn machine learning framework (Pedregosa et al. 2011).

A unigram and bigram vocabulary is used, with the n-gram range set to $(1,2)$, and TF-IDF weights are estimated with respect to the entire training corpus, yielding a sparse V-dimensional lexical feature vector $x_{\text{lex}} \in \mathbb{R}^V$ for each document, where V is the vocabulary size. In the fusion stage, x_{lex} is combined with an embedding of the document's rhetorical role distribution to form the final input representation x_{in}.

3.5.2 Feature Fusion Mechanism and Network Architecture

To merge the high-dimensional TF-IDF lexical features with the low-dimensional Rhetorical Role (RR) distribution features, a weighted concatenation mechanism is employed. Before concatenation, the raw 7-dimensional RR distribution vector is projected into a 128-dimensional representation space via a non-linear mapping function, which increases its representational capacity while keeping the structural feature vector relatively compact.

$$\phi(x_{\text{rr}}) = W_2^{\text{fuse}}\,\text{ReLU}(W_1^{\text{fuse}} x_{\text{rr}} + b_1^{\text{fuse}}) + b_2^{\text{fuse}} \tag{20}$$

where $W_1^{\text{fuse}} \in \mathbb{R}^{64\times 7}$ and $W_2^{\text{fuse}} \in \mathbb{R}^{128\times 64}$ are the trainable weight matrices of the two fully connected layers, and $b_1^{\text{fuse}} \in \mathbb{R}^{64}$, $b_2^{\text{fuse}} \in \mathbb{R}^{128}$ are their bias terms.
The projected RR representation is scaled by a learnable scalar weight γ and concatenated with the TF-IDF feature vector x_{lex}:

$$x_{\text{in}} = x_{\text{lex}} \oplus \left(\gamma \cdot \phi(x_{\text{rr}})\right) \tag{21}$$

where $\oplus$ denotes the concatenation operator, and γ regulates the overall influence of structural RR features relative to purely lexical evidence.
The fused vector $x_{\text{in}} \in \mathbb{R}^{V+128}$ is fed into a Multi-Layer Perceptron (MLP) with two hidden layers (dimensions 512 and 256), each followed by ReLU activation and dropout (probability 0.3) to mitigate overfitting. The overall process of combining TF-IDF lexical features with rhetorical role distribution features is illustrated in Fig. 7.

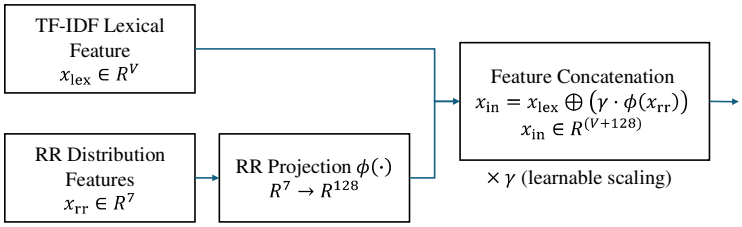


Fig. 7 Feature fusion architecture combining TF-IDF lexical features with projected rhetorical role distribution features via learnable scaling and concatenation before MLP classification.

3.5.3 Classification Probabilities and Explanation Target

For an input document x, let the classifier produce a logit vector $y = [y_1, y_2, \ldots, y_C]$, where C denotes the number of target classes. In the primary SCOTUS classification setting used in this study, $C = 13$. The dataset is based on the SCOTUS task included in LexGLUE (Chalkidis et al. 2022). The classification probability for target class c is computed using the softmax function:

$$f_c(x) = \frac{\exp(y_c)}{\sum_{i=1}^{C} \exp(y_i)} \tag{22}$$

Within the SPS-LIME pipeline, $f_c(x)$ serves as the precise decision signal that the local surrogate model aims to approximate. All segment-level and sentence-level perturbed samples are evaluated against this black-box output to quantify the relative attribution of different semantic units.
For the auxiliary evaluation, we derive a document-level binary classification task from the ECtHR corpus of Chalkidis et al. (2019). A case is labelled as violation if at least one Convention article is identified as violated and as non-violation otherwise. This binary formulation differs from the original multi-label article-level classification task.

3.6 Summary of the Methodology

This section establishes the methodological foundation of SPS-LIME. By elevating the unit of perturbation from isolated tokens to rhetorically cohesive semantic segments, the framework sharply reduces the effective search space while preserving the argumentative integrity of long legal documents. Building on this segment-based representation, a two-layer explanation mechanism and a hybrid feature classifier together provide a structurally aware paradigm for interpreting complex legal text classifications.

3.6.1 Module Dependency and Data Flow

The pipeline begins with the Rhetorical Role Tagger and Semantic Segmentation module, which converts raw judgment text into a sequence of M rhetorically homogeneous semantic segments $\{S_1, S_2, \dots, S_M\}$. These segments serve as the atomic units for all downstream operations, compressing the perturbation search space from $O(2^N)$ to $O(2^M)$ with $M \ll N$.

The Semantic-Preserved Perturbation Strategy constructs a local neighborhood of $K = 1000$ binary-masked perturbed samples, each synthesized by independently retaining or omitting complete semantic segments with masking probability $p_{\text{mask}} = 0.4$. The Local Surrogate Model Fitting module fits a weighted ridge regression over these samples, using L_1-distance-based proximity weights and L_2 regularization ($\alpha_{\text{seg}} = 1.0$) to produce the segment-level attribution weights w_{seg}.

3.6.2 Dual-Layer Explanation Output

The Dual-Layer Explanation Mechanism translates segment-level attribution weights into a structured, two-layer explanation report. Layer 1 ranks the M semantic segments by $|w_j|$ and surfaces the top-K_{attr} most influential argumentative blocks. Layer 2 performs sentence-level localization within these top segments via a recursive surrogate fitting procedure, rendering the result as a hierarchical heatmap projected back onto the original judgment text.

3.6.3 Architectural Basis of the Rhetorical Role Tagger

The rhetorical-role tagging model is adapted from the rhetorical-role and shift-aware architecture proposed by Malik et al. (2022). It contains two parallel BiLSTM branches. The rhetorical-role branch models contextual sentence representations, whereas the shift-detection branch captures transitions between adjacent sentences. The outputs of the two branches are concatenated and transformed into emission scores through a linear projection layer. A CRF layer then models dependencies among adjacent labels and produces the final rhetorical-role sequence. This architecture jointly incorporates sentence-level semantics and discourse-transition information into rhetorical-role classification.

The effectiveness of SPS-LIME is evaluated in Section 4 through experiments on SCOTUS and ECtHR. The evaluation examines classification performance, explanation consistency, comprehensiveness, sufficiency, legal role alignment, and the effect of segment granularity. SPS-LIME is compared with token-level and sentence-level

LIME variants to assess the benefits and limitations of rhetorical segment-based perturbation.

4 Experimental Results

Recent surveys document the growing use of NLP for legal document analysis and judgment classification (Zhong et al. 2020), while benchmark suites such as LexGLUE provide standardized evaluation across multiple legal text classification tasks (Chalkidis et al. 2022). Research on explainable AI in law further emphasizes the need to clarify how opaque models derive classifications in legally consequential settings (Richmond et al. 2024). Accordingly, this section evaluates SPS-LIME from two perspectives. First, it examines the classification performance of the underlying model and the contribution of rhetorical-role information. Second, it compares Token LIME, Sentence LIME, and SPS-LIME on the SCOTUS and ECtHR datasets using Consistency, Comprehensiveness, Sufficiency, and Legal Role Alignment at k, denoted LRA@ k. These evaluations assess both classification performance and the types of information captured by the resulting explanations.

4.1 Dataset and Black-box Classifier Setup

Experiments are conducted on two publicly available legal datasets. The primary evaluation uses the official 13-class LexGLUE-SCOTUS dataset (Chalkidis et al. 2022), which formulates Supreme Court opinion classification as a 13-class issue-area classification task. The original chronological training, validation, and test partitions are retained.

The auxiliary evaluation uses a document-level binary violation-classification task derived from the ECtHR corpus of Chalkidis et al. (2019), in which each case is classified according to whether the Court found at least one violation of the European Convention on Human Rights. For both datasets, the original chronological train, validation, and test partitions are retained.

Both tasks exhibit class imbalance. We therefore report micro-F1 and macro-F1. Micro-F1 aggregates decisions across all instances and is dominated by frequently occurring classes, whereas macro-F1 assigns equal weight to each class and provides a more balanced assessment of minority-class performance.

For reference, we report the SCOTUS results of representative transformer-based and long-document models from the LexGLUE benchmark, including BERT (Devlin et al. 2019), RoBERTa (Liu et al. 2019), DeBERTa (He et al. 2021), Longformer (Beltagy et al. 2020), BigBird (Zaheer et al. 2020), and Legal-BERT (Chalkidis et al. 2020). These results are taken from prior work rather than reproduced under the present implementation.

The primary black-box classifier used for explanation is the proposed TF-IDF+RR+MLP model, which combines lexical TF-IDF features with rhetorical-role representations. The same classifier is used for Token LIME, Sentence LIME, and SPS-LIME to ensure that differences in explanation quality arise from the explanation method rather than from the underlying classification model.

4.2 Experimental Settings

The TF-IDF representation uses a maximum vocabulary size of 10,000 and includes unigrams and bigrams. The rhetorical-role distribution vector is projected through two fully connected layers with dimensions $7 \rightarrow 64 \rightarrow 128$, and the fused representation is classified using an MLP with hidden dimensions of 512 and 256. ReLU activation and dropout with a probability of 0.3 are applied after the hidden layers.

The classifier is trained using Adam with an initial learning rate of 1×10^{-3}, a batch size of 32, and a maximum of 50 epochs. Cross-entropy loss is used as the training objective. The learning rate is reduced by a factor of 0.5 when the validation loss does not improve for two consecutive epochs, and early stopping is applied with a patience of five epochs. Classification experiments are repeated using five random seeds, namely 13, 21, 42, 87, and 100, and the results are reported as mean and standard deviation.

Token LIME, Sentence LIME, and SPS-LIME are applied to the same trained black-box classifier. All three methods generate 1,000 perturbed samples per document with a masking probability of 0.4. For SPS-LIME, the proximity-kernel width is fixed to $\sigma = 1$, and the weighted ridge regression coefficient is set to $\alpha_{\text{seg}} = 1.0$. Token LIME is configured to return 20 explanatory features. Consistency is computed using the top 10 explanatory units, while Comprehensiveness and Sufficiency are averaged across selection ratios of 1%, 5%, 10%, 20%, and 50%.

For sentence-level localization, SPS-LIME selects the top $\max(1, \lfloor M/2 \rfloor)$ semantic segments. Within each selected segment, sentence-level masks are sampled with a retention probability of 0.5. The number of sentence-level perturbation samples is set to $K_{\text{sent}} = \min(100, K/2)$, and the sentence-level ridge coefficient is fixed to $\alpha_{\text{sent}} = 1.0$. The experiments are implemented in Python using PyTorch and executed on an NVIDIA GeForce RTX 3050 Laptop GPU with 4 GB of memory.

4.3 Classification Metrics

To evaluate classification performance, this paper reports micro-F1 and macro-F1. Let TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives for class c, respectively. Micro-precision and micro-recall are defined as follows:

$$Precision_{micro} = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FP_c} \tag{23}$$

$$Recall_{micro} = \frac{\sum_{c \in C} TP_c}{\sum_{c \in C} TP_c + \sum_{c \in C} FN_c} \tag{24}$$

The F1 score for class c is:

$$F1_c = 2 \times \frac{Precision_c \times Recall_c}{Precision_c + Recall_c} \tag{25}$$

Accordingly, micro-F1 and macro-F1 are computed as:

$$Micro\text{-}F1 = 2 \times \frac{Precision_{micro} \times Recall_{micro}}{Precision_{micro} + Recall_{micro}} \tag{26}$$

$$Macro\text{-}F1 = \frac{1}{\mid C \mid} \sum_{c \in C} F1_c \tag{27}$$

Micro-F1 emphasizes overall accuracy across all instances, whereas macro-F1 gives equal importance to each class and is therefore more sensitive to performance on minority labels. This distinction is particularly important in imbalanced legal datasets such as ECtHR, where a model may achieve a high micro-F1 by favoring the majority class while still performing poorly on minority classes.

4.4 Competitiveness of the Black-Box Classifier

Before evaluating explanations, it is necessary to establish that the adopted black-box classifier has reasonable classification performance. Table 1 compares the proposed RR+TF-IDF+MLP model with representative baseline models previously reported on SCOTUS. The results show that RR+TF-IDF+MLP achieves 0.734 micro-F1 and 0.648 macro-F1, outperforming several general-purpose embedding-based and long-document models such as BERT, RoBERTa, DeBERTa, BigBird, and Longformer, while remaining slightly below domain-specific Legal-BERT. Because the baseline results are taken from the LexGLUE study, this comparison is intended to provide a reference point rather than a fully controlled head-to-head evaluation.

Table 1 Comparison with previously reported results on the LexGLUE-SCOTUS dataset.

Model	Micro-F1	Macro-F1
BERT	0.683	0.583
RoBERTa	0.716	0.620
BigBird	0.728	0.620
DeBERTa	0.711	0.627
Longformer	0.729	0.640
RR+TF-IDF+MLP	**0.734**	**0.648**
Legal-BERT	0.764	0.665

This result suggests that combining TF-IDF lexical features with rhetorical-role information provides a competitive basis for legal document classification. Although Legal-BERT achieves stronger classification performance, RR+TF-IDF+MLP is used as the primary black-box classifier because it provides explicit control over lexical and rhetorical-role features and supports the subsequent input and feature-ablation analyses. All explanation methods are applied to the same classifier to ensure a controlled comparison.

4.5 Input Ablation by Rhetorical Role

To examine which parts of legal documents contribute most to classification, this paper performs an input ablation study in which only selected rhetorical-role segments are

retained. Table 2 reports the results on SCOTUS when the model is restricted to ROD+RPC, Argument only, Fact only, Fact+Argument, or Full Text.

Table 2 Classification performance of RR+TF-IDF+MLP under different input settings.

Input	Micro-F1	Macro-F1
Only ROD+RPC	0.630	0.456
Only Argument	0.669	0.575
Only Fact	0.716	0.626
Only Fact+Argument	0.729	0.644
Full Text	**0.734**	**0.648**

A notable result is that models using only Fact or Fact+Argument substantially outperform the model using only ROD+RPC. The Fact-only setting achieves a micro-F1 of 0.716 and a macro-F1 of 0.626, whereas the ROD+RPC setting achieves 0.630 and 0.456, respectively. This result indicates that factual and argumentative sections contain stronger classification signals for SCOTUS issue-area classification than sections explicitly presenting the court's reasoning and ruling. Although Ratio of the Decision and Ruling by Present Court are central to explaining a judicial decision, issue-area classification is not equivalent to reconstructing the court's doctrinal reasoning. Factual narratives and party arguments may directly reveal the subject matter and legal category of a dispute.

This pattern is consistent with prior findings that classification relevance and legal-rationale relevance may diverge. Chalkidis et al. (2021) showed that paragraph-level rationale extraction for ECtHR cases remains challenging and that sparsity, continuity, and comprehensiveness constraints do not uniformly improve rationale quality. Staliūnaitė et al. (2024) reported weak alignment between model attribution scores and rationales derived from judicial decisions. Namazov et al. (2026) similarly found substantial disagreement between automatically extracted rationales and legal-expert assessments despite favorable faithfulness scores. The strong performance of Fact-based inputs therefore does not by itself establish shortcut learning, but it is consistent with the possibility that the classifier relies more heavily on dataset-specific factual patterns than on passages that explicitly articulate legal reasoning.

4.6 Ablation on RR Integration

This paper further investigates how rhetorical-role features interact with TF-IDF features under different integration settings. On SCOTUS, Table 3 shows that adding RR to TF-IDF consistently improves performance over the TF-IDF+MLP baseline. Among the tested initializations, $\gamma = 20$ yields the highest mean performance on SCOTUS, whereas $\gamma = 10$ yields the highest micro-F1 on ECtHR. In contrast, using RR alone performs very poorly, with only 0.3156 micro-F1 and 0.0877 macro-F1. The poor performance of the RR-only model indicates that rhetorical role distributions alone are insufficient for SCOTUS classification, which relies primarily on lexical features.

Table 3 Ablation results for different classifier configurations (SCOTUS)

Model	Micro-F1	Macro-F1
TF-IDF+MLP (Baseline)	0.7180±0.0020	0.6286±0.0043
RR+TF-IDF+MLP (initial $\gamma = 0.0$)	0.7211±0.0035	0.6322±0.0063
RR+TF-IDF+MLP (initial $\gamma = 5.0$)	0.7290±0.0035	0.6403±0.0034
RR+TF-IDF+MLP (initial $\gamma = 10.0$)	0.7264±0.0040	0.6371±0.0095
RR+TF-IDF+MLP (initial $\gamma = 20.0$)	**0.7336±0.0029**	**0.6477±0.0087**
RR+MLP	0.3156±0.0063	0.0877±0.0009

The situation is markedly different on ECtHR. As shown in Table 4, adding RR to TF-IDF increases micro-F1 but tends to reduce macro-F1 relative to the TF-IDF baseline. More strikingly, using RR alone produces the highest micro-F1, 0.8470, but also the lowest macro-F1, 0.4586. The RR-only model achieves high micro-F1 but very low macro-F1, indicating severe majority-class bias and limited balanced discrimination. This result shows that rhetorical-role distributions alone are insufficient for reliable ECtHR classification. Although the pattern is consistent with a potential shortcut, it does not by itself establish shortcut learning.

Table 4 Ablation results for different classifier configurations (ECtHR)

Model	Micro-F1	Macro-F1
TF-IDF+MLP (Baseline)	0.8326±0.0055	**0.6377±0.0088**
RR+TF-IDF+MLP (initial $\gamma = 0.0$)	0.8356±0.0071	0.6281±0.0067
RR+TF-IDF+MLP (initial $\gamma = 5.0$)	0.8392±0.0033	0.6331±0.0149
RR+TF-IDF+MLP (initial $\gamma = 10.0$)	0.8440±0.0030	0.6249±0.0088
RR+TF-IDF+MLP (initial $\gamma = 20.0$)	0.8404±0.0056	0.6248±0.0187
RR+MLP	**0.8470±0.0000**	0.4586±0.0000

This interpretation is consistent with the distinction between micro-F1 and macro-F1. Because micro-F1 aggregates over all instances, a model that heavily favors the majority label can still achieve a high score. Macro-F1, by contrast, penalizes such imbalance because it gives equal weight to each class. Therefore, the ECtHR results indicate that rhetorical-role features can become disproportionately influential under class imbalance, leading to apparently strong performance that is partly driven by majority-class bias rather than balanced discrimination.

The best-performing initialization of γ differs across the two datasets, with 20 for SCOTUS and 10 for ECtHR, suggesting that the contribution of rhetorical role information is dataset dependent and may be influenced by differences in task formulation, class distribution, and document structure.

4.7 Explanation Evaluation Metrics

To evaluate explanation quality, this paper adopts three metrics, namely Consistency, Comprehensiveness (COMP), and Sufficiency (SUFF), together with the proposed Legal Role Alignment at k (LRA@k).

Consistency measures the stability of an explanation method across repeated executions with different random perturbation samples. For a given document, let $E_i$and $E_j$denote the sets containing the top-k explanatory units obtained from the i-th and j-th runs, respectively. Their agreement is measured using Jaccard similarity:

$$J(E_i, E_j) = \frac{|E_i \cap E_j|}{|E_i \cup E_j|} \tag{28}$$

Let N_{iter} denote the number of independent explanation runs for the same document. Consistency is defined as the average Jaccard similarity over all pairs of runs:

$$\text{Consistency} = \frac{2}{N_{\text{iter}}(N_{\text{iter}} - 1)} \sum_{i<j} J\left(E_i, E_j\right) \tag{29}$$

In the experiments, the top-k parameter used for Consistency is set to $k = 10$. A higher Consistency score indicates that the explanation method selects more similar explanatory units across repeated random runs.

Comprehensiveness and Sufficiency evaluate the faithfulness of the selected explanatory units by measuring changes in the black-box classifier's confidence (DeYoung et al. 2020). Let x denote the original document, c the class assigned to x, and $f_c(x)$ the classifier's confidence in class c. Let E_ρ denote the highest-ranked explanatory units selected at ratio ρ, $x \setminus E_\rho$ the document obtained after removing these units, and x_{E_ρ} the document retaining only these units.

Comprehensiveness is defined as:

$$\text{COMP}_\rho = f_c(x) - f_c(x \setminus E_\rho) \tag{30}$$

A higher COMP value indicates that removing the selected units causes a larger decrease in classification confidence and therefore that these units exert greater influence on the original classification.

Sufficiency is defined as:

$$\text{SUFF}_\rho = f_c(x) - f_c\left(x_{E_\rho}\right) \tag{31}$$

A lower SUFF value indicates that the selected units alone preserve more of the original classification confidence. Because the raw confidence difference is retained without clipping, SUFF may be negative when the classifier assigns higher confidence to the retained explanatory units than to the full document.

To reduce dependence on a single explanation size, COMP and SUFF are evaluated at the ratio set

$$R_S = \{0.01, 0.05, 0.10, 0.20, 0.50\} \tag{32}$$

and the final scores are averaged across these ratios:

$$\text{COMP} = \frac{1}{|R_S|} \sum_{\rho \in R_S} \text{COMP}_\rho \tag{33}$$

$$\mathrm{SUFF} = \frac{1}{|R_S|} \sum_{\rho \in R_S} \mathrm{SUFF}_\rho \tag{34}$$

In addition, LRA@k evaluates whether the selected important units align with rhetorically central legal roles. Roles such as Ratio of the Decision (ROD) and Ruling by Present Court (RPC) are assigned higher heuristic weights, while Fact receives a lower weight because it generally provides background information rather than doctrinal reasoning. Formally,

$$LRA@k = \frac{1}{k} \sum_{v=1}^{k} W\left(L\left(u_{(v)}\right)\right) \tag{35}$$

where $u_{(v)}$ is the v-th selected explanatory unit, $L(\cdot)$ denotes its rhetorical role, and $W(\cdot)$ is the corresponding role weight.

The weighting design in LRA is intentionally role-sensitive. Specifically, Ratio of the Decision (ROD) is given the highest weight (1.0), followed by Ruling by Present Court (RPC, 0.9), because these two roles most directly encode the court's reasoning process and final adjudicative outcome. Argument and Precedent are both assigned 0.8, reflecting their substantive contribution to legal justification, whereas Statute is assigned 0.7 as an explicit normative basis for judicial analysis. In contrast, Ruling by Lower Court (0.5) and Fact (0.3) receive lower weights because they are comparatively more contextual than dispositive in explaining the final classification. The complete role-specific weighting scheme is summarized in Table 5.

Table 5 Importance weights $W(r)$ of rhetorical roles in the LRA metric.

Rhetorical Role r	Weight $W(r)$
Ratio of the Decision, ROD	1.0
Ruling by Present Court, RPC	0.9
Argument	0.8
Precedent	0.8
Statute	0.7
Ruling by Lower Court	0.5
Fact	0.3

This weighting scheme is informed by jurisprudential accounts of ratio decidendi, which emphasize the legal propositions and material facts necessary to the court's decision (Goodhart 1930; MacCormick 1978; Cross and Harris 1991). In the present study, ROD and RPC are assigned higher weights as proxies for explicitly adjudicative reasoning and the final disposition. The lower weight assigned to Fact does not imply that facts are legally irrelevant; rather, it reflects the metric's emphasis on discourse roles that explicitly articulate the court's reasoning. Because the cited accounts do not prescribe the specific numerical values used in LRA@k, the weights reported in Table 5 are heuristic and should be interpreted as a role-based legal-salience proxy rather than as a direct measure of doctrinal correctness.

For Token LIME, each selected token inherits the rhetorical-role label of the sentence in which it occurs. Sentence-level units use their assigned sentence roles, while SPS-LIME segments use their shared segment roles.

4.8 Explanation Results on SCOTUS

Table 6 compares Token LIME, Sentence LIME, and SPS-LIME on SCOTUS. SPS-LIME achieves the highest mean score on both Consistency and Comprehensiveness, with scores of 0.7591 and 0.0859, respectively. These results indicate that segment-level perturbation based on rhetorical-role continuity shows higher average Consistency and identifies units whose removal more strongly affects model classifications. In other words, SPS-LIME provides numerically stronger results than token- or sentence-level baselines in capturing classifier-sensitive evidence on SCOTUS.

Table 6 Comparison of explanation quality for Token LIME, Sentence LIME, and SPS-LIME (SCOTUS).

Metric	Token LIME	Sentence LIME	SPS-LIME
Consistency↑	0.4502±0.2358	0.5545±0.2197	**0.7591±0.1826**
COMP↑	0.0559±0.1237	0.0631±0.1005	**0.0859±0.1473**
SUFF↓	0.1777±0.2041	**0.0168±0.1527**	0.0283±0.1389

For Sufficiency, however, Sentence LIME obtains the best score. One possible explanation is that some SCOTUS opinions contain highly discriminative issue-area cues within individual sentences. Verifying this hypothesis would require analysing the textual coverage and role distribution of the selected sentences.

4.9 Legal Role Alignment on SCOTUS

Table 7 presents LRA@k results for SCOTUS. Surprisingly, Token LIME achieves the highest LRA@k scores across all values of k, whereas SPS-LIME ranks second and Sentence LIME ranks last. At first glance, this could suggest that token-level explanations are better aligned with legally important roles. However, such a conclusion would be misleading if taken at face value.

Table 7 LRA@k comparison (SCOTUS)

Metric	Token LIME	Sentence LIME	SPS-LIME
LRA@3	**0.6103±0.1556**	0.4317±0.1425	0.4173±0.1277
LRA@5	**0.6099±0.1321**	0.4298±0.1143	0.4511±0.1110
LRA@10	**0.6071±0.0826**	0.4340±0.0951	0.4952±0.0963

One possible explanation is that Token LIME selects multiple lexical cues located within highly weighted rhetorical roles. However, the current quantitative results do not distinguish this possibility from other effects, including differences in textual coverage across explanation granularities.

4.10 Explanation Results on ECtHR

Table 8 shows the explanation-quality results on ECtHR. In this dataset, SPS-LIME achieves the highest mean Consistency and Comprehensiveness scores and the lowest mean Sufficiency score. Compared with Token LIME and Sentence LIME, SPS-LIME shows higher average Consistency and identifies segments whose removal produces the largest mean decrease in classification confidence. This indicates that, for violation/non-violation classification, grouping consecutive sentences with the same rhetorical role provides numerically stronger results for extracting decisive evidence.

Table 8 Comparison of explanation quality for Token LIME, Sentence LIME, and SPS-LIME (ECtHR).

Metric	Token LIME	Sentence LIME	SPS-LIME
Consistency↑	0.8583±0.2239	0.5667±0.2460	**0.9317±0.1667**
COMP↑	0.1938±0.3312	0.0781±0.1438	**0.3028±0.3502**
SUFF↓	0.1333±0.2717	0.0275±0.1188	**-0.0051±0.0791**

The negative mean SUFF score indicates that the classifier assigns slightly higher confidence to the selected segments alone than to the full document. This may reflect the concentration of classification evidence within the selected segments, although alternative explanations such as distribution shift or confidence variation cannot be excluded.

4.11 Legal Role Alignment on ECtHR

Table 9 reports LRA@k results on ECtHR. As in SCOTUS, Token LIME achieves the highest scores overall, SPS-LIME is a close second, and Sentence LIME performs worst. Notably, SPS-LIME approaches Token LIME very closely at $k = 10$, with LRA@10 values of 0.5021 and 0.5057, respectively. This suggests that once a slightly larger set of explanatory units is considered, SPS-LIME approaches the token-level baseline in LRA@10 while retaining role-continuous segments by construction.

Table 9 LRA@k comparison (ECtHR)

Metric	Token LIME	Sentence LIME	SPS-LIME
LRA@3	**0.5112±0.1620**	0.3753±0.1235	0.4552±0.1340
LRA@5	**0.5077±0.1431**	0.3868±0.1165	0.4835±0.1258
LRA@10	**0.5057±0.1311**	0.3954±0.1054	0.5021±0.1264

Taken together, the SCOTUS and ECtHR results reveal an important methodological tension. If legal alignment is judged only by role-weighted Top-k overlap, token-level explanations may appear superior. Yet if stability, semantic continuity, and deletion-based faithfulness are also taken into account, SPS-LIME becomes more compelling, especially on ECtHR. Therefore, no single metric is sufficient to characterize explanation quality in legal NLP; different metrics capture different aspects of what it means for an explanation to be useful, faithful, and legally meaningful.

4.12 Segment Granularity and Model Behavior

Table 10 summarizes the effect of varying MaxSegmentSize on the behaviour of SPS-LIME across our four evaluation metrics. In this experiment, MaxSegmentSize is defined as the maximum number of sentences allowed within a single semantic segment; for a fixed judgment, smaller values therefore induce a larger number of shorter segments, whereas larger values yield fewer but longer segments. When MaxSegmentSize is set to 20, SPS-LIME achieves the highest Consistency score. Comprehensiveness increases across the tested settings and reaches its highest value under the Unlimited condition. By contrast, Sufficiency and LRA@10 vary non-monotonically. The lowest Sufficiency score is obtained when MaxSegmentSize is 25, whereas the Unlimited setting obtains the highest LRA@10 score. These results indicate that segment granularity produces metric-specific trade-offs and that no single setting is uniformly optimal.

Table 10 Effect of MaxSegmentSize on SPS-LIME for SCOTUS.

Metric/Size	5	20	25	Unlimited
Consistency↑	0.7551	**0.8118**	0.7932	0.7591
COMP↑	0.0376	0.0505	0.0620	**0.0859**
SUFF↓	0.0643	0.0303	**0.0247**	0.0283
LRA@10	0.4365	0.3985	0.4215	**0.4952**

4.13 Discussion

The experiments reported in this section support three main observations. First, rhetorical-role information is useful but task-dependent. On SCOTUS, combining rhetorical-role and lexical features improves classification performance. On ECtHR, however, rhetorical-role information is associated with a stronger tendency toward majority-class decisions, which may indicate reliance on shortcut patterns. Second, SPS-LIME achieves the highest average Consistency and Comprehensiveness on both datasets and the lowest Sufficiency on ECtHR. On SCOTUS, however, Sentence LIME obtains a lower Sufficiency score, indicating a task-dependent trade-off between segment-level stability and compact retained rationales. Third, legal-role alignment does not necessarily improve together with stability or semantic coherence. Token-level explanations may obtain higher alignment scores by selecting isolated cues within highly weighted rhetorical roles without preserving complete legal reasoning units.

These tensions are consistent with prior work distinguishing classification performance from legal justification. Mumford et al. (2021) argue that explanations for legal machine-learning systems should relate model classifications to recognizable forms of legal argument rather than merely describe the statistical basis of a classifier. Xu et al. (2021) show that sentence position and legal-domain sentence embeddings improve the classification of issues, reasons, and conclusions, while sentence classification remains substantially more difficult in full-text decisions. The present findings extend this line of research by showing that, even when rhetorical roles are explicitly modeled, highly influential explanations may still concentrate on fact-heavy segments rather than segments labeled as Ratio of the Decision or Ruling by Present Court.

Beyond these observations, the MaxSegmentSize experiments show that SPS-LIME is sensitive to segment granularity. A MaxSegmentSize of 20 yields the highest Consistency, the Unlimited setting obtains the highest Comprehensiveness and LRA@10 scores, and a MaxSegmentSize of 25 produces the lowest Sufficiency score. Because the effects on Sufficiency and legal-role alignment are non-monotonic, no single segment size is uniformly optimal. Segment construction therefore involves trade-offs among explanation stability, deletion-based faithfulness, semantic continuity, and legal-role alignment.

More broadly, SPS-LIME can be situated within longstanding AI and Law research on the relationship between data-driven classification and explicit computational representations of legal reasoning (Bench-Capon et al. 2012). Related work further argues that explanations in legal AI should connect model outputs to recognizable argumentative and justificatory structures rather than provide only numerical feature attributions (Atkinson et al. 2020). By examining whether model-sensitive content corresponds to rhetorical roles, the present study provides an empirical analysis of this relationship.

Taken together, the findings indicate that explanations for legal AI should not be evaluated solely according to their effects on model classifications or their concentration within selected rhetorical roles. Explanation quality is multidimensional and includes faithfulness to classifier behavior, stability, semantic continuity, and task-specific legal-role relevance. SPS-LIME is therefore useful not only because it improves several quantitative measures, but also because it reveals potential mismatches between the content most influential to a classifier and the content identified as legally salient by the proposed role-based proxy.

5 Conclusion and Future Work

Perturbation-based explanation methods for text classification commonly operate on fixed units such as individual tokens or sentences, which can fragment discourse structure or produce semantically incomplete rationales in long documents. This work introduced Semantic-Preserved Segment LIME, or SPS-LIME, a model-agnostic framework that constructs variable-length explanation units by merging consecutive sentences assigned the same rhetorical role. SPS-LIME combines segment-level attribution with sentence-level localization, allowing discourse-continuous segments to be retained as explanation units while still identifying influential sentences within them.

On LexGLUE-SCOTUS and a document-level binary violation-classification task derived from the ECtHR corpus, SPS-LIME is compared with Token LIME and Sentence LIME, which use tokens and individual sentences as perturbation units, respectively. The results show that SPS-LIME achieves the highest mean Consistency and Comprehensiveness scores on both datasets. It also obtains the lowest Sufficiency score on ECtHR, whereas Sentence LIME performs best on Sufficiency for SCOTUS. The MaxSegmentSize analysis further demonstrates that segmentation granularity produces metric-specific trade-offs. A moderately constrained setting yields the highest Consistency, while the Unlimited setting achieves the highest Comprehensiveness and LRA@10. These findings indicate that explanation granularity should be treated as a

design parameter rather than assuming that finer explanation units are uniformly more faithful or informative.

The experiments further show that faithfulness to classifier behavior, semantic continuity, and legal-role alignment capture different properties of an explanation. Token LIME obtains higher LRA@k scores on both datasets, although one possible explanation is that individual tokens located within highly weighted rhetorical roles can receive high alignment scores without representing complete discourse units. The quantitative and qualitative analyses also indicate that classifier-sensitive SPS-LIME segments may be dominated by factual or contextual content rather than by Ratio of the Decision or Ruling by Present Court segments. These results provide evidence of a potential mismatch between model-sensitive content and the legal salience approximated by the proposed role-based evaluation.

The findings also expose limitations in current explanation evaluation. Deletion-based metrics measure model sensitivity to selected content but do not determine whether that content constitutes substantively appropriate legal reasoning. Similarly, LRA@k depends on heuristic role weights and should be interpreted as a role-based legal-salience proxy rather than as a substitute for expert legal assessment. The comparison across explanation granularities is also unit-relative, because equal numbers or proportions of tokens, sentences, and segments do not necessarily correspond to equal textual coverage.

Future work should incorporate expert-annotated rationales, evaluate explanation methods under matched textual-coverage budgets, examine alternative role-weighting and alignment formulations, and test robustness across rhetorical-role taggers, classifier architectures, and additional long-document datasets. Extending segment-based perturbation to other structured domains may also clarify whether discourse-aware explanation units provide similar benefits beyond legal text classification.

References

1. Ahmed, S.U., Ahmad, N.: A rhetorical role relatedness (RRR) framework for legal case brief generation. Procedia Computer Science 258, 3333–3344 (2025). https://doi.org/10.1016/j.procs.2025.04.591
2. Alvarez-Melis, D., Jaakkola, T.S.: On the robustness of interpretability methods. In: Proceedings of the ICML Workshop on Human Interpretability in Machine Learning, Stockholm, Sweden (2018); arXiv:1806.08049.
3. Ashley, K.D.: Artificial Intelligence and Legal Analytics: New Tools for Law Practice in the Digital Age. Cambridge University Press, Cambridge, UK (2017).
4. Atanasova, P., Simonsen, J.G., Lioma, C., Augenstein, I.: A diagnostic study of explainability techniques for text classification. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pp. 3256–3274. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.emnlp-main.263
5. Atkinson, K., Bench-Capon, T.J.M., Bollegala, D.: Explanation in AI and law: Past, present and future. Artificial Intelligence 289, 103387 (2020). https://doi.org/10.1016/j.artint.2020.103387
6. Aumiller, D., Almasian, S., Lackner, S., Gertz, M.: Structural text segmentation of legal documents. arXiv:2012.03619 (2020).

7. Bambroo, P., Adhikary, S., Bhattacharya, P., Chakraborty, A., Ghosh, S., Ghosh, K.: MARRO: Multi-headed attention for rhetorical role labeling in legal documents. Artificial Intelligence and Law (2025). https://doi.org/10.1007/s10506-025-09449-7
8. Belfathi, A., Hernandez, N., Monceaux, L., Bonnard, W., Lavissiere, M.C., Jacquin, C., Dufour, R.: Coupling local context and global semantic prototypes via a hierarchical architecture for rhetorical roles labeling. In: Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2986–3004. Association for Computational Linguistics, Rabat, Morocco (2026). https://doi.org/10.18653/v1/2026.eacl-long.137
9. Beltagy, I., Peters, M.E., Cohan, A.: Longformer: The long-document transformer. arXiv:2004.05150v1 (2020).
10. Bench-Capon, T., Araszkiewicz, M., Ashley, K.D., Atkinson, K., Bex, F., Karlsson, F., Prakken, H., Sartor, G., Wardeh, M., Wyner, A.: A history of AI and law in 50 papers: 25 years of the international conference on AI and law. Artificial Intelligence and Law 20(3), 215–319 (2012). https://doi.org/10.1007/s10506-012-9131-x
11. Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A.: DeepRhole: Deep learning for rhetorical role labeling of sentences in legal case documents. Artificial Intelligence and Law 31(1), 53–90 (2023). https://doi.org/10.1007/s10506-021-09304-5
12. Bhattacharya, P., Paul, S., Ghosh, K., Ghosh, S., Wyner, A.: Identification of rhetorical roles of sentences in Indian legal judgments. In: Legal Knowledge and Information Systems—JURIX 2019, pp. 3–12. IOS Press, Amsterdam, The Netherlands (2019). https://doi.org/10.3233/FAIA190301
13. Brożek, B., Furman, M., Jakubiec, M., Kucharzyk, B.: The black box problem revisited: Real and imaginary challenges for automated legal decision making. Artificial Intelligence and Law 32, 427–440 (2024). https://doi.org/10.1007/s10506-023-09356-9
14. Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., Androutsopoulos, I.: LEGAL-BERT: The muppets straight out of law school. In: Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898–2904. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.261
15. 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. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 4310–4330. Association for Computational Linguistics, Dublin, Ireland (2022). https://doi.org/10.18653/v1/2022.acl-long.297
16. Chalkidis, I., Androutsopoulos, I., Aletras, N.: Neural legal judgment prediction in English. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4317–4323. Association for Computational Linguistics, Florence, Italy (2019). https://doi.org/10.18653/v1/P19-1424
17. Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I., Malakasiotis, P.: Paragraph-level rationale extraction through regularization: A case study on European Court of Human Rights cases. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 226–241. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.22
18. Collenette, J., Atkinson, K., Bench-Capon, T.: Explainable AI tools for legal reasoning about cases: A study on the European Court of Human Rights. Artificial Intelligence 317, 103861 (2023). https://doi.org/10.1016/j.artint.2023.103861
19. Cross, R., Harris, J.W.: Precedent in English Law, 4th edn. Clarendon Press, Oxford, UK (1991).

20. de Arriba-Pérez, F., García-Méndez, S., González-Castaño, F.J., González-González, J.: Explainable machine learning multi-label classification of Spanish legal judgements. Journal of King Saud University—Computer and Information Sciences 34(10), 10180–10192 (2022). https://doi.org/10.1016/j.jksuci.2022.10.015
21. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, MN, USA (2019). https://doi.org/10.18653/v1/N19-1423
22. DeYoung, J., Jain, S., Rajani, N.F., Lehman, E., Xiong, C., Socher, R., Wallace, B.C.: ERASER: A benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.408
23. Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A.: Shortcut learning in deep neural networks. Nature Machine Intelligence 2(11), 665–673 (2020). https://doi.org/10.1038/s42256-020-00257-z
24. Goodhart, A.L.: Determining the ratio decidendi of a case. Yale Law Journal 40(2), 161–183 (1930).
25. Goodrich, P.: Rhetoric as jurisprudence: An introduction to the politics of legal language. Oxford Journal of Legal Studies 4(1), 88–121 (1984). https://doi.org/10.1093/ojls/4.1.88
26. Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S.R., Smith, N.A.: Annotation artifacts in natural language inference data. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 107–112. Association for Computational Linguistics, New Orleans, LA, USA (2018). https://doi.org/10.18653/v1/N18-2017
27. Hachey, B., Grover, C.: A rhetorical status classifier for legal text summarisation. In: Proceedings of the Workshop on Text Summarization Branches Out, pp. 35–42. Association for Computational Linguistics, Barcelona, Spain (2004).
28. Hachey, B., Grover, C.: Sequence modelling for sentence classification in a legal summarisation system. In: Proceedings of the 20th ACM Symposium on Applied Computing, pp. 292–296. Association for Computing Machinery, Santa Fe, NM, USA (2005). https://doi.org/10.1145/1066677.1066746
29. He, P., Liu, X., Gao, J., Chen, W.: DeBERTa: Decoding-enhanced BERT with disentangled attention. In: Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Event (2021).
30. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12(1), 55–67 (1970). https://doi.org/10.1080/00401706.1970.10488634
31. Jacovi, A., Goldberg, Y.: Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4198–4205. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.386
32. Kalamkar, P., Tiwari, A., Agarwal, A., Karn, S., Gupta, S., Raghavan, V., Modi, A.: Corpus for automatic structuring of legal documents. In: Proceedings of the 13th Language Resources and Evaluation Conference, pp. 4420–4429. European Language Resources Association, Marseille, France (2022).
33. Kolt, N., Caputo, N., Boeglin, J., O'Keefe, C., Bommasani, R., Casper, S., Cuéllar, M.-F., Feldman, N., Gabriel, I., Hadfield, G.K., Hammond, L., Henderson, P., Kasirzadeh, A.,

Lazar, S., Reuel, A., Wei, K., Zittrain, J.L., et al.: Legal alignment for safe and ethical AI. Transactions on Machine Learning Research (2026); arXiv:2601.04175.
34. Lafferty, J., McCallum, A., Pereira, F.C.N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: Proceedings of the 18th International Conference on Machine Learning, pp. 282–289. Morgan Kaufmann, Williamstown, MA, USA (2001).
35. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692 (2019).
36. Lundberg, S.M., Lee, S.-I.: A unified approach to interpreting model predictions. In: Advances in Neural Information Processing Systems 30, pp. 4765–4774 (2017).
37. MacCormick, N.: Legal Reasoning and Legal Theory. Clarendon Press, Oxford, UK (1978).
38. Malik, V., Sanjay, R., Guha, S.K., Hazarika, A., Nigam, S.K., Bhattacharya, A., Modi, A.: Semantic segmentation of legal documents via rhetorical roles. In: Proceedings of the Natural Legal Language Processing Workshop 2022, pp. 153–171. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (2022). https://doi.org/10.18653/v1/2022.nllp-1.13
39. Mamakas, D.: Bag of words BERT-based models for large legal document classification. B.Sc. thesis, Department of Informatics, Athens University of Economics and Business, Athens, Greece (2022).
40. Mumford, J., Atkinson, K., Bench-Capon, T.: Machine learning and legal argument. In: Proceedings of the 21st Workshop on Computational Models of Natural Argument, CEUR Workshop Proceedings, vol. 2937, pp. 47–56 (2021).
41. Namazov, M., Koref, T., Habernal, I.: Legal experts disagree with rationale extraction techniques for explaining ECtHR case outcome classification. arXiv:2601.12419 (2026).
42. Nigam, S.K., Dubey, T., Sharma, G., Shallum, N., Ghosh, K., Bhattacharya, A.: LegalSeg: Unlocking the structure of Indian legal judgments through rhetorical role classification. In: Findings of the Association for Computational Linguistics: NAACL 2025, pp. 1129–1144. Association for Computational Linguistics, Albuquerque, NM, USA (2025). https://doi.org/10.18653/v1/2025.findings-naacl.63
43. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011).
44. Ribeiro, M.T., Singh, S., Guestrin, C.: 'Why should I trust you?': Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1135–1144. Association for Computing Machinery, San Francisco, CA, USA (2016). https://doi.org/10.1145/2939672.2939778
45. Richmond, K.M., Muddamsetty, S.M., Gammeltoft-Hansen, T., Olsen, H.P., Moeslund, T.B.: Explainable AI and law: An evidential survey. Digital Society 3(1), 1 (2024). https://doi.org/10.1007/s44206-023-00081-z
46. Rong, Y., Leemann, T., Nguyen, T.-T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E.: Towards human-centered explainable AI: A survey of user studies for model explanations. IEEE Transactions on Pattern Analysis and Machine Intelligence 46(4), 2104–2122 (2024). https://doi.org/10.1109/TPAMI.2023.3331846
47. Rychener, Y., Renard, X., Seddah, D., Frossard, P., Detyniecki, M.: On the granularity of explanations in model agnostic NLP interpretability. In: Koprinska, I., et al. (eds.) Machine Learning and Principles and Practice of Knowledge Discovery in Databases: ECML PKDD 2022. Communications in Computer and Information Science, vol. 1752, pp. 498–512. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-23618-1_33

48. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing & Management 24(5), 513–523 (1988). https://doi.org/10.1016/0306-4573(88)90021-0
49. Santosh, T.Y.S.S., Isaia, A., Hong, S., Grabmair, M.: HiCuLR: Hierarchical curriculum learning for rhetorical role labeling of legal documents. In: Findings of the Association for Computational Linguistics: EMNLP 2024, pp. 7357–7364. Association for Computational Linguistics, Miami, FL, USA (2024). https://doi.org/10.18653/v1/2024.findings-emnlp.433
50. Saravanan, M., Ravindran, B., Raman, S.: Automatic identification of rhetorical roles using conditional random fields for legal document summarization. In: Proceedings of the Third International Joint Conference on Natural Language Processing (IJCNLP 2008), pp. 481–488. Asian Federation of Natural Language Processing, Hyderabad, India (2008).
51. Saravanan, M., Ravindran, B.: Identification of rhetorical roles for segmentation and summarization of Indian legal judgments. Artificial Intelligence and Law 18(1), 45–76 (2010). https://doi.org/10.1007/s10506-010-9087-7
52. Staliūnaitė, I., Valvoda, J., Satoh, K.: Comparative study of explainability methods for legal outcome prediction. In: Proceedings of the Natural Legal Language Processing Workshop 2024, pp. 243–258. Association for Computational Linguistics, Miami, FL, USA (2024). https://doi.org/10.18653/v1/2024.nllp-1.20
53. Valvoda, J., Cotterell, R.: Towards explainability in legal outcome prediction models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pp. 7269–7289. Association for Computational Linguistics, Mexico City, Mexico (2024). https://doi.org/10.18653/v1/2024.naacl-long.404
54. Valvoda, J., Pimentel, T., Stoehr, N., Cotterell, R., Teufel, S.: What about the precedent: An information-theoretic analysis of common law. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2275–2288. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.181
55. Waltl, B., Bonczek, G., Scepankova, E., Matthes, F.: Semantic types of legal norms in German laws: Classification and analysis using local linear explanations. Artificial Intelligence and Law 27(1), 43–71 (2019). https://doi.org/10.1007/s10506-018-9228-y
56. Wu, M.L., Liu, C.W.: Toward deployment-oriented long legal document classification: A segmentation-based framework for distributed legal evidence integration. IEEE Access 14, 72103–72116 (2026). https://doi.org/10.1109/ACCESS.2026.3691973
57. Xu, H., Savelka, J., Ashley, K.D.: Accounting for sentence position and legal domain sentence embedding in learning to classify case sentences. In: Bex, F., Henderson, P. (eds.) Legal Knowledge and Information Systems (JURIX 2021), pp. 33–42. IOS Press, Amsterdam, The Netherlands (2021).
58. Zafar, U., Wu, F.: Methodological challenges in explainable AI for fraud detection: A systematic literature review. Artificial Intelligence Review 59(4), 115 (2026). https://doi.org/10.1007/s10462-026-11516-7
59. Zaheer, M., Guruganesh, G., Dubey, K.A., Ainslie, J., Alberti, C., Ontañón, S., Pham, P., Ravula, A., Wang, Q., Yang, L., Ahmed, A.: Big Bird: Transformers for longer sequences. In: Advances in Neural Information Processing Systems 33, pp. 17283–17297 (2020).
60. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does NLP benefit legal system: A summary of legal artificial intelligence. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 5218–5230. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.466

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendix.docx