Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.074141

ARTICLE

Effective Token Masking Augmentation Using Term-Document Frequency for Language Model-Based Legal Case Classification

Ye-Chan Park[1], **Mohd Asyraf Zulkifley**[2], **Bong-Soo Sohn**[3] and **Jaesung Lee**[4,*]

[1]Department of Artificial Intelligence, Chung-Ang University, Seoul, 06974, Republic of Korea
[2]Department of Electrical, Electronic and Systems Engineering, Universiti Kebangsaan Malaysia, Bangi, 43600, Malaysia
[3]School of Computer Science and Engineering, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, 06974, Republic of Korea
[4]AI/ML Innovation Research Center, Chung-Ang University, 84 Heukseok-ro, Dongjak-gu, Seoul, 06974, Republic of Korea

*Corresponding Author: Jaesung Lee. Email: curseor@cau.ac.kr

Received: 03 October 2025; Accepted: 20 November 2025; Published: 10 February 2026

ABSTRACT: Legal case classification involves the categorization of legal documents into predefined categories, which facilitates legal information retrieval and case management. However, real-world legal datasets often suffer from class imbalances due to the uneven distribution of case types across legal domains. This leads to biased model performance, in the form of high accuracy for overrepresented categories and underperformance for minority classes. To address this issue, in this study, we propose a data augmentation method that masks unimportant terms within a document selectively while preserving key terms from the perspective of the legal domain. This approach enhances data diversity and improves the generalization capability of conventional models. Our experiments demonstrate consistent improvements achieved by the proposed augmentation strategy in terms of accuracy and F1 score across all models, validating the effectiveness of the proposed method in legal case classification.

KEYWORDS: Legal case classification; class imbalance; data augmentation; token masking; legal NLP

1 Introduction

Legal case classification is a key task in legal natural language processing (NLP), which aims to organize judicial documents in terms of their factual content and underlying legal principles. This enhances legal information retrieval, decision making, and analysis by enhancing access precision to relevant precedents and facilitating a systematic understanding of legal trends [1–3]. In practice, certain types of cases such as contract disputes and criminal theft are more frequent because of their prevalence in society. In contrast, other types such as antitrust and intellectual property cases are relatively rare. This class imbalance leads to biased model predictions, where models tend to favor majority classes and perform poorly for underrepresented categories, reducing their generalizability [4].

Data augmentation is a well-known strategy for improving model performance. Masking-based augmentation methods, which remove unimportant terms selectively while retaining key legal expressions, have garnered significant attention over recent years [5]. In such methods, the masking algorithm determines the candidate terms to be masked based on statistical salience, such as term frequency (TF) and inverse document frequency (IDF), which are popular concepts in text analysis. Notable methods include TF-IDF-based masking [6], which improves learning efficiency by removing less significant terms but may

inadvertently eliminate legally critical ones; difference masking [7], which refines term selection; and iterative mask filling (IMF) [8], which generates augmented documents using masked language models.

In legal documents, different sets of essential terms for case classification appear in specific cases. For example, "revocation" and "rescission" are commonly used in civil or administrative cases, yet they are typically absent from criminal law because they pertain to the legal validity of contracts, registrations, or administrative actions, which are central to civil proceedings but irrelevant in the context of criminal offenses. Importantly, these terms inherently exhibit low TF values and, consequently, low TF-IDF values, as they usually appear only once in the corresponding document. Thus, existing masking methods are likely to mask these essential terms, degrading case classification performance [9,10].

To address this issue, we propose a new masking method that selectively masks unimportant terms while preserving key legal expressions. Specifically, to protect essential legal terms that occur infrequently in the corresponding documents (i.e., with low TF values), the proposed algorithm uses term frequency-document frequency (TF-DF) instead of TF-IDF to assign masking likelihood to terms. The main contributions of this study are as follows:

- We proposed a TF-DF-based augmentation method tailored for legal text classification.
- We provided a comprehensive analysis of why TF-IDF-based masking fails in legal domains.
- We conducted an in-depth analysis of the actual masked outputs based on legal domain knowledge, offering practical insights for legal practitioners.

2 Related Work

Legal text classification is a fundamental task in legal NLP that facilitates information retrieval, case analysis, and judicial decision support. Early studies primarily relied on general NLP models without domain adaptation, which limited their ability to capture the nuances of legal terminology and reasoning. Foundational work such as Katz [11] highlighted the importance of quantitative approaches for legal prediction, underscoring the need for domain-adapted methods in legal NLP. In recent times, transformer-based architectures, e.g., Bidirectional Encoder Representations from Transformers (BERT) [1], and domain-specific variants, e.g., LegalBERT [2], have demonstrated significant performance improvements through pretraining on large-scale legal corpora that better reflect legal language. Further specialization has been achieved with encoders such as CaseLaw-BERT [12], which are tailored to judicial opinions and show improved performance on benchmark datasets like EURLEX and LexGLUE [3]. Nevertheless, despite these advances, transformer-based models continue to struggle with under-represented or low-resource legal categories, which motivates ongoing research into few-shot and zero-shot learning paradigms [4] as a means of enhancing generalization.

Data augmentation is a popular technique used to address data sparsity and improve model generalization, especially in scenarios with limited labeled data. Early work such as back-translation [13] illustrated the effectiveness of simple cross-lingual transformations for expanding training corpora. Rule-based techniques (e.g., back-translation and synonym replacement) are widely adopted in general NLP [14]. These approaches offer interpretability and ease of implementation, but often introduce noise or distort legal semantics when applied directly to legal texts, which typically exhibit rigid syntactic structures and formal language. Token-level strategies, such as term replacement, random swapping, and POS-guided deletion [5], aim to perturb input sequences without significantly altering their meaning. These methods are sensitive to the structural roles of tokens—especially in legal documents, where function words and modifiers may carry substantive legal implications. To address this, self-supervised approaches have been proposed. Contextual consistency training [15] encourages models to produce consistent outputs under augmented inputs, while manifold-based methods, such as SSMBA [16], perturb hidden representations to improve robustness against

out-of-distribution data. Although promising, these approaches are primarily validated on general NLP tasks (e.g., sentiment classification or QA) and may fail to account for the domain-specific precision required in legal applications. Lightweight schemes like AEDA (Easier Data Augmentation) [17] offer computational efficiency by randomly inserting punctuation or replacing characters. While such methods improve training diversity, they risk violating the syntactic and semantic constraints of legal text, leading to unnatural or misleading outputs.

Masking-based augmentation removes and replaces tokens selectively to facilitate pattern learning. TF-IDF masking has been shown effective in sentiment analysis [6], since it highlights discriminative words and suppresses redundant ones. Importantly, in the legal domain it is prone to semantic drift, as legally decisive terminology often appears infrequently and thus receives disproportionately low scores. Alternative strategies such as Different masking [7] and IMF [8] refine token selection and replacement, yet neither explicitly safeguards critical legal expressions nor addresses the persistent issue of class imbalance in legal text classification.

Domain-aware alternatives have also been proposed. LegalBERT is pretrained on legal corpora [10], enhancing representation quality, and TF-IDF representations occasionally outperform neural embeddings in legal classification [9]. Token deletion guided by corpus-passage frequency has shown promise in general-domain dense retrieval settings [18], motivating further exploration in domain-specific contexts such as legal NLP. Active learning pipelines further reduce annotation costs [19]. More recently, Ghosh et al. [20] introduced DALE, a selective masking approach tailored to legal language, and Kasthuriarachchy et al. [21] further refined this line of work through meaning-sensitive masking. Duffy et al. [22] examined a hybrid approach combining rule-based and generative augmentation in contract document classification, showing that simpler rules can sometimes outperform more complex generators. Sheik et al. [23] employed prompt engineering and pseudo-labeled data generation in overrule prediction, demonstrating that augmented models consistently outperformed non-augmented baselines and even surpassed few-shot GPT-3 in F1 score. Despite these advances, none of the existing methods integrates frequency-based token importance with the preservation of essential legal terms, both of which are crucial for robust legal text augmentation.

To address the limitations of prior methods, we propose a masking strategy that preserves legally salient terms while filtering peripheral ones. By leveraging corpus-level token statistics, our method enhances semantic fidelity and improves classification robustness in legal NLP tasks.

3 Proposed Method

In this section, we first explain the rationale behind the proposed masking strategy by comparing it with conventional TF-IDF-based masking strategies. Next, the procedure is introduced, and the details of the legal case dataset are presented. Finally, we describe the proposed masking method.

3.1 Rationale

Fig. 1 illustrates the masking tendencies of different strategies for different legal terms based on their TF and DF values. Existing augmentation methods often rely on TF-IDF to identify key legal terms. However, case-specific legal terms, e.g., “cancellation” and “inheritance,” tend to appear in only a narrow range of cases, thereby exhibiting low TF-IDF values. Consequently, they are unintentionally treated as unimportant and may be masked during augmentation, degrading classification performance. This mismatch highlights the need for a domain-aware weighting scheme that can distinguish between genuinely irrelevant terms and legally decisive yet sparsely distributed expressions. We contend that masking terms with low TF-IDF values is not necessarily the same as masking unimportant terms in the legal domain.

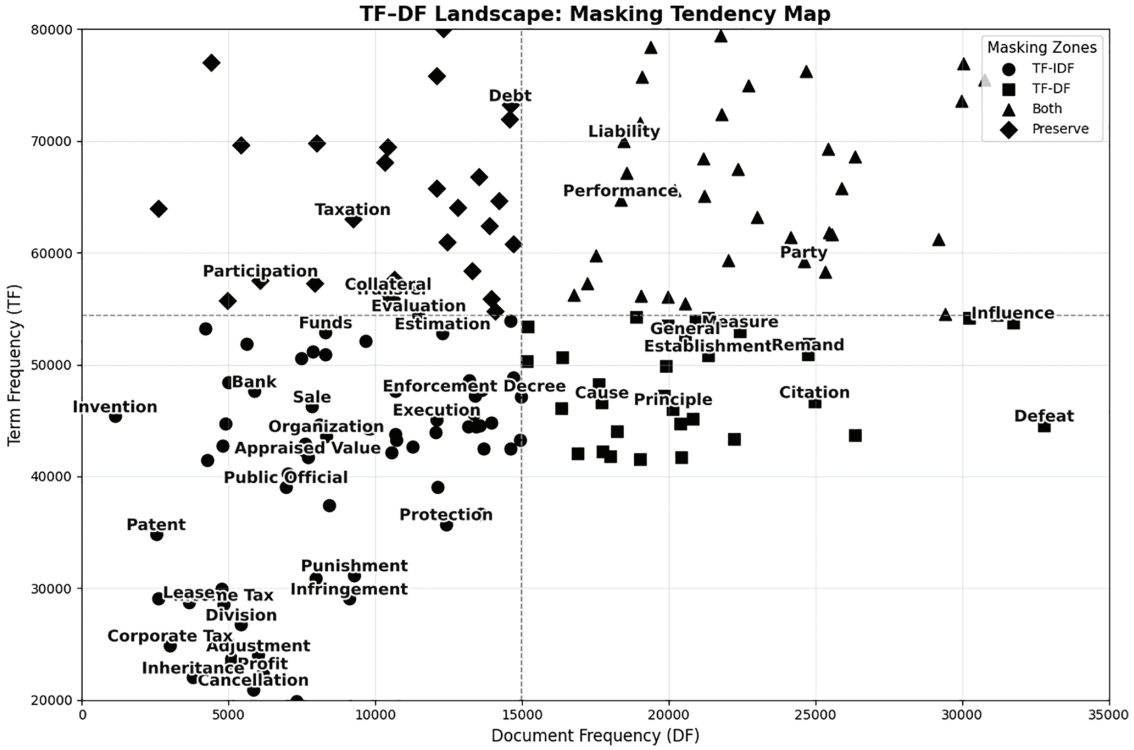

Figure 1: TF-DF masking tendency map. Each term is positioned based on its TF and DF within the corpus. Essential legal terms (e.g., "Cancellation," "Inheritance," "Public Official") appear in the circle zone, showing that TF-IDF masking may incorrectly remove them. The diamond zone denotes terms preserved by both strategies, while the triangle zone indicates generic terms masked by both

The x-axis represents document frequency (DF), whereas IDF decreases as DF increases (IDF = $\log \frac{N}{DF}$); therefore, the right-hand side of the plot indicates tokens with smaller IDF values, corresponding to more common terms within the corpus. Each marker in Fig. 1 corresponds to the TF-DF coordinate of an individual token, illustrating token-level masking outcomes rather than predefined regions: circles (•) indicating their likelihood to be masked by TF-IDF, squares (■) denote tokens masked only by the proposed TF-DF strategy, triangles (▲) represent tokens masked by both methods, and diamonds (⧫) mark tokens preserved by both. This clarification distinguishes token-specific masking behavior from the broader conceptual zones described in the rationale, ensuring consistent interpretation of Fig. 1.

- The term "Cancellation," located in the lower-left region of the TF-DF landscape (close to the x-axis, indicating both low term frequency and low document frequency), is essential for identifying cases involving administrative disposition cancellations, and facilitates their distinction from other administrative or tax-related matters.
- The term "Inheritance," also situated in the lower-left region near the x-axis, pertains to civil-law disputes over succession and property division. It appears in many judgments, but only a few times per document (low TF), placing it close to the x-axis. Therefore, TF-IDF tends to mask it despite it being a strong signal for the "inheritance-related civil case" category. Ideally, it should be retained.
- The term "Public Official," positioned in the lower-central region (moderate TF and relatively low DF), is an indicator of cases involving the responsibilities or duties of public officials and helps classify administrative disputes.

- The term “Cause,” appearing in the lower-right region (high TF and low-to-moderate DF),that appears in almost every opinion. Its high DF and moderate TF motivate TF-DF to down weight it; therefore, masking this token removes generic terminology that does not aid fine-grained classification.
- The term “Principle,” located slightly rightward in the lower-right region, is frequently used in abstract phrases such as “principle of good faith” or “principle of proportionality”. As it is ubiquitous across judgments and rarely decisive in the outcome, TF-DF correctly identifies it as a low–salience token to be masked.
- The term “Remand,” placed at the far right of the lower region (high TF and relatively low DF), signals procedural posture—namely, when an appellate court sends a case back to a lower court. Although it can dominate term counts in Appellate opinions, it conveys little substantive information about the underlying legal issue; therefore, masking it prevents the model from relying on procedural cues rather than topic-specific content.

Fig. 1 reveals a critical drawback of the TF-IDF strategy—many domain-specific and case-specific terms, e.g., “Cancellation,” “Inheritance,” and “Public Official” are located in the lower-left area (low TF and low DF), indicating their likelihood to be masked by TF-IDF. Although they appear infrequently across the corpus, these terms are essential for determining the legal context of specific cases. In contrast, the proposed TF-DF-based masking exhibits improved sensitivity to such legal relevance by preserving these terms. This figure supports the rationale behind the proposed TF-DF-based masking strategy and highlights its difference from existing TF-IDF-based preservation, demonstrating its ability to differentiate legally significant terms from generic ones. This figure substantiates the rationale of the proposed TF-DF approach and highlights its ability to distinguish legally salient tokens from generic ones, ensuring that data augmentation does not distort the core legal reasoning in case texts.

3.2 Data Preparation

Data preparation in this study follows a systematic multi-stage pipeline designed to transform raw judicial texts into a standardized format suitable for classification, as illustrated in Fig. 2. The dataset is derived from South Korean court rulings, originally provided by the Korean Ministry of Government Legislation via Law Open Data (https://open.law.go.kr, accessed on 18 November 2025), and consists of 87,160 legal cases spanning more than seven decades, from 13 January 1952 to 29 February 2024. This long temporal coverage ensures that the dataset captures the evolution of judicial language and legal reasoning in Korea, providing a valuable resource for both historical and contemporary analyses.

Table 1 presents the number of cases for each class along with descriptive statistics, including quartiles, means, and Std of the token lengths per case. In this study, we denote the Intellectual Property Law (IP Property) for brevity. This category encompasses legal disputes concerning patents, trademarks, copyrights, and design rights. A substantial class imbalance is observed: while the largest class (Civil Law) contains 39,830 cases, the smallest category has fewer than 1273 cases, reflecting the uneven distribution of case types in real-world judicial practice and making it challenging for models to learn minority categories effectively.

A primary challenge in preparing this dataset lies in the accurate separation of judicial decisions from legal reasoning, particularly in cases involving complex or multi-layered arguments. To address this, case facts, claims, and judicial decisions were extracted selectively using custom Python scripts, which systematically remove irrelevant metadata such as judge names, court divisions, and procedural annotations. This ensures that the processed text focuses on substantive legal content. The process begins with data extraction from court archives and relevant sources, followed by preprocessing steps that filter out extraneous information and personal identifiers to maintain textual consistency. The cleaned documents are then

organized and assigned to their corresponding classification categories, resulting in a corpus that is both structured and legally interpretable.

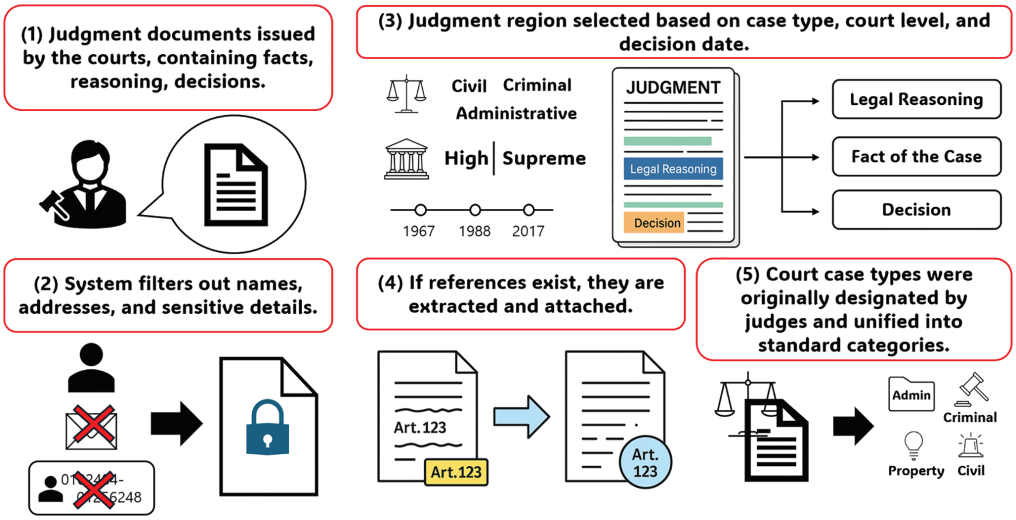

Figure 2: Overview of the legal judgment preprocessing and labeling pipeline, from raw court documents to standardized classification. The process includes personal information masking, legal reference extraction, segment selection, and global case type unification

Table 1: Dataset characteristics across legal categories. Q1, Median, and Q3 represent the 25th, 50th, and 75th percentiles of the number of tokens per case. Mean and standard deviation (Std) are also reported

Law class	Cases	Q1 (25%)	Median	Q3 (75%)	Max	Mean ± Std
Civil	39,830	316	570	1000	29,976	845.30 ± 989.64
Criminal	20,454	228	443	858	111,734	931.65 ± 2501.44
Administrative	12,660	318	569	1018	42,415	860.93 ± 1071.01
Taxation	9656	264	445	772	26,339	668.09 ± 816.49
Intellectual property	3287	251	384	668.5	16,567	643.36 ± 873.08
Family	1273	214	392.5	772	7835	620.45 ± 674.98
Total	87,160	281	513	934	111,734	838.18 ± 1505.41

Given the hybrid nature of the South Korean legal system, which integrates statutory law from civil law traditions with precedent-based reasoning from common law, the dataset requires meticulous structuring [24]. Legal phrase identification is performed to accurately segment case decisions into linguistically meaningful units. This is particularly important in Korean, an agglutinative language where grammatical particles and suffixes convey critical semantic cues. To ensure ethical compliance, personally identifiable information (e.g., party names, addresses, and references to individuals) is systematically removed, thereby safeguarding privacy while preserving the essential content required for downstream NLP tasks.

Unlike existing datasets [25], which primarily focus on criminal and civil law, the proposed dataset encompasses a broader spectrum of legal domains, including family law, intellectual property law, taxation, and administrative law. This broader coverage not only improves the representativeness of the dataset but also

enables the training of models capable of handling the diversity of real-world legal cases. Table 2 summarizes the outputs at each stage of the dataset refinement pipeline.

Table 2: Example of the output at each processing stage (The content of the legal document is shortened for clarity.)

Stage	Output example
Original	Plaintiff: Hong Gil-dong, Defendant: Kim Cheol-soo. Judgment Outline: The plaintiff's claim is dismissed. Supreme Court Justice OO, Judge OO, Presiding Judge.
Case filtering	The case interprets law under supplementary provisions (Article 23(1), Labor Standards Act).
Add legal ref	Key Laws: Labor Standards Act 23 Provision: Dismissal is prohibited without just cause.
Labeling	Detailed Categories: Wrongful Dismissal Claim. Referenced Cases: 2010Da98765, 2015Da12345. Text: Plaintiff's claim is dismissed for lack of just cause.

3.3 Proposed Masking Method

The proposed augmentation method aims to preserve legally important expressions, while introducing corpus-aware variability via probabilistic masking. Instead of relying on uniform or randomly applied masking, we leverage corpus-level statistics to determine the tokens that should be replaced by `[MASK]` symbols. The full masking algorithm is described in Algorithm 1. The algorithm begins by creating an empty container $\mathcal{D}'$ to store the augmented corpus, thereby establishing a dedicated repository for masked documents that will be generated in the subsequent steps (Line 1). This initialization step, though seemingly simple, is crucial in ensuring that the augmented data are systematically collected and preserved in a manner that is completely separable from the original corpus, thus preventing unintended data leakage or overwriting.

Algorithm 1: Proposed masking method.

Require: Dataset $\mathcal{D}$, tokenizer $\mathcal{T}$, masking scale α
Ensure: Augmented dataset $\mathcal{D}'$
1: $\mathcal{D}' \leftarrow \varnothing$ ▷ initialize empty container
2: Tokenize every document in $\mathcal{D}$ using $\mathcal{T}$
3: Compute the document frequency $\mathrm{df}(\omega)$ for each token ω
4: **for all** document $d \in \mathcal{D}$ **do**
5: Compute $\mathrm{tf}(\omega)$ for all tokens in d
6: $w(\omega) = \mathrm{tf}(\omega) \cdot \log(1 + \mathrm{df}(\omega))$ ▷ raw importance score
7: $\tilde{w}(\omega) = \dfrac{w(\omega) - \min(w)}{\max(w) - \min(w) + \epsilon}$ ▷ min–max normalization
8: $d' \leftarrow d$ ▷ create a working copy
9: **for all** token ω in d' **then**
10: Draw $r \sim \mathcal{U}(0,1)$
11: **if** $r \leq \alpha \cdot \tilde{w}(\omega)$ **then**
12: Replace ω with `[mask]`

(Continued)

Algorithm 1 (continued)

13: **end if**
14: **end for**
15: $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{d'\}$ ▷ append masked document
16: **end for**
return $\mathcal{D}'$

Next, every legal document in the corpus is tokenized using the KLUE/BERT WordPiece tokenizer, which is designed to segment words into subword units suitable for transformer-based language models (Line 2). This tokenization is not merely a mechanical preprocessing step but an essential foundation for frequency-based analysis, since subword segmentation captures rare and morphologically complex legal terms more effectively than word-level tokenization. Once tokenized, the document frequency $\mathrm{df}(\omega)$ of each token ω is computed at the corpus level (Line 3). Here, $\mathrm{tf}(\omega)$ denotes the number of occurrences of ω within a document, and $\mathrm{df}(\omega)$ the number of documents containing ω. These corpus-level statistics play a pivotal role in guiding subsequent masking decisions, because they reveal how widely distributed each token is across documents rather than just within a single text.

For each individual document $d \in \mathcal{D}$, the algorithm then computes the term frequency $\mathrm{tf}(\omega)$ for all tokens and assigns an importance score to each token based on a TF-DF formulation:

$$w(\omega) = \mathrm{tf}(\omega) \cdot \log\big(1 + \mathrm{df}(\omega)\big). \tag{1}$$

The logarithmic smoothing prevents excessively large df values from dominating the weighting scheme, thereby avoiding a situation where frequent but uninformative tokens (e.g., procedural markers such as "submitted," "record," or "hearing") overwhelm more discriminative but less frequent terms. It yields smoother scaling across large corpora and reduces sensitivity to corpus size, maintaining consistent importance estimation across datasets. Adding 1 inside the logarithm ensures numerical stability by avoiding undefined values when $df(\omega) = 0$. This design mirrors the stabilizing role of the logarithmic component in TF-IDF while reversing its intent—to emphasize legally meaningful expressions that recur across multiple cases rather than penalizing them.

To ensure comparability and stable scaling across tokens, the scores are subsequently normalized using min–max scaling:

$$\tilde{w}(\omega) = \frac{w(\omega) - \min(w)}{\max(w) - \min(w) + \epsilon}, \tag{2}$$

where ϵ is a small constant introduced for numerical stability. This normalization compresses all weights into the interval $[0,1]$, thus allowing them to be directly interpreted as probabilistic scaling factors for masking (Lines 4–7). Subsequently, for each document d, a copy d' is created to preserve the original text (Line 8). This duplication ensures that the original legal record remains intact for reference and evaluation, while all augmentation operations are confined to the copy. For each token ω in d', the algorithm samples a random value $r \sim \mathcal{U}(0,1)$ and applies the masking rule: if $r \leq \alpha, \tilde{w}(\omega)$, where α is a user-defined masking intensity parameter, then the token is replaced with `[MASK]` (Lines 9–13). This stochastic mechanism introduces controlled randomness into the augmentation process. By linking the masking probability directly to $\tilde{w}(\omega)$, the algorithm biases masking toward less-informative and frequently occurring tokens, while simultaneously lowering the likelihood of masking legally significant expressions such as "trust Asset" or "unjust dismissal."

A notable advantage of the stochastic masking strategy is its ability to reduce redundancy in the augmented corpus. Deterministic masking would consistently replace the same tokens across documents,

resulting in a less diverse dataset [15]. In contrast, the stochastic rule introduces variability between augmented instances, enriching the training distribution with multiple plausible variants of the same document. In this study, we empirically set the masking intensity parameter $\alpha = 0.2$, following prior works [17,20,26] that demonstrated its effectiveness in balancing coverage and diversity in stochastic masking and a sensitivity analysis was performed to validate the stability of this choice.

Once all tokens in a document have been processed under this stochastic masking regime, the resulting augmented document d' is appended to the container $\mathcal{D}'$ (Line 14). This process repeats for every document in the input corpus, gradually populating $\mathcal{D}'$ with augmented versions that maintain the core semantic and legal reasoning of the originals while discarding extraneous information. After all documents have been processed, the fully constructed augmented corpus $\mathcal{D}'$ is returned as output (Lines 15–16). This corpus serves as the foundation for downstream training, offering a richer and more balanced dataset for classification tasks.

Through stochastic masking guided by TF-DF weighting, the proposed method suppresses peripheral terms such as dates or procedural phrases while preserving legally decisive expressions, thereby operationalizing the algorithmic rationale described above in a concrete and systematic manner. Fig. 3 illustrates the overall workflow of this masking process in greater detail, highlighting how corpus-level statistics are used to determine token salience and guide the replacement of low-importance terms during augmentation.

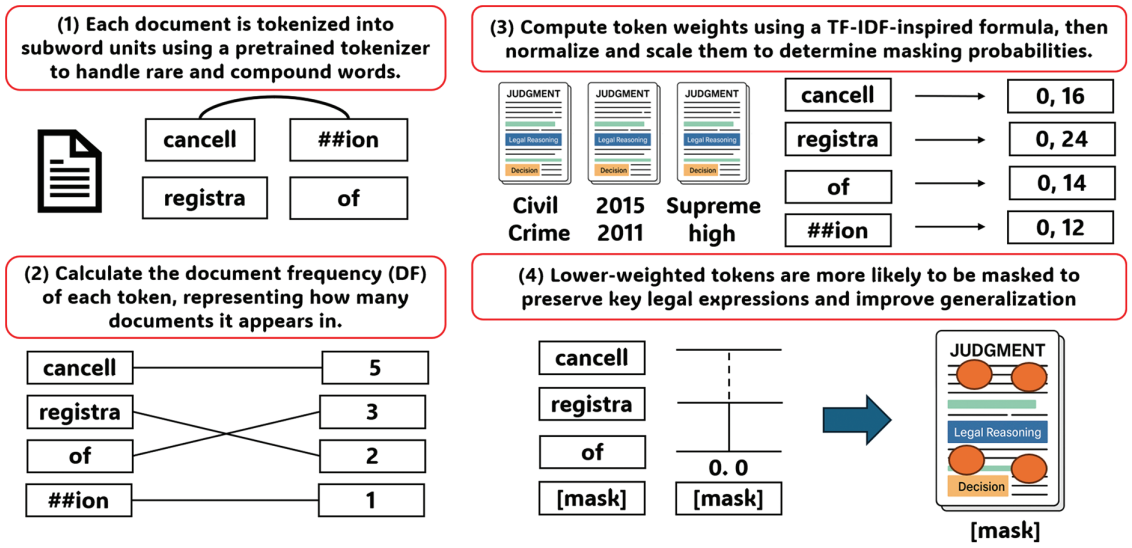

Figure 3: Overview of proposed augmentation process. The pipeline includes tokenization, document frequency calculation, weight computation, and selective masking

Furthermore, Table 3 presents illustrative examples of masked outputs, demonstrating that high-frequency but legally irrelevant expressions—such as dates, specific object names, procedural details, or other peripheral descriptors frequently appearing in judicial texts—are effectively suppressed, whereas pivotal legal terms remain intact, thereby preserving the core semantic structure required for correct legal interpretation. These observations highlight the importance of maintaining semantic fidelity in legal text augmentation, as even minor hallucinations can compromise downstream classification and retrieval tasks by subtly altering the factual framing or doctrinal meaning of a case.

Table 3: Examples of masked terms across different categories

Type	Augmentation (Before)	Augmentation (After)
Date, Time, Monetary values	2017, 1981. 11. 16	The wedding ceremony on May [MASK], 2010
Specific entities, Individuals	(with a pencil sharpener)	(with a [MASK] sharpener)
Legal procedural details	(Tax notice for revocation of seizure disposition)	(Tax notice for revocation of [MASK] disposition)

Table 4 shows that while AEDA introduces only minor structural noise, DALE fundamentally alters the factual and legal nature of the case itself. Such distortions illustrate why generative augmentation methods are unsuitable for high-fidelity legal datasets, where even subtle lexical substitutions can invert judicial meaning.

Table 4: Examples of semantic and structural distortions caused by AEDA [17] and DALE [20] in legal judgments

Method	Description and Example
TF-DF masking (Proposed)	*Original:* "The defendant shall bear the litigation costs." *Augmented:* "The [MASK] shall bear the litigation costs."
AEDA	*Original:* "According to Article 54-2 Paragraph 1 *Augmented:* "According to Article 54-2, Paragraph 1
DALE	*Original:* "The court annuls the defendant's **damages disposition**." *Generated:* "The court annuls the defendant's **information disclosure refusal**."

4 Experimental Results

This section presents the experimental validation of the effectiveness of the proposed method in terms of legal text classification performance. The performances of classification models are assessed, with and without data augmentation.

4.1 Experimental Settings

For evaluation, the dataset is split into three subsets: 60% for training, 20% for validation, and 20% for testing. To mitigate the imbalance problem, balanced augmentation [27] is applied, adjusting each category to match the size of the largest class rather than simply oversampling smaller ones. Balanced augmentation was applied after generating augmented samples through the TF-DF masking procedure. For each category, the number of cases was adjusted to match the largest class by adding non-redundant masked documents, thereby balancing the training distribution without simple duplication. In addition, balanced augmentation was applied solely to equalize the number of samples across legal categories, without modifying the textual content of any document. This setting inherently isolates the contribution of each augmentation strategy. All augmentations were evaluated under identical conditions using the same balanced dataset and hyperparameter settings. Therefore, any observed performance differences arise solely from the augmentation strategy itself rather than from variations in training data or optimization.

To ensure fair comparison, we evaluated four configurations. The 'No Augmentation' baseline denotes training on the original unaltered dataset without any augmentation, serving as a reference for evaluating

the contribution of each augmentation method. And a POS Deletion baseline that randomly removes part-of-speech–based tokens, a TF-IDF Masking baseline that masks tokens according to their inverse document frequency, and the proposed TF-DF Masking method. This strategy ensures that underrepresented categories are not disproportionately overlooked. Fig. 4 illustrates the distributions before and after augmentation.

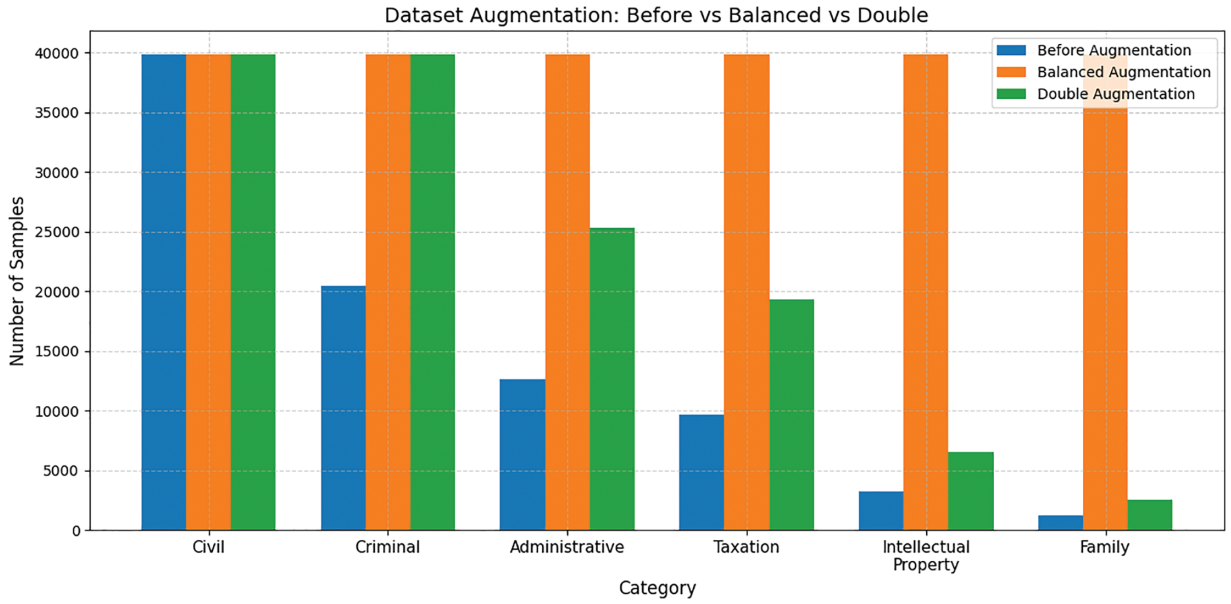

Figure 4: Number of samples per category before and after augmentation. Augmentation is applied to balance all categories to match the largest class, rather than simply duplicating smaller ones

Transformer-based models fine-tuned for legal case classification are considered. The models are trained using an AdamW optimizer with hyperparameters $\beta_1 = 0.9$; $\beta_2 = 0.999$; weight decay = 0.01; initial learning rate = 1×10^{-5}; adjusted via a linear scheduler; the number of epoch = 20; and batch size = 16. Each experiment is repeated 10 times to ensure robust and reliable evaluation. Given the four basic statistics, true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), for classification tasks, we evaluate model performance by calculating accuracy and F1 score, which are two widely used metrics in text classification tasks. Accuracy measures the proportion of correctly classified cases relative to all legal categories and is defined as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \tag{3}$$

where TP, TN, FP, and FN denote true positives, true negatives, false positives, and false negatives, respectively.

The F1 score, which represents the harmonic mean of precision $\left(\frac{TP}{TP+FP}\right)$ and recall $\left(\frac{TP}{TP+FN}\right)$, is defined as

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{4}$$

This metric provides a balanced evaluation of classification performance by combining both precision and recall.

4.2 Performance Comparison

We evaluated the effectiveness of the proposed TF-DF–based masking augmentation method using two representative models, BERT [1] and LegalBERT [2]. Tables 5 and 6 summarize the results in terms of accuracy and macro F1. Paired t-tests conducted on accuracy across ten runs confirmed statistically significant improvements for all models ($p < 0.01$). Specifically, LegalBERT ($p \approx 0.007$), and BERT ($p < 0.0001$) all showed meaningful accuracy gains following TF-DF masking. For LegalBERT, the proposed method achieved an accuracy of 0.8551, outperforming TF-IDF (0.8384), POS-based deletion (0.7775), and no augmentation (0.7436). Similarly, the macro F1 score reached 0.8453, showing a clear improvement over TF-IDF (0.7916), POS (0.6913), and no augmentation (0.5952). For BERT, the proposed method also gave the best performance, with an accuracy of 0.9704 and a macro F1 score of 0.9640. The proposed augmentation outperformed TF-IDF and POS-based methods across both models.

Table 5: Comparison of accuracy ± standard deviation under augmentation methods: proposed method, TF-IDF, POS deletion, and no augmentation

Model	Proposed	TF-IDF	POS	No augmentation
LegalBERT	0.8551 ± 0.0090	0.8384 ± 0.0103	0.7775 ± 0.0128	0.7436 ± 0.0091
BERT	0.9704 ± 0.0028	0.9591 ± 0.0007	0.9643 ± 0.0013	0.9203 ± 0.0012

The macro F1 improvements observed in Table 6 are primarily attributable to category-level gains highlighted in Table 7. Most notably, the Administrative Law category showed substantial improvement with the proposed method (0.9013 ± 0.0100) compared to TF-IDF (0.8171 ± 0.0030). Given that Administrative Law represents a large portion of the dataset, this gain had an outsized effect on the overall macro F1. In addition, the Family Law category—characterized by cultural specificity and nuanced linguistic expressions—benefited from TF-DF masking, improving from 0.7795 ± 0.0390 (TF-IDF) to 0.8990 ± 0.0076. This demonstrates that the proposed method effectively preserves contextually critical tokens (e.g., kinship or familial roles) that are often decisive in classification but may be indiscriminately masked under TF-IDF. Similarly, the Taxation category also exhibited meaningful gains (0.8936 ± 0.0149 → 0.9384 ± 0.0047).

Table 6: Comparison of macro F1 scores ± standard deviation under augmentation methods: proposed method, TF-IDF, POS deletion, and no augmentation

Model	Proposed	TF-IDF	POS	No augmentation
LegalBERT	0.8453 ± 0.0100	0.7916 ± 0.0155	0.6913 ± 0.0198	0.5952 ± 0.0327
BERT	0.9640 ± 0.0031	0.9475 ± 0.0017	0.9530 ± 0.0034	0.8973 ± 0.0023

Table 7: Comparison of F1 score ± standard deviation between the proposed method and TF-IDF-based masking for each category using BERT model-based classification

Category	Patterns	Proposed	TF-IDF
Civil	23,898	0.9626 ± 0.0040	0.9306 ± 0.0005
Criminal	12,272	0.9921 ± 0.0015	0.9810 ± 0.0010
Administrative	7596	0.9013 ± 0.0100	0.8171 ± 0.0030
Taxation	5794	0.9384 ± 0.0047	0.8936 ± 0.0149

(Continued)

Table 7 (continued)

Category	Patterns	Proposed	TF-IDF
IP Property	1972	0.9853 ± 0.0055	0.9663 ± 0.0022
Family	763	0.8990 ± 0.0076	0.7795 ± 0.0390

Table 8 presents a qualitative comparison of classification results obtained using different masking strategies. This comparison illustrates how the proposed strategy enhances classification reliability by selectively masking peripheral expressions while retaining critical legal terms, thus maintaining semantic fidelity in the augmented corpus. Each block of three rows corresponds to one case: the original judgment text, the version augmented with the proposed TF-DF masking, and the version augmented with TF-IDF masking. In the "Sentence" column, tokens surrounded by brackets (e.g., [**word**]) indicate the terms that would have been masked during augmentation according to each strategy. The last column reports the ground-truth label and the prediction produced by LegalBERT when trained with the corresponding input. As shown in Cases 1, 3, the proposed masking strategy preserves legally decisive terms such as "seizure invalid", "gift tax", enabling the model to predict the correct category. In contrast, TF-IDF masking often removes these essential expressions, which leads to semantic drift and incorrect predictions (e.g., misclassifying taxation as civil law or family law). When TF-IDF masking eliminates key tax-related legal terms such as 'tax', 'seizure', and 'trust Asset', the decisive linguistic markers necessary to situate the dispute within the domain of taxation law are lost. With the fiscal and administrative context removed, the residual sentence can be interpreted merely as a conflict concerning property ownership or possession. Consequently, the model fails to recognize the case as a taxation dispute and instead misclassifies it as a matter falling within the scope of general civil law, particularly property rights disputes.

Table 8: Comparison of legal case classification results obtained using different masking strategies. Sentences with [word] indicate tokens that would have been masked during training if selected by the corresponding strategy. The proposed masking strategy retains critical legal expressions, allowing LegalBERT to predict correctly, while TF-IDF masking often removes essential terms

No.	Version	Sentence (with [word])	Ground truth predicted result
1	Original	The tax authority seized property based on unpaid taxes, but the property was trust asset of the plaintiff, making the seizure invalid.	Taxation
	Proposed + LegalBERT	The tax authority seized [**property**] based on unpaid taxes, but the property was trust asset of the plaintiff, making the seizure invalid.	Taxation
	TF-IDF + LegalBERT	The tax authority seized [**property**] based on unpaid [**taxes**], but the property was [**trust asset**] of the plaintiff, making the [**seizure invalid**].	Civil law
	Original	The defendant unlawfully entered the victim's house at night with a pencil sharpener and stole valuable items.	Criminal law

(Continued)

Table 8 (continued)

No.	Version	Sentence (with [word])	Ground truth predicted result
2	Proposed + LegalBERT	The defendant unlawfully entered the [**house**] at night with a [**pencil**] sharpener and committed theft of valuable items.	Criminal law
	TF-IDF + LegalBERT	The [**defendant**] unlawfully entered the [**house**] at night with a [**pencil**] sharpener and [**committed theft**] of valuable items.	Administrative
3	Original	The tax office imposed gift taxes on family members, but some charges were revoked because the recipients were minors.	Taxation
	Proposed + LegalBERT	The tax office imposed [**gift tax**] on family members, but some charges were revoked because recipients were [**minors**].	Taxation
	TF-IDF + LegalBERT	The tax office imposed [**gift tax**] on [**family members**], but some charges were revoked because recipients were [**minors**].	Civil law
4	Original	The tenant failed to pay rent for two months, so the landlord terminated the lease contract and claimed delivery of the building.	Civil law
	Proposed + LegalBERT	The tenant failed to pay rent [**for two months**], so the landlord [**terminated**] the lease [**contract**] and claimed [**building delivery**].	Civil law
	TF-IDF + LegalBERT	The [**tenant**] failed to pay [**rent**] for two months, so the [**landlord**] [**terminated**] the [**lease**] contract and claimed [**building delivery**].	Administrative

By contrast, the proposed masking strategy preserves critical legal expressions such as 'seizure', 'trust Asset', and 'invalid', thereby maintaining the fiscal and administrative character of the dispute. Even though some peripheral expressions are masked, the presence of these domain-specific tokens enables the model to correctly identify the case as involving the validity of a tax levy, specifically the annulment of a taxation disposition. Taken together, the substantial improvement in Administrative Law, along with the enhanced handling of culturally specific Family Law cases and the robust recognition of domain-critical terms in Taxation, collectively explain the consistent macro F1 gains of the proposed method across both LegalBERT and BERT.

To assess parameter robustness, we varied the masking intensity parameter α across four levels: 0.05, 0.1, 0.2, and 0.3. As shown in Fig. 5, the F1 scores remained highly consistent across settings, demonstrating that the proposed TF-DF masking is insensitive to small perturbations in α. Both BERT and LegalBERT achieved their highest performance at $\alpha = 0.2$, suggesting that a moderate masking intensity provides an optimal balance between lexical diversity and semantic retention.

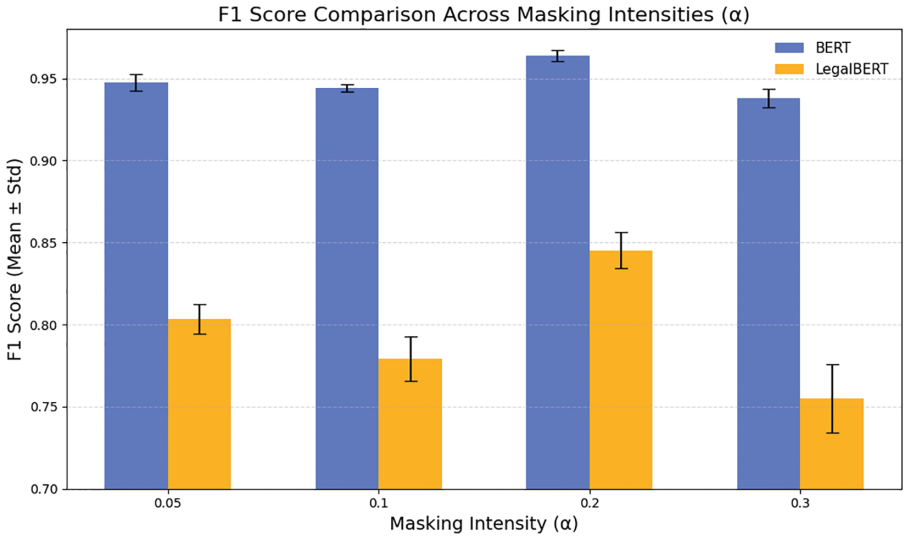

Figure 5: F1 score comparison across different masking intensities (α) for BERT and LegalBERT. The performance remained stable across $\alpha \in \{0.05, 0.1, 0.2, 0.3\}$

5 Discussion

As the size of the training corpus increases, the marginal benefit of data augmentation naturally diminishes, since larger datasets already encompass diverse lexical and syntactic patterns. TF-DF remains particularly useful for low-resource categories or specialized sub-domains where data imbalance continues to limit model generalization.

In terms of computational complexity, the proposed TF-DF masking operates in linear time with respect to the corpus size, requiring only a single pass to compute term and document frequencies, i.e., $O(N \times L)$ where N and L denote the number of documents and the average number of tokens per document. By contrast, generative augmentation frameworks such as DALE or Meaning-Sensitive Masking require repeated model inference for each masked token, resulting in $O\ (N \times L \times M)$ complexity, where M represents the cost of a forward pass through a large language model. Consequently, TF-DF achieves comparable semantic fidelity with approximately 3–5× lower preprocessing time while avoiding additional GPU-based fine-tuning.

Although the TF-DF framework effectively captures corpus-level token importance, its reliance on statistical weighting may reduce stability when applied to extremely short legal texts such as claims or briefs, where term occurrences are sparse. In such cases, contextual or embedding-based weighting could complement TF-DF by providing semantic cues independent of token frequency. Moreover, proposed TF-DF masking relies solely on corpus-level token and document statistics rather than language-specific lexical features, it is inherently language-agnostic and can be applied to legal corpora across different jurisdictions. This design enables the method to generalize beyond Korean texts without requiring additional linguistic adaptation. In addition, we verified its applicability on the LEGAR [3] English legal dataset, where it achieved an F1 improvement from 0.8555 ± 0.0045 to 0.8629 ± 0.0064, further confirming that the method generalizes well across different jurisdictions and linguistic contexts.

Regarding tokenizer variations, our study primarily relied on the WordPiece tokenizer; we acknowledge this as a limitation and suggest that future research explore the impact of alternative tokenization strategies on masking behavior across languages. Future work could integrate embedding-based or contextual weighting to improve performance in such settings. In addition, the present study focuses on South Korean legal

judgments; future extensions will include multilingual and cross-jurisdictional corpora to test the generality of the proposed approach.

6 Conclusions

In this study, we propose a TF-DF-based masking method as a novel data augmentation technique designed to address data imbalances in legal text classification. Unlike conventional augmentation methods, which suffer from semantic drift or struggle to preserve domain-specific legal terminology, the proposed approach selectively masks unimportant terms while preserving key legal expressions.

We evaluate the effectiveness of the proposed method for transformer-based models, including BERT and LegalBERT, in a large-scale legal classification task. The results demonstrate consistent improvements in accuracy and F1 score, particularly corresponding to underrepresented legal categories. LegalBERT exhibits the most substantial performance improvements, highlighting the strength of domain-adaptive augmentation.

Besides quantitative evaluation, the proposed method is compared with rule-based POS deletion. Although both methods exhibit similar metrics, our approach preserves essential legal semantics more reliably. For example, it retains crucial modifiers, e.g., "Asset" in "trust Asset". The proposed method also corrects previous misclassifications, especially in semantically complex domains, such as taxation and civil Law.

Although the proposed method relies on statistical weighting and does not fully guarantee the preservation of all legally essential terms, it reduces the likelihood of masking them significantly compared to traditional approaches. The current study focuses on South Korean legal texts. Future works should explore their generalizability to other legal systems and languages. In addition, we intend to investigate adaptive augmentation strategies using contextual embedding and attention-based token selection to further enhance performance in legal NLP tasks.

Acknowledgement: The authors would like to thank Chung-Ang University for providing computational resources and administrative support that contributed to this work.

Funding Statement: This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [RS-2021-II211341, Artificial Intelligence Graduate School Program (Chung-Ang University)], and by the Chung-Ang University Graduate Research Scholarship in 2024.

Author Contributions: The authors confirm contribution to the paper as follows: Conceptualization, Ye-Chan Park; methodology, Ye-Chan Park; software, Ye-Chan Park; validation, Ye-Chan Park, Mohd Asyraf Zulkifley, and Bong-Soo Sohn; investigation, Ye-Chan Park; resources, Ye-Chan Park; writing—original draft preparation, Ye-Chan Park; writing—review and editing, Mohd Asyraf Zulkifley, Bong-Soo Sohn, and Jaesung Lee; visualization, Ye-Chan Park; supervision, Jaesung Lee; project administration, Jaesung Lee. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available at https://huggingface.co/datasets/Yeeachan/korleg (accessed on 18 November 2025).

Ethics Approval: Not applicable. This study did not involve human participants or animals. All legal case documents used were publicly available and anonymized to remove personal identifiers.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. 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. Radnor, PA, USA: Association for Computational Linguistics; 2019. p. 4317–23.
2. 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. Radnor, PA, USA: Association for Computational Linguistics; 2020. p. 2898–904.
3. Chalkidis I, Jana A, Hartung D, Bommarito M, Androutsopoulos I, Katz D, et al. LexGLUE: a benchmark dataset for legal language understanding in English. In: Muresan S, Nakov P, Villavicencio A, editors. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics. Radnor, PA, USA: Association for Computational Linguistics; 2022. p. 4310–30.
4. Hakimi Parizi A, Liu Y, Nokku P, Gholamian S, Emerson D. A comparative study of prompting strategies for legal text classification. In: Proceedings of the Natural Legal Language Processing Workshop 2023. Singapore: Association for Computational Linguistics; 2023. p. 258–65.
5. Chen J, Tam D, Raffel C, Bansal M, Yang D. An empirical survey of data augmentation for limited data learning in NLP. Tran Assoc Comput Linguist. 2023;11:191–211. doi:10.1162/tacl_a_00542.
6. Hsu TW, Chen CC, Huang HH, Chen HH. Semantics-preserved data augmentation for aspect-based sentiment analysis. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Radnor, PA, USA: Association for Computational Linguistics; 2021. p. 4417–22.
7. Wilf A, Akter S, Mathur L, Liang P, Mathew S, Shou M, et al. Difference-masking: choosing what to mask in continued pretraining. In: Findings of the Association for Computational Linguistics: EMNLP 2023. Radnor, PA, USA: Association for Computational Linguistics; 2023. p. 13222–34.
8. Kesgin HT, Amasyali MF. Iterative mask filling: an effective text augmentation method using masked language modeling. In: Proceedings of International Conference on Advanced Engineering, Technology and Applications. Cham, Swizterland: Springer; 2023. p. 450–63.
9. Costa JAF, Dantas NCD, Silva EDSA. In: Evaluating text classification in the legal domain using BERT embeddings. Cham, Switzerland: Springer Nature; 2023. p. 51–63.
10. Nair I, Modani N. Exploiting language characteristics for legal domain-specific language model pretraining. In: Findings of the Association for Computational Linguistics: EACL 2023. Radnor, PA, USA: Association for Computational Linguistics; 2023. p. 2516–26.
11. Katz DM. Quantitative legal prediction-or-how i learned to stop worrying and start preparing for the data-driven future of the legal services industry. Emory LJ. 2012;62:909.
12. Bender EM, Koller A. Climbing towards NLU: on meaning, form, and understanding in the age of data. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Radnor, PA, USA: Association for Computational Linguistics; 2020. p. 5185–98.
13. Sennrich R, Haddow B, Birch A. Improving neural machine translation models with monolingual data. In: Erk K, Smith NA, editors. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Radnor, PA, USA: Association for Computational Linguistics; 2016. p. 86–96 doi:10.1162/tacl_a_00395.
14. Wei J, Zou K. EDA: easy data augmentation techniques for boosting performance on text classification tasks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. Radnor, PA, USA: Association for Computational Linguistics; 2019. p. 6382–8.
15. Xie Q, Dai Z, Hovy E, Luong T, Le Q. Unsupervised data augmentation for consistency training. Adv Neural Inform Process Syst. 2020;33:6256–68.
16. Ng N, Cho K, Ghassemi M. SSMBA: self-supervised manifold based data augmentation for improving out-of-domain robustness. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Radnor, PA, USA: Association for Computational Linguistics; 2020. p. 1268–83.

17. Karimi A, Rossi L, Prati A. AEDA: an easier data augmentation technique for text classification. In: Findings of the Association for Computational Linguistics: EMNLP 2021. Radnor, PA, USA: Association for Computational Linguistics; 2021. p. 2748–54.
18. Kim KM. A study of data augmentation for dense passage retrieval using corpus-passage frequency-based token deletion [master's thesis]. Seoul, Republic of Korea: Chung-Ang University; 2024.
19. Mamooler S, Lebret R, Massonnet S, Aberer K. An efficient active learning pipeline for legal text classification. In: Proceedings of the Natural Legal Language Processing Workshop 2022. Radnor, PA, USA: Association for Computational Linguistics; 2022. p. 345–58.
20. Ghosh S, Evuru CKR, Kumar S, Ramaneswaran S, Sakshi S, Tyagi U, et al. DALE: generative data augmentation for low-resource legal NLP. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore: Springer; 2023. p. 8511–65.
21. Kasthuriarachchy B, Chetty M, Shatte A, Walls D. Meaning-sensitive text data augmentation with intelligent masking. ACM Trans Intell Syst Technol. 2023;14(6):1–20. doi:10.1145/3623403.
22. Duffy W, O'Connell E, McCarroll N, Sloan K, Curran K, McNamee E, et al. Evaluating rule-based and generative data augmentation techniques for legal document classification. Knowl Inform Syst. 2025;67(9):7825–46. doi:10.1007/s10115-025-02454-x.
23. Sheik R, Siva Sundara K, Nirmala SJ. Neural data augmentation for legal overruling task: small deep learning models vs. large language models. Neural Process Lett. 2024;56(2):121. doi:10.1007/s11063-024-11574-4.
24. Kim MC, Penrod SD. Legal decision making among Korean and American legal professionals and lay people. Int J Law Crime Justice. 2010;38(4):175–97. doi:10.1016/j.ijlcj.2011.01.004.
25. Hwang W, Lee D, Cho K, Lee H, Seo M. A multi-task benchmark for Korean legal language understanding and judgement prediction. Adv Neural Inform Process Syst. 2022;35:32537–51.
26. Mizrahi D, Bachmann R, Kar O, Yeo T, Gao M, Dehghan A, et al. 4m: massively multimodal masked modeling. Adv Neural Inform Process Syst. 2023;36:58363–408.
27. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57. doi:10.1613/jair.953.