Master's Thesis Presentation for the 144th Dissertation Evaluation

A Study on Token Masking Augmentation Using Term-Document Frequency For Language Model-Based Legal Case Classification

단어·문서 빈도 기반 토큰 마스킹 증강을 활용한 언어 모델 법률 사례 분류 연구

Ye-Chan Park

qkrd1285@cau.ac.kr

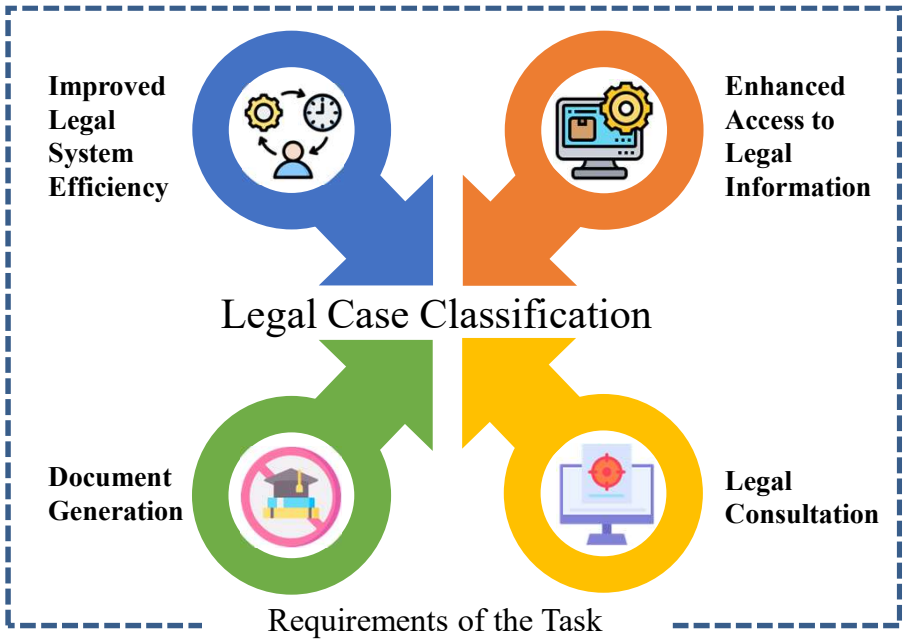
9th jun, 2025

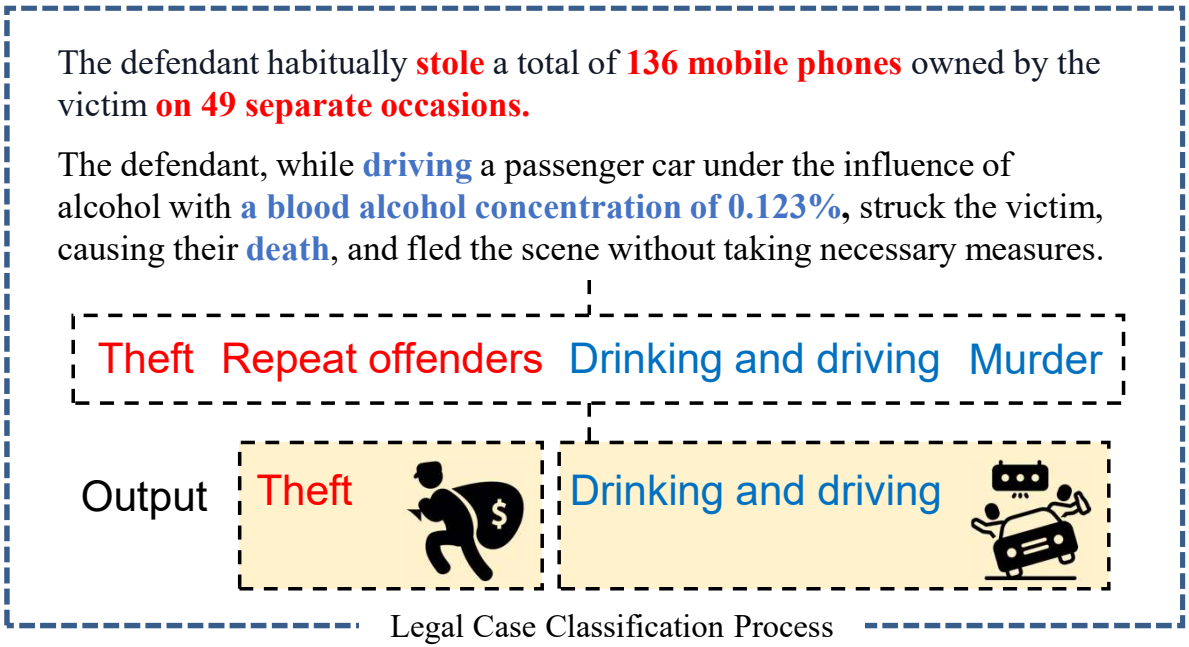
Contents

Introduction

Legal Case Classification

- Legal case classification supports enhanced legal system efficiency, improved legal consultations, better access to legal information, and individual empowerment. Accurate and swift case classification systems are essential for these applications.
- Legal case classification predicts the appropriate legal category for a given case document by analyzing its factual descriptions and legal arguments to support tasks such as retrieval, analytics, and decision-making in judicial systems.

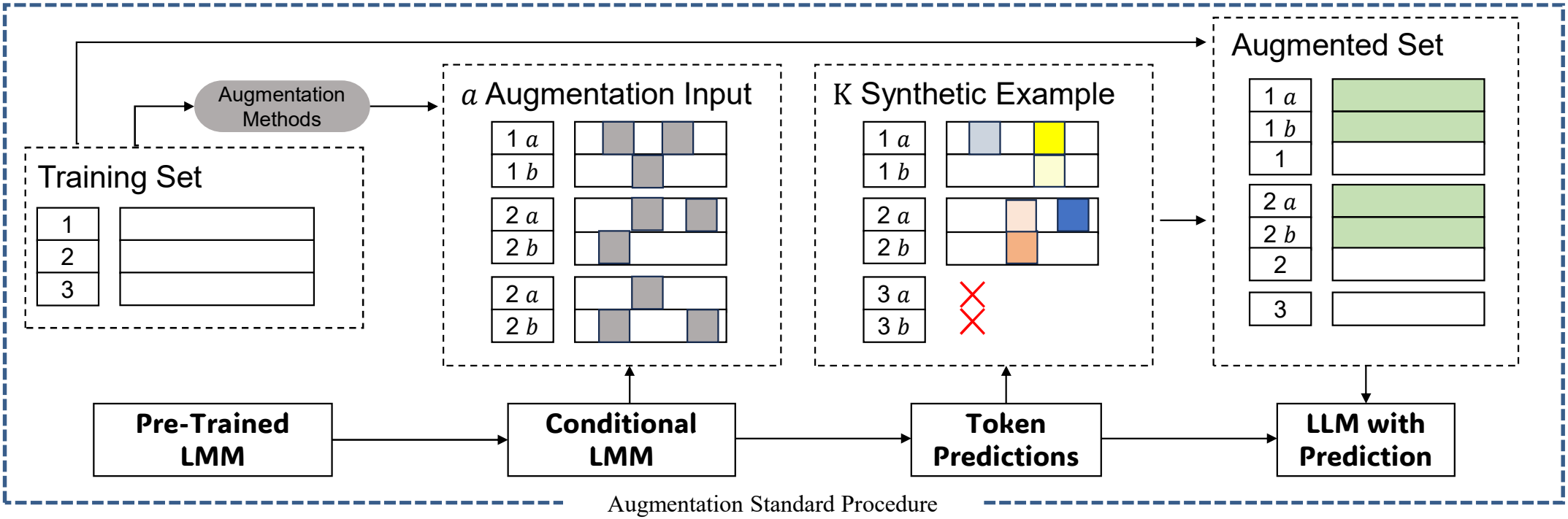




Introduction

Objective and Standard Procedure

- Due to the lack of comprehensive information in legal data and the need to address complex issues, generative models are often employed to create datasets and tackle the challenge of maintaining performance with the introduction of novel case categories.
- Conventional methods apply data augmentation by randomly masking tokens or replacing them using synonym dictionaries or contextual embeddings to diversify training inputs.

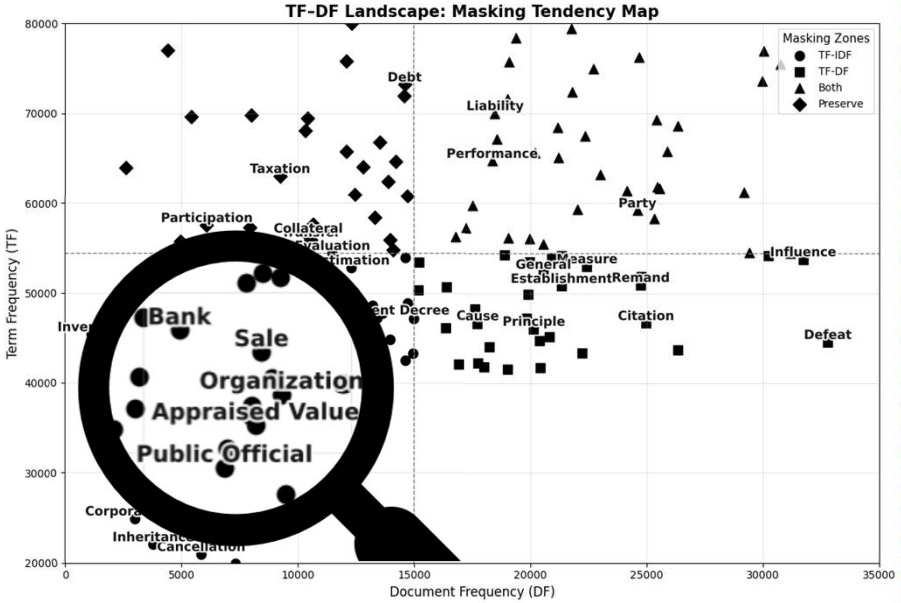


Augmentation Standard Procedure

Introduction

Problem and Strategy

- Conventional masking-based augmentation methods degrade classification performance because they frequently mask legally significant expressions that are essential for determining the correct legal category of a given case document.
- This thesis adopts a term frequency–document frequency (TF-DF) masking strategy that preserves low-frequency but legally important terms while selectively masking only the generic or less informative expressions in legal case documents.



TF-DF landscape

Tokens near X-axis (low TF, variable DF)

- **Cancellation**
  → Crucial for detecting administrative disposition cancellation cases.
- **Inheritance**
  → Indicates succession and property division in civil cases; often masked due to low TF despite high legal relevance.
- **Public Official**
  → Helps identify disputes involving duties or responsibilities of officials.

Tokens with high DF and moderate/high TF

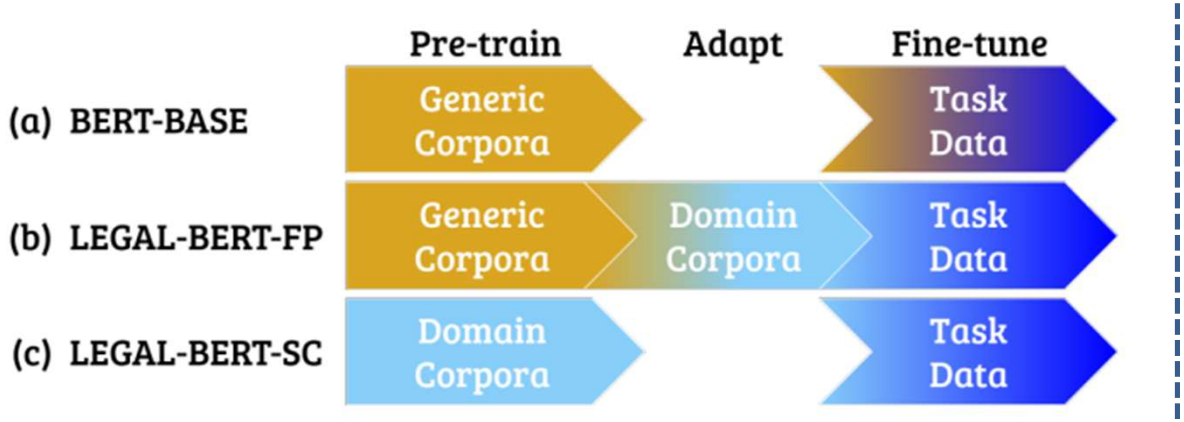
- **Cause**
  → Common connective phrase with low classification value.
- **Principle**
  → Abstract legal phrases (“principle of proportionality”); generally non-decisive.
- **Remand**
  → Procedural term in appellate opinions; masking prevents topic confusion.

Legally important terms

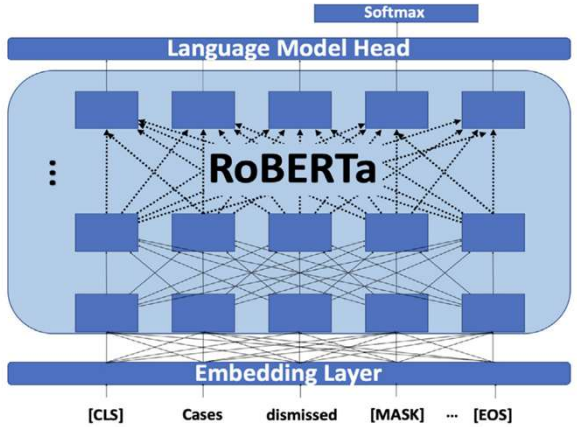
Related Work

Transformer-based Legal Text Classification

- Transformer-based models such as BERT, Legal-BERT, and CaseLaw-BERT improve legal text classification by leveraging large-scale legal corpora and domain-specific pretraining to enhance semantic understanding of legal documents.
- However, these models consistently underperform on minority legal classes due to data sparsity and lack effective mechanisms to address class imbalance through input-level augmentation.



Example of Domain-specific Model Feature



Example of Domain-specific Model Structure

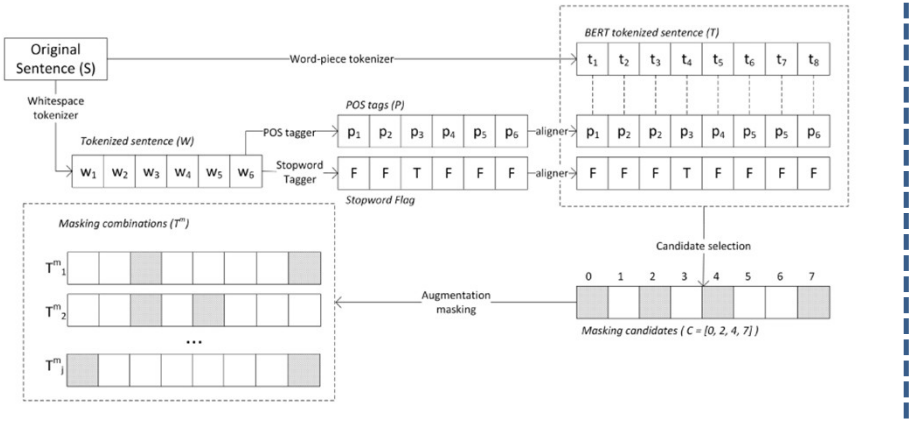
Song, D., Vold, A., Madan, K., & Schilder, F. (2022). Multi-label legal document classification: A deep learning-based approach with label-attention and domain-specific pre training. *Information Systems*, *106*, 101718.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-BERT: The Muppets straight out of law school. Findings of the Association for Computational Linguistics: EMNLP 2020, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Related Work

Rule-Based Data Augmentation

- General NLP studies employ rule-based and self-supervised augmentation strategies, such as back-translation, synonym replacement, POS deletion, and manifold-based sampling, to increase input diversity and improve generalization.
- However, these methods are not tailored to legal texts and often fail to preserve domain-specific legal expressions that are critical for accurate classification.



Example of POS-based Augmentation

Original: "The defendant intentionally breached the contractual obligation without prior notice."

After: "The defendant breached the obligation without notice."

Example of POS deletion Augmentation

Original: "The registration shall be revoked by the court."

After: "The [MASK] shall be revoked by the [MASK]."

Example of Synonym Replacement Augmentation

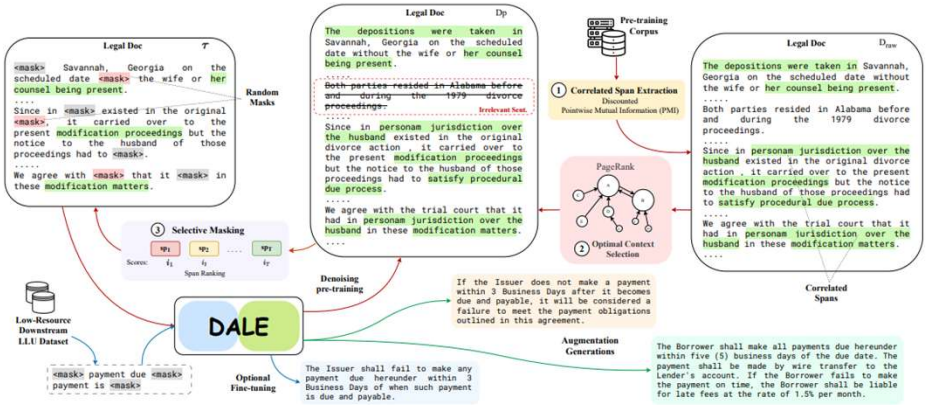
Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) Workshop*, 638–644. https://doi.org/10.18653/v1/D19-1670

Kasthuriarachchy, B., Chetty, M., Shatte, A., & Walls, D. (2023). Meaning-Sensitive Text Data Augmentation with Intelligent Masking. *ACM Transactions on Intelligent Systems and Technology*, *14*(6), 1-20.

Related Work

Masking-Based Augmentation Strategies

- Recent studies use masking-based augmentation techniques like TF-IDF masking, difference masking, and iterative mask filling to selectively remove or replace tokens and generate diverse training instances.
- However, these approaches rely on frequency heuristics that tend to discard low-frequency but legally significant terms, thus weakening the classifier's ability to capture legal semantics essential for accurate prediction.



Example of Masking Process

Original: "The agreement shall be canceled due to fraud."
After: "The agreement shall be [MASK] due to [MASK]."

Example of TF-IDF Augmentation

Original: "The building must be delivered by the defendant."
After: "The [MASK] must be delivered by the [MASK]."
("building" Remained)

Example of Meaning-Sensitive Masking Augmentation

Hsu, Y.-C., Lin, S.-H., & Glass, J. (2021). AEDA: An Easier Data Augmentation Technique for Text Classification. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 10853–10862. https://doi.org/10.18653/v1/2021.emnlp-main.846

Ghosh, S., Evuru, C. K. R., Kumar, S., Ramaneswaran, S., Sakshi, S., Tyagi, U., & Manocha, D. (2023). DALE: Generative Data Augmentation for Low-Resource Legal NLP. In The 2023 Conference on Empirical Methods in Natural Language Processing.

Related Work

Common Drawback of Conventional Methods

- Conventional augmentation methods commonly rely on surface-level statistical heuristics such as term frequency or TF-IDF, which fail to reflect the legal significance of terms and therefore tend to mask legally essential expressions during data generation.
- As a result of this masking behavior, classification performance deteriorates, particularly for underrepresented legal categories where critical terms are sparse but crucial for accurate prediction.

Augmentation Method	Reference (Author, Year)	Venue	Summary	Limitation from Augmentation Perspective
AEDA (TF-IDF-guided Masking)	Hsu et al., 2021	ACL/IJCNLP	Selectively masks low-TF-IDF tokens to improve performance over random masking.	Lacks analysis on semantic distortion or structural integrity post-masking.
DALE (Domain-Aware Legal Aug.)	Ghosh et al., 2023	ACL	Pre-defines masking range and position to retain meaning in legal texts.	Focuses on semantic preservation but lacks explicit integration of frequency-based term salience.
EDA (Rule-Based Text Augmentation)	Wei & Zou, 2019	EMNLP Workshop	Introduces simple rule-based methods like synonym replacement and insertion.	Ineffective for structured domains like legal/medical due to lack of semantic consistency.
MSM (Meaning-Sensitive Masking)	Kasthuriarachchy et al., 2023	ACM Transactions on Intelligent Systems and Technology	Adjusts masking positions based on contextual clues to preserve meaning.	Enhances consistency but lacks structural or legal-specific interpretability or analysis.

Table 1. Counterpart Summary

Proposed Method

Notation and Definition

- Let $D = \{x_i, y_i\}_{i=1}^{N}$ be a labeled dataset, where x_i denotes a legal document and $y_i \in$ y is its corresponding category label. The label typically reflects case type to which the document belongs, such as Civil, Criminal, Administrative, Family, Intellectual Property .
- Given x_i =[w_1, w_2,…, w_T], Method calculates p(w_t) from TF-DF scores and generates an augmented document $\tilde{x}_i$ by masking tokens with low TF-DF values while preserving legally critical terms.

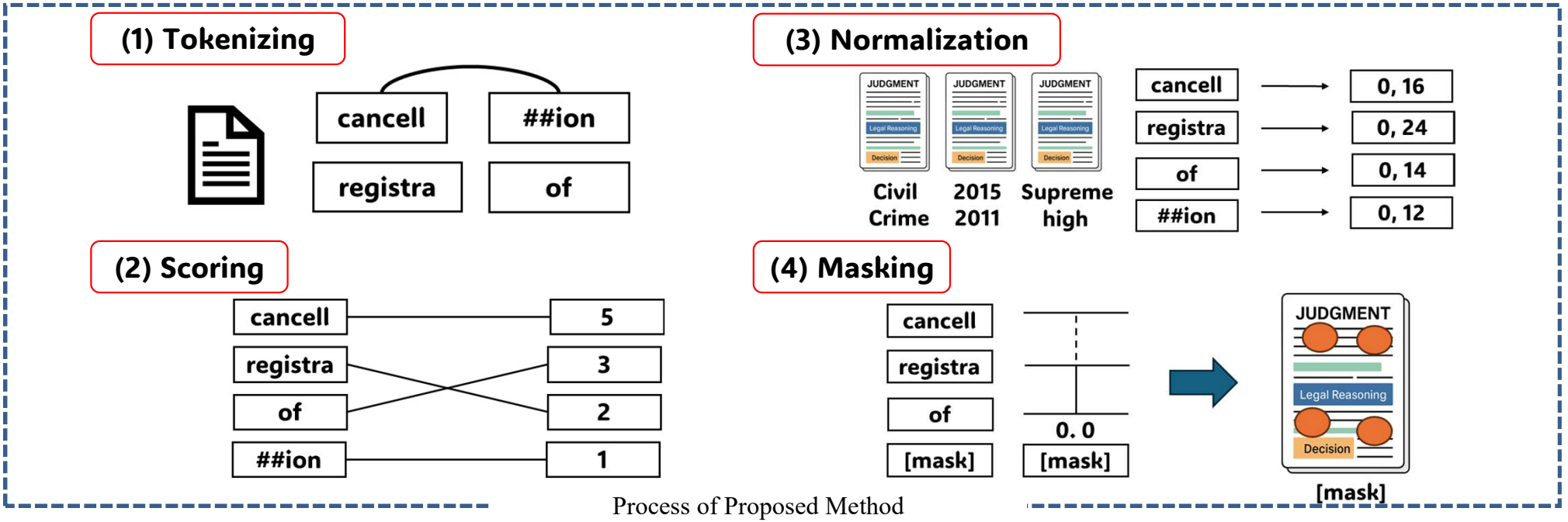
Notation	Name	Usage
D	Dataset	A set of labeled legal documents x_i, y_i ; input to the augmentation method.
x_i,	Input Document	A legal document consisting of a token sequence [w_1, w_2,…, w_T].
y_i	Label	The legal category assigned to document x_i; used for classification.
ω_t	Token	A token at position t in the input document; subject to importance scoring and masking.
$tf(\omega)$	Term Frequency	Number of times token ww appears in a document; used in importance score calculation.
$df(\omega)$	Document Frequency	Number of documents in which token ww appears; used to reflect corpus-level significance.
$w(\omega)$	Raw Score	TF-DF based importance score for token $w(\omega)$; defined as $tf(\omega) \cdot \log(1 + df(\omega))$
$\tilde{w}(\omega)$	Normalized Score	Min–max normalized importance score; determines masking probability.
α	Masking Scale	Hyperparameter controlling overall masking strength; used to scale masking probability.
$\tilde{x_i}$	Augmented Document	Output document after masking low-importance tokens in x_i; used for training.
$u(0,1)$	Uniform Distribution	Used to sample random values for stochastic masking decisions.

Table 2. List of Notations

Proposed Method

Procedural Overview

- Proposed method computes TF-DF-based importance scores for each token in a legal document and applies min–max normalization to generate token-specific masking probabilities that reflect legal relevance across the corpus.
- Based on these normalized scores, the method stochastically masks low-importance tokens to create augmented training samples that preserve legally critical expressions and improve model generalization.



Process of Proposed Method

Proposed Method

Details

Require: Dataset $\mathcal{D}$, tokenizer $\mathcal{T}$, masking scale α
Ensure: Augmented dataset $\mathcal{D}'$

$\mathcal{D}' \leftarrow \emptyset$ {▷ initialize empty container}
Tokenize every document in $\mathcal{D}$ using $\mathcal{T}$
Compute the document frequency $\mathrm{df}(\omega)$ for each token ω
for all document $d \in \mathcal{D}$ **do**
  Compute the term frequency $\mathrm{tf}(\omega)$ for all tokens in d
  $w(\omega) = \mathrm{tf}(\omega) \cdot \log(1 + \mathrm{df}(\omega))$ {▷ raw importance score}
  $\tilde{w}(\omega) = \dfrac{w(\omega) - \min(w)}{\max(w) - \min(w) + \epsilon}$ {▷ min–max normalization}
  $d' \leftarrow d$ {▷ create a working copy}
  for all token ω in d' **do**
    Draw $r \sim \mathcal{U}(0, 1)$
    if $r \leq \alpha \cdot \tilde{w}(\omega)$ **then**
      Replace ω with `[mask]`
    end if
  end for
  $\mathcal{D}' \leftarrow \mathcal{D}' \cup \{d'\}$ {▷ append masked document}
end for
return $\mathcal{D}'$

Proposed Masking Method

Experimental Settings

Dataset

- The data is prepared through a comprehensive multi-stage pipeline that includes legal text extraction, cleaning and normalization during preprocessing, morpheme restoration, and final document classification, as illustrated in Appendix 1
- The dataset originates from South Korean court rulings, obtained via Law Open Data provided by the Korean Ministry of Government Legislation.

Category	Number of Cases	Q1 (25%)	Median	Q3 (75%)	Max	Mean	Std
Civil	39,830	316	570	1,000	29,976	845.30	± 989.64
Criminal	20,454	228	443	858	111,734	931.65	± 2,501.44
Admin	12,660	318	569	1,018	42,415	860.93	± 1,071.01
Taxation	9,656	264	445	772	26,339	668.09	± 816.49
Intellectual Property	3,287	251	384	668.5	16,567	643.36	± 873.08
Family	1,273	214	392.5	772	7,835	620.45	± 674.98
Total	87,160	281	513	934	111,734	838.18	1505.41

Table 3. Data Characteristics

Cross Validation

- Number of Iterations: 10
- Data Split Ratio (Train: Validation: Test) = 6 : 2 : 2

Experimental Settings

Hyperparameter Settings

- This thesis set settings following prior works that have demonstrated its effectiveness in balancing modality contributions and improving generalization in stochastic masking.

- Batch size: 16
- Optimizer : AdamW
- Epochs : 20
- Learning Rate: 5×10^{-5}
- Loss function: CrossEntropyLoss

Evaluation Measures

- *Accuracy*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- *Precision*

$$Precision = \frac{TP}{TP + FP}$$

- *Recall*

$$Precision = \frac{TP}{TP + FN}$$

- *F1 Score*

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

TP: True Positives / TN: True Negatives / FP: False Positives/ FN: False Negatives

Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., & Zamir, A. (2023). 4m: Massively multimodal masked modeling. *Advances in Neural Information Processing Systems*, *36*, 58363-58408.

Experimental Results

Performance Comparison

- The proposed algorithm consistently improved both accuracy and F1 scores across all models compared to existing methods, showing particularly significant gains for models like LegalBERT, which initially exhibited high performance variance.
- This highlights the effectiveness of the selective masking strategy, which preserves legally important expressions while filtering out less relevant tokens.

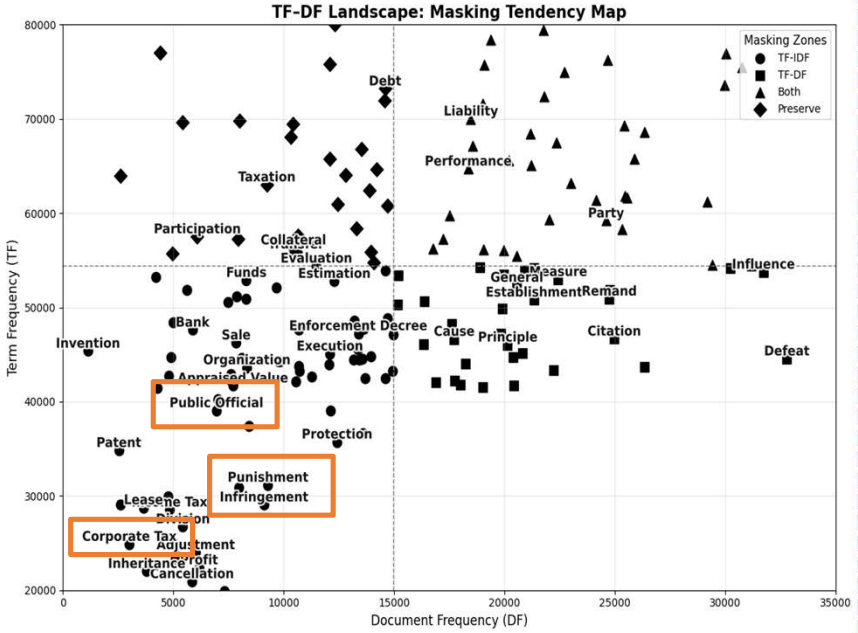
Model	Original	TF-IDF	POS	Proposed	Δ
LBox	0.9563 ± 0.0035	0.9805 ± 0.0008	0.9776 ± 0.0075	**0.9861 ± 0.0005**	+0.0298
KoGPT	0.9541 ± 0.0007	0.9811 ± 0.0035	0.9827 ± 0.0011	**0.9860 ± 0.0009**	+0.0319
LegalBERT	0.5952 ± 0.0327	0.7916 ± 0.0155	0.6913 ± 0.0198	**0.8453 ± 0.0100**	+0.2601
KLUE/BERT	0.9566 ± 0.0033	0.9840 ± 0.0007	0.9837 ± 0.0019	**0.9874 ± 0.0008**	+0.0667
BERT	0.8973 ± 0.0023	0.9475 ± 0.0017	0.9530 ± 0.0034	**0.9640 ± 0.0031**	+0.0308

Table 4. Performance Comparison (F1 Score)

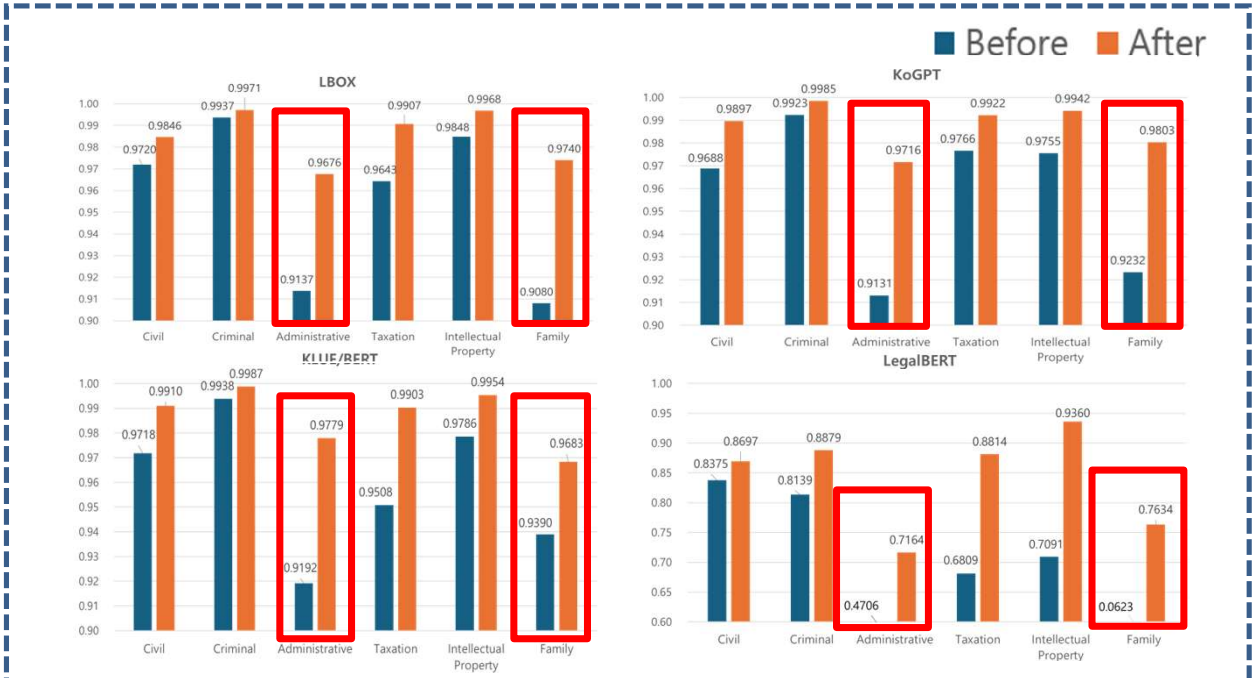
Experimental Results

In-depth Analysis

- The TF-DF landscape illustrates how token importance varies across documents, highlighting which terms are preserved or masked based on their frequency-informed significance.
- As a result, the proposed method improves classification performance across legal categories, particularly enhancing accuracy for underrepresented types like Administrative and Family Law while maintaining stability in well-performing classes.



TF-DF landscape



Per-Category Accuracy Comparison Before and After Data Augmentation

Experimental Results

In-depth Analysis

- The proposed masking method outperforms POS deletion and TF-IDF strategies by preserving legally meaningful expressions and selectively masking low-impact tokens, resulting in more semantically faithful augmented texts.
- Although the numerical performance is comparable, proposed method achieves greater legal reliability and robustness across models by mitigating the distortion of case meanings and ensuring consistent gains in classification metrics.

No	Version	Summary (with/without [MASK])	Classification Result
1	Original	The defendant seized property based on tax claims against MirDNI, but the property was entrusted to the plaintiff, making the seizure invalid.	Taxation
	Proposed (TF-DF)	Key expressions like “[MASK] - seizure,” “[MASK] - trust property,” and “[MASK] - invalid” are retained.	Taxation
	TF-IDF Masked	Excessive masking obscures legal reasoning by removing pivotal terms such as “[MASK] - trust” and “[MASK] - invalid.”	Civil
2	Original	The plaintiff was unaware of the appellate process due to public service and missed the appeal deadline. The court ruled this excusable and overturned the ruling.	Family
	Proposed (TF-DF)	Procedural terms such as “[MASK] - public notice,” “[MASK] - appeal deadline,” and “[MASK] - non-extendable period” are preserved.	Family
	TF-IDF Masked	Core procedural justice terms are masked, such as “[MASK] - notice” and “[MASK] - appeal deadline,” misleading the model.	Administrative
3	Original	The tax office imposed gift taxes on family members for alleged property transfers. Some charges were revoked due to overestimation and minor status of recipients.	Taxation
	Proposed (TF-DF)	Retains fiscal phrases like “[MASK] - gift tax,” “[MASK] - assessment,” and references to minors.	Taxation
	TF-IDF Masked	Financial and legal references like “[MASK] - valuation,” “[MASK] - donation,” and “[MASK] - tax amount” are removed.	Civil Law

Table 5. Mask Results for Augmentation Methods

Conclusion and Contribution

- Thesisudy presents a Proposed method as a novel data augmentation strategy to improve performance in legal classification. It enhances data diversity while preserving legally meaningful terms through selective masking based on term-document statistics.
- The proposed method addresses the limitations of conventional augmentation by minimizing semantic loss and better handling class imbalance. By weighting and filtering unimportant tokens, it preserves key expressions critical to accurate classification.
- Experimental results show that proposed method outperforms TF-IDF and POS deletion across transformer-based models. Notably, LegalBERT achieved the largest gain, and statistical tests confirmed the significance of improvements in both accuracy and F1-score.

Achievement

- **In International Conferences**
  - Justice AI: Implementation and Evaluation of Legal Case Retrieval AI Using Dense Passage Retrieval, ICCE, 2025
- **Patents**
  - Aspect-level Classification Program Based on Capturing the Structural Characteristics of the Target, *C-2023-024388*, 2023
- **In Journal**
  - Test Reliability in Case Law by Analysis of Korean Precedent Judgment Criteria on the Degree of Assault and Intimidation, Jahr-European Journal of Bioethics, 2024 (Scopus)
  - Deep Learning for Predicting Korean Court Judgments Based on Judges' Cognitive Reasoning, Journal of Artificial Intelligence Humanities, 2023 (KCI)
  - Analysis of Supreme Court of Korea's Judgment Criteria on the Degree of Assault and Intimidation - For Empirical Verification of the Court's Judgment Process, Journal of Artificial Intelligence Humanities, 2022 (KCI)
  - Effective Token Masking Augmentation Using Term-Document Frequency For Language Model-Based Legal Classification, Knowledge-based System, 2025 (SCIE)(submitted)

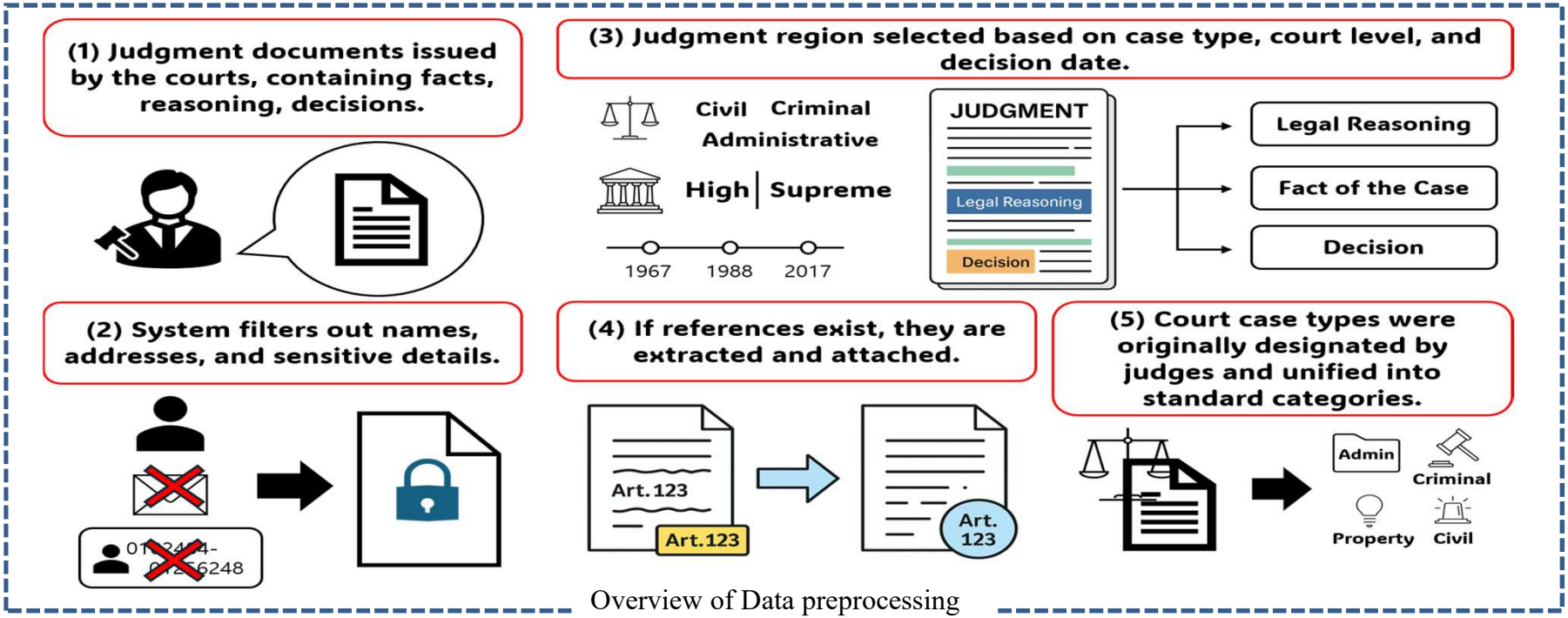
Thank you

Appendix 1

Data preprocessing

- Proposed preprocessing pipeline transforms raw court judgments into structured classification data by performing personal information masking, extracting legal references, and selecting text segments that contain relevant legal reasoning.

- The processed cases are then standardized through global case type unification, enabling consistent label assignment across diverse formats and ensuring compatibility with legal text classification models.



Overview of Data preprocessing

Reference

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N., & Androutsopoulos, I. (2020). Legal-BERT: The Muppets straight out of law school. Findings of the Association for Computational Linguistics: EMNLP 2020, 2898–2904. https://doi.org/10.18653/v1/2020.findings-emnlp.261

Chalkidis, I., Jana, A., Hartung, D., Thorne, J., & Aletras, N. (2021). LexGLUE: A benchmark dataset for legal language understanding in English. https://doi.org/10.48550/arXiv.2110.00976

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463

Ghosh, S., Xu, J., Dey, M., Ali, M., & Aji, A. F. (2023). DALE: Domain-Aware Legal Text Augmentation via Selective Masking. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/2023.acl-long.245

Wei, J., & Zou, K. (2019). EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP) Workshop, 638–644. https://doi.org/10.18653/v1/D19-1670

Kasthuriarachchy, B., Chetty, M., Shatte, A., & Walls, D. (2023). Meaning-Sensitive Text Data Augmentation with Intelligent Masking. ACM Transactions on Intelligent Systems and Technology, 14(6), 1-20.

Hsu, Y.-C., Lin, S.-H., & Glass, J. (2021). AEDA: An Easier Data Augmentation Technique for Text Classification. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10853–10862. https://doi.org/10.18653/v1/2021.emnlp-main.846

Mizrahi, D., Bachmann, R., Kar, O., Yeo, T., Gao, M., Dehghan, A., & Zamir, A. (2023). 4m: Massively multimodal masked modeling. Advances in Neural Information Processing Systems, 36, 58363-58408.