COSE474-2024F: Final Project Proposal
“A model for predicting Korean court rulings using RoBERTa”

Moogeun Park

1. Introduction

1.1. Motivation

The increasing complexity and volume of legal disputes place a significant burden on judicial systems worldwide. In the context of the Korean legal system, courts face mounting caseloads, often requiring considerable time and resources to reach verdicts. Simultaneously, legal practitioners and stakeholders struggle to navigate a vast and growing body of case precedents and judicial decisions. This underscores the pressing need for innovative tools that can assist in understanding and anticipating legal outcomes, thereby supporting more efficient legal processes.

The application of artificial intelligence in the legal domain offers promising solutions to address these challenges. By leveraging advancements in natural language processing (NLP) and machine learning, AI systems can analyze large volumes of legal data to uncover patterns and predict outcomes with a degree of accuracy that complements human expertise. In this context, developing a deep learning model capable of predicting case outcomes based on detailed claims, facts, and supporting information could serve as a transformative tool for legal professionals. Such a system could provide insights into potential outcomes, inform legal strategies, and enhance access to legal foresight, particularly for individuals or organizations with limited resources.

1.2. Problem Definition

Predicting the outcomes of court cases is a multifaceted problem involving both structured and unstructured data. Legal cases typically consist of claims and counterclaims articulated by the plaintiff and defendant, a set of relevant facts, and any supporting evidence or context. The challenge lies in accurately interpreting and integrating this diverse data to produce a reliable prediction of whether the plaintiff or defendant will prevail.

The problem is further compounded by the linguistic and contextual nuances inherent in legal documents. Korean court documents, in particular, often contain highly formalized and context-dependent language, which requires sophisticated NLP techniques to process effectively. Additionally, ensuring that the predictive model is both interpretable and unbiased is essential, given the ethical implications of deploying AI in legal contexts. Therefore, the central theme of this study is as follows: Can a deep learning model be designed to predict or assist in predicting case outcomes within the Korean judicial system by utilizing textual data and its structure? Furthermore, if such a model is designed, what methods can ensure its high accuracy, interpretability, and fairness?

1.3. Concise Description of Contribution

This study addresses the aforementioned challenges by proposing a comprehensive deep learning framework tailored for the prediction of legal case outcomes within the Korean judicial system. The contributions of this work are as follows:

Dataset Development: The creation and curation of a dataset comprising Korean court case data, including plaintiffs’ and defendants’ claims, case-related facts, and other foundational information. This dataset represents a valuable resource for advancing legal AI research in Korea.

Model Design: The development of a novel deep learning architecture that integrates NLP for processing unstructured textual data with structured data analysis techniques. This architecture is designed to capture the intricate relationships between claims, evidence, and case outcomes effectively.

Empirical Evaluation: A thorough evaluation of the model’s performance on the curated dataset, including comparisons with baseline methods and analysis of its generalizability. The study also examines the interpretability and ethical considerations of the model’s predictions, highlighting its potential for real-world applications.

By addressing the unique characteristics of Korean legal data and proposing a robust AI framework, this work contributes to the growing body of research on judicial AI. It aims to pave the way for more efficient and accessible legal decision-making tools, ultimately enhancing the fairness and transparency of the judicial process.

2. Methods

2.1. Significance/Novelty/Main challenges

This study introduces an approach to legal outcome prediction by utilizing a RobertaBinaryClassifier as the primary framework. The RoBERTa model, a transformer-based architecture known for its exceptional performance in natural language processing tasks, is fine-tuned on a curated dataset of Korean court cases.

The significance of this approach lies in its ability to adapt to the linguistic and contextual intricacies of Korean legal documents, which are characterized by formal and domain-specific language. Previous studies often relied on simpler or rule-based models, but this work leverages the deep contextual embeddings of RoBERTa to analyze claims, counterclaims, and case-related facts comprehensively. By doing so, this method not only enhances prediction accuracy but also advances interpretability and scalability in the field of judicial AI.

Legal case data typically includes unstructured textual elements (e.g., claims) and structured metadata (e.g., case type, filing date). Integrating these different types of information into a single predictive model is challenging. To address this issue, we handle unstructured text elements and structured metadata separately, applying different approaches to each type of data.

2.2. Main figure/Reproducibility

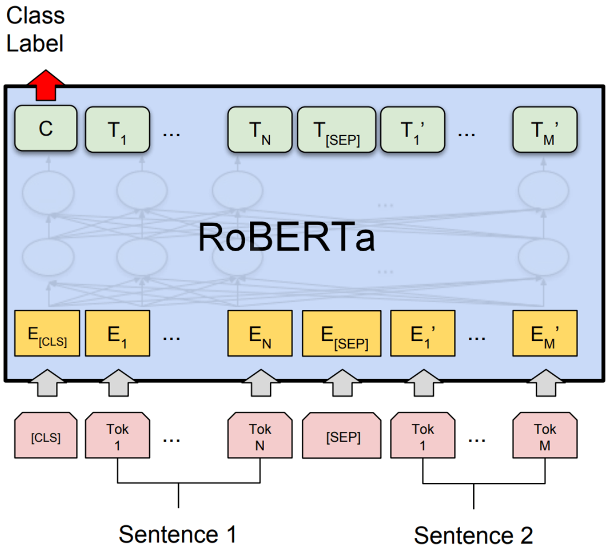

Figure 1. RoBERTa

2.3. Reproducibility

The RoBERTa model tokenizes the input text, converts it into embeddings, and uses a Transformer encoder to understand the meaning of the text. Finally, it generates predictions through a fine-tuned approach. During this process, the model learns the relationships between tokens and produces accurate outputs based on the given context.

The Binary Cross-Entropy with Logits Loss (BCEWithLogitsLoss) is a loss function commonly used for binary classification tasks. It combines the sigmoid activation function and binary cross-entropy loss into a single efficient step, making it numerically stable and computationally efficient. By applying the sigmoid function internally, BCEWithLogitsLoss transforms raw logits into probabilities and calculates the loss based on the divergence between predicted probabilities and ground truth labels. This approach is particularly suitable for models outputting logits directly, ensuring stability in gradient calculations.

The AdamW (Adaptive Moment Estimation with Weight Decay) optimizer is an advanced variant of the Adam optimizer. It incorporates a decoupled weight decay regularization technique, which independently applies weight decay to the parameters during optimization. This adjustment prevents the weight decay term from interfering with the adaptive learning rate updates of Adam, thereby improving generalization and convergence. AdamW is particularly effective in training deep learning models, offering a balance between speed and stability by adapting learning rates for each parameter based on first- and second-order moments.

3. Experiments

3.1. Dataset

Court rulings are crawled from the official website of the Republic of Korea courts, categorized by field and format. The data is then divided into foundational case information, parties involved, plaintiff/defendant claims, and case-related facts. The labels are defined as 1 for defendant victories and 0 for plaintiff victories.

3.2. Computing resource

The training was conducted on Google Colab using a single T4 GPU. The necessary libraries for PyTorch were installed, and the code was implemented accordingly.

3.3. Experimental design/setup

The data crawled from the court's website was split into 800 training samples and 200 test samples. The training set was further divided into training and validation data in an 8:2 ratio. The batch size was set to 32, and the dropout parameter was specified as 0.3. The loss function used was BCEWithLogitsLoss, the optimizer was AdamW, and the learning rate was set to 1e-5. By fine-tuning these hyperparameters and iterating through multiple training processes, the model's performance was improved.

3.4. Quantitative results

After multiple training and testing iterations, the results showed some variability due to the relatively small dataset size. However, the average accuracy exceeded 0.5, with a maximum of approximately 0.6. In addition to the RoBERTa model, other models were tested, and various loss functions and optimizers were explored. Among them, the experimental design described above proved to be the most effective. By fine-tuning the hyperparameters, the accuracy gradually improved over time.

3.5. Figures/Analysis

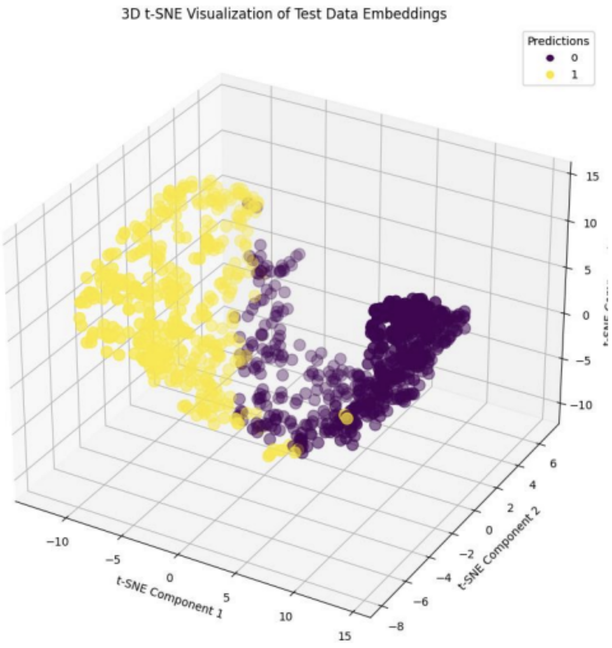

Figure 2. 3D t-SNE

The above figure illustrates the embedding vectors extracted from the test dataset, classified according to their labels. As shown in the figure, apart from the contact area between label 0 and label 1, the data is relatively well-separated. High-accuracy results from other models and training processes are also well-represented by t-SNE. This demonstrates that the claims of the plaintiff and defendant, related facts, and the verdict can be tokenized into embedding vectors without significant loss of information, enabling effective prediction.

3.6. Discussion

While the results appear fairly reasonable, it cannot be considered entirely successful. For AI to be implemented in actual judicial decision-making, it must provide correct predictions or defer judgment in almost all cases. However, the current accuracy is insufficient, and the results of the embedding vectors remain challenging to interpret. Therefore, instead of directly using AI models for verdicts, they should be employed as auxiliary tools to support judicial decision-making.

4. Future Direction

This study has contributed to the growing field of AI in the legal domain by proposing a framework that uses deep learning techniques to predict legal case outcomes in the context of the Korean judicial system. The creation of a dataset comprising detailed claims, case facts, and verdicts has provided a valuable resource for future research. Additionally, the model design, which integrates NLP techniques with structured data analysis, offers a novel approach to handling legal case data. Finally, the empirical evaluation demonstrated the model's potential, although room for improvement remains.

Despite the promising results, this study faces several limitations. The relatively small size of the dataset and the inherent complexity of legal language affected the model's generalizability and performance. Furthermore, the lack of interpretability in the model's decision-making process poses a challenge for practical deployment in real-world legal settings. Addressing these issues will require the incorporation of larger datasets, improved model transparency, and the development of methods for mitigating bias.

The future direction of AI in legal judgment models offers exciting opportunities for improving efficiency and fairness in the legal system. However, to realize the full potential of AI in this domain, ongoing efforts will be required to enhance model performance, ensure fairness, and integrate these technologies in a manner that complements human expertise. By addressing the challenges identified above, AI models can play a transformative role in the legal industry, ultimately fostering more equitable and accessible legal decision-making processes.

References

Park Ye Chan, L. J. Deep learning for predicting korean court judgments based on judges' cognitive reasoning. *Journal of AI Humanities*, 13:73–107, 2023.

Sung Won Kim, G. R. P. Deep learning based semantic similarity for korean legal field. *KTSDE*, 11(2):93–100, 2022.