MLOps-Driven Natural Language Search for Legal Case Retrieval: A Cost-Effective and Scalable Approach [*]

Hae Won Park

Abstract

In the rapidly evolving landscape of machine learning applications, industries are in dire need of efficient, cost-effective solutions tailored to their unique challenges. This paper delves into the development and implementation of an MLOps-driven natural language search service, specifically designed for the retrieval of legal cases. Unlike traditional case search services, this innovative approach harnesses the power of AI to continuously update its database, ensuring the provision of not only the most relevant case references but also associated violation articles. By integrating MLOps principles, this system promises reduced maintenance costs, timely updates, and unparalleled scalability. Furthermore, the fusion of advanced NLP techniques enables users to query in natural language, democratizing access to intricate legal information and bridging the gap between legal professionals and the general public.

***K*eywords** MLOps · Natural Language Processing · Legal Case Retrieval · RoBERTa

1 Introduction

The modern legal domain, characterized by a dynamic proliferation of cases, legal principles, and doctrines, presents a daunting challenge for both professionals and the general populace. This flux underscores the need for a reliable and swift mechanism that not only remains updated with the latest legal changes but also discerns the intricate nuances embedded within legal verbiage.

The challenge, however, isn't merely about volume; it's about complexity. Legal documents are dense, brimming with terminologies and structures that often require years of training to decipher. For a layperson, navigating this labyrinth can be disorienting. For professionals, while their training aids in comprehension, the sheer volume and pace of new case laws can make efficient retrieval a time-consuming endeavor.

Enter the realm of MLOps-driven Natural Language Search—a paradigm shift that promises to transform how we interact with legal datasets. By merging the analytical prowess of Machine Learning (ML) with the operational efficiency of DevOps, an MLOps-driven approach ensures that not only are search results accurate and contextually relevant, but the underlying systems remain agile, scalable, and updated.

Furthermore, with the integration of Natural Language Processing (NLP), this approach holds the potential to democratize access to legal knowledge. Rather than grappling with legal jargon, users can frame queries in everyday language and still retrieve pertinent legal cases or articles. This shift signifies more than just convenience—it heralds a more inclusive legal ecosystem where legal knowledge isn't gated behind professional expertise or technical know-how.

This paper delves into an innovative project that embodies this vision. By leveraging the synergies of ML and Operations (Ops) in the legal landscape, we aim to elucidate a roadmap for a future where legal information retrieval is not a hurdle but a seamless, intuitive experience.

[*] *Citation*: To be updated post publication.

2 Background and Prior works

The legal domain has always been a repository of vast amounts of textual data, with its roots tracing back to ancient civilizations where legal codes were inscribed on stone tablets. As societies evolved, so did the complexity and volume of legal documents. The digital age brought about a revolution in how legal data is stored, accessed, and analyzed. However, the tools to search and retrieve relevant legal information have not kept pace with the exponential growth of data.

2.1 Traditional Legal Search Tools

Traditional legal search tools primarily rely on keyword-based search algorithms. These tools require users to input specific keywords or phrases to retrieve relevant documents. While this method can be effective for users familiar with legal terminologies, it poses challenges for laypersons. Moreover, keyword-based searches often return a plethora of results, many of which may not be directly relevant to the user's query.

2.2 Emergence of AI in Legal Research

With advancements in artificial intelligence (AI) and machine learning (ML), there have been attempts to incorporate these technologies into legal search tools. AI-powered tools can understand the context behind a query, making them more effective than traditional keyword-based searches. However, these tools are still in their nascent stages and have their own set of challenges, such as understanding the nuances of legal language.

2.3 Natural Language Processing (NLP) in Legal Domain

NLP, a subfield of AI, focuses on the interaction between computers and human language. In the legal domain, NLP can be used to analyze legal documents, extract relevant information, and even predict legal outcomes. Some preliminary tools have started using NLP to simplify legal language for laypersons. However, there's still a long way to go in terms of achieving high accuracy and reliability.

3 Method

3.1 MLOps

MLOps, which combines the principles of DevOps with machine learning, is revolutionizing the way we manage and deploy ML projects. By ensuring continuous integration, delivery, and deployment of ML models, MLOps streamlines the end-to-end machine learning lifecycle. This approach not only optimizes costs but also reduces the time to market, ensuring that models are scalable, maintainable, and always up-to-date with the latest data.

3.1.1 System Architecture

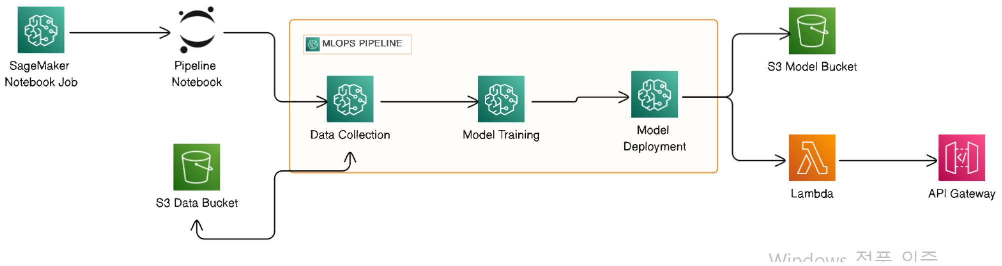

Figure 1: Overall System Architecture of the MLOps-driven Legal Case Retrieval System.

The architecture is modular and scalable, starting with data collection and ending with real-time querying. Each component is designed to work seamlessly with the others, ensuring a cohesive system that meets the demands of modern legal professionals.

3.1.2 Structure of MLOps Pipeline

The core of our system is built upon AWS SageMaker Studio's pipeline. This pipeline is triggered by scheduled notebook instances that run at predefined intervals. The pipeline is divided into several stages:

1. Data Collection: Fetching the latest legal cases from national databases.
2. Preprocessing: Cleaning and structuring the data for training.
3. Model Training: Using the preprocessed data to train our RoBERTa model.
4. Deployment: Hosting the trained model for real-time querying.
5. Evaluation: Continuously monitoring and evaluating the model's performance.

All model artifacts, including training checkpoints and model weights, are stored in S3 buckets. AWS Lambda serves the trained models as RESTful APIs, ensuring that they are accessible for real-time querying.

3.1.3 SageMaker and Scalability

AWS SageMaker pipelines are inherently scalable. They are built on top of other AWS services that allow for easy scaling, both vertically and horizontally. Integration with platforms like Kubeflow further ensures that the pipeline can handle increased loads, be it from more data or more users. The use of SageMaker also allows for seamless integration with other AWS services, ensuring a cohesive and unified system.

3.2 Dataset

National legal information APIs provide the primary source of our data. As of September 5, 2023, the API contains a vast collection of about 82,038 cases. After rigorous preprocessing and labeling, a subset of 21,280 cases was curated for model training and validation.

The preprocessing involved several steps:

- **Data Cleaning:** Removal of any irrelevant or redundant information, ensuring that the dataset is free from any noise or inconsistencies.
- **Tokenization:** Breaking down the legal text into smaller chunks or tokens, making it easier for the model to process and understand.
- **Labeling:** Manual annotation of the cases to categorize them based on their legal significance, relevance, and context.
- **Data Augmentation:** Techniques were employed to artificially increase the size of the training dataset, ensuring a diverse range of data for the model to learn from.

The chosen base model for this project is the KLUE RoBERTa-base, an advanced variant from the BERT series. RoBERTa models are known for their efficiency in understanding context within textual data, making them ideal for tasks like legal case retrieval. The model was fine-tuned on our curated dataset, ensuring that it is tailored to understand and retrieve relevant legal cases based on user queries. The choice of RoBERTa was also influenced by its proven track record in various NLP tasks and its ability to capture intricate nuances in textual data, especially in the complex domain of legal texts.

3.3 Fine Tuning Model

The KLUE RoBERTa-base model, which serves as our primary model, was fine-tuned specifically for the task of legal case retrieval. Fine-tuning is a process where a pretrained model, which has been trained on a large general dataset, is further trained on a smaller, task-specific dataset. This allows the model to adapt its knowledge to the specific nuances and requirements of the new task.

For our system, the RoBERTa-base model was fine-tuned using the curated dataset of 21,280 legal cases. The fine-tuning process involved adjusting the model's weights using our dataset while keeping the base architecture unchanged. This approach ensures that the model retains the vast knowledge it gained during its initial training while adapting to the specificities of legal case retrieval.

3.3.1 Model Architecture and Details

The search tool is powered by the KLUE RoBERTa-base model, an advanced variant of the BERT series. This model is known for its efficiency in understanding context within textual data, making it ideal for legal case retrieval tasks. While the RoBERTa-base model serves as our primary model, we are currently exploring lightweight models to optimize the system further. These models aim to maintain high performance while being computationally less intensive.

Model Component	Description
Architecture	RoBERTa
Training Data	KLUE
Language	Korean
Tasks	Text Classification
Layers	12
Hidden Size	768
Attention Heads	12
Total Parameters	125M

Table 1: Details of the KLUE RoBERTa-base model

The below table provides a detailed overview of the fine-tuning process and the resources utilized for the KLUE RoBERTa-base model. The chosen model for the task is the KLUE RoBERTa-base, a renowned variant of the BERT series, recognized for its adeptness in grasping contextual nuances within textual data. This model underwent fine-tuning specifically tailored for the legal case retrieval task, utilizing a curated dataset comprising 21,280 legal cases. The entire fine-tuning procedure was both time-efficient and cost-effective, culminating in approximately 2.7 hours and incurring a cost of around $101. This is a stark contrast to the extensive time and financial resources required to train large models from the ground up.

To guarantee optimal performance and efficient training, the fine-tuning was executed on AWS's `ml.p3.8xlarge` instance. This particular instance is fortified with four NVIDIA® V100 Tensor Core GPUs, specifically engineered for high-performance computing and intricate AI workloads. The integration of such formidable GPUs ensures rapid model training capable of managing the intricacies of the expansive legal dataset. The synergy between the RoBERTa model and these high-caliber resources ensures precise and swift legal case retrievals, solidifying the system's robustness and reliability.

Aspect	Details
Model	KLUE RoBERTa-base
Fine-tuning Data	21,280 legal cases
Cost	$101
Time	2.7 hours
AWS Resource	ml.p3.8xlarge
GPU	NVIDIA® V100 Tensor Core (4)

Table 2: Details of the fine-tuning process and resources used

While the RoBERTa-base model offers impressive performance, its size and computational requirements can be a concern, especially for real-time applications. We are currently exploring the possibility of using smaller, more efficient models. We are in the process of conducting experiments to determine the feasibility and performance of these lightweight models in the context of legal case retrieval. Experiments are underway to evaluate the performance of these lightweight models in the context of legal case retrieval, and preliminary results are promising.

Model	Advantages	Size/Performance
DistilBERT	60% faster, retains 95% of BERT's performance	40% smaller
TinyBERT	9.4x faster than BERT-base	7.5x smaller
MobileBERT	Optimized for mobile devices	4.3x smaller than BERT-base

Table 3: Comparison of potential lightweight models for future experiments

4 Conclusion

The integration of MLOps and advanced NLP techniques in the realm of legal case retrieval signifies a monumental shift in how legal information is accessed and understood. By allowing users to employ natural language for their queries, the barrier of legal jargon is effectively dismantled, making intricate legal knowledge accessible to all. The success of this project not only underscores the transformative potential of MLOps in real-world applications but also sets a precedent for its broader adoption across various sectors. As industries continue to grapple with the challenges posed by vast data volumes and the need for real-time insights, solutions like the one presented in this paper emerge as beacons of innovation, pointing the way forward. The future beckons a world where technology and human expertise converge, making information not just available but also comprehensible to all, irrespective of their background or expertise.

References