Legal Text Similarity Analysis Using Deep Learning Based Model with Enhanced Feature Extraction

I.CHANDRAKALA[1], A. N. DINESH KUMAR[2]
[1]*PG Student, Vemu Institute of Technology, P. Kothakota*
[2]*Assistant professor, Vemu Institute of Technology, P. Kothakota*

Abstract: **The application of artificial intelligence in the legal domain has received significant attention from legal professionals and AI researchers in recent years. The intelligent judge system has made remarkable progress due to advancements in natural language processing, particularly deep learning. Matching similar cases has enormous potential with significant implications for the legal domain. Matching and analyzing similar cases helps legal professionals make more reasonable judgments, ensuring fairness, consistency, and accuracy in law applications. The existing methods did not fully use representation-based and interaction-based text matching in the feature extraction. This paper presents an innovative approach that employs ensemble learning with multiple models to enhance the prediction of legal case similarity. The method comprises two sub-networks: a similarity representation sub-network and a binary classification judgment sub-network. The similarity representation sub-network is trained using contrastive learning, focusing on semanticizing the similarity between sample features to distinguish between dissimilar samples and reduce the distance between similar ones. Furthermore, the binary classification judgment sub-network integrates sample pairs to facilitate feature interaction between text pairs during extraction. The training of these two sub-networks employs different information processing and optimization loss, which allows ensemble learning to capitalize on the strengths of both models and significantly improve the accuracy of predicting the similarity of legal cases. The accuracy of our method on the test set is 74.53%, outperforming other existing methods on the public dataset CAIL2019-SCM.**

INTRODUCTION

Similar case matching (SCM) leverages artificial intelligence (AI) to compare legal cases based on their principles and factual details, significantly enhancing the efficiency and consistency of legal decision-making. By identifying relevant precedents, SCM ensures that judgments are aligned across similar cases, reducing human bias and subjectivity. This technology streamlines legal processes by minimizing the need for extensive manual analysis, thus saving time and resources for both lawyers and judges. AI-driven case matching enables a more objective approach to legal decisions, grounded in established principles and evidence, and promotes fairness in the legal system. Overall, SCM not only improves the accuracy of legal outcomes but also optimizes the workflow, making legal proceedings more efficient and reliable.

LITERATURE SURVEY

1. Fang et al. (2022) - BERT-based Multi-task Learning for Legal Document Analysis

- This study explores a deep learning-based approach using Bidirectional Encoder Representations from Transformers (BERT) for legal text classification and case similarity matching.
- The model is trained on large legal judgment datasets, improving accuracy in multi-task learning scenarios.
- A key limitation is the high computational cost associated with training and deploying BERT-based models in real-world applications.

2. Zhong et al. (2020) - Low-resource Legal Case Matching

- The research investigates methods for legal text matching in low-resource environments, leveraging transfer learning and pre-trained embeddings.
- The model demonstrates effectiveness in handling jurisdictions with limited legal data.
- However, scalability across different legal systems and domains remains a challenge due to data constraints.

3. Francia et al. (2022) - Survey of Text Mining Techniques for Judicial Decision Prediction

- This study provides a comprehensive review of text mining approaches applied to judicial decision-making.

- It categorizes various machine learning and deep learning methods used in legal text processing.
- Despite its breadth, the study emphasizes the need for further validation of techniques in complex case scenarios.

4. Tran et al. (2020) - Encoded Summarization for Legal Case Retrieval

- The research focuses on generating case summaries using neural network-based encoding techniques.
- Encoded summarization facilitates faster and more accurate retrieval of relevant legal cases.
- The effectiveness of the method depends on the quality of case encoding, requiring high-quality datasets

5. Mandal et al. (2021) - Unsupervised Textual Similarity for Legal Case Reports

- The study introduces an unsupervised approach to measuring textual similarity between court case reports.
- The method is effective for identifying related cases without labeled training data.
- However, variations in legal language and terminology pose challenges in achieving high accuracy across different case types.

PROBLEM STATEMENT

Existing legal case similarity prediction methods fail to fully utilize representation and interaction-based text matching, leading to suboptimal accuracy and consistency. An improved approach is required to better assist legal professionals in making informed judgments

PROPOSED METHOD

Legal documents similarity can help intelligent judges can know about similarity between cases and based on similarity so they can improve verdict by removing errors in previous judgements. Growing AI algorithms popularity attracting legal professional to apply AI algorithms to predict and find matching legal documents. Matching and analysing similar cases helps legal professionals make more reasonable judgments, ensuring fairness, consistency, and accuracy in law applications. The existing methods did not fully use representation-based and interaction-based text matching in the feature extraction. This paper presents an innovative approach that employs ensemble learning with multiple models to enhance the prediction of legal case similarity. The method comprises two sub-networks: a similarity representation sub-network and a binary classification judgment sub-network. The similarity representation sub-network is trained using contrastive learning, focusing on semanticizing the similarity between sample features to distinguish between dissimilar samples and reduce the distance between similar ones. Furthermore, the binary classification judgment sub-network integrates sample pairs to facilitate feature interaction between text pairs during extraction.

METHODOLOGY

1. DATA AUGMENTATION

Data augmentation is a technique to increase the size and diversity of a data set in the legal domain for similar case matching. Its goal is to identify similar cases based on their underlying legal concepts and reasoning. Data augmentation can generate new case summaries or legal issues by applying different legal rules or regulations from existing cases. It can help the model understand how legal cases can be affected by different legal contexts and be more.

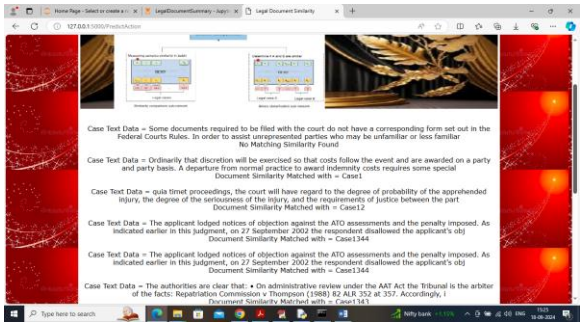
2.SUB-NETWORK BASED ON CONTRASTIVE LEARNING FOR SCM

In recent years, contrastive learning has gained increasing attention in computer vision and natural language processing. In text contrastive learning, multiple samples are compared with each other to bring similar texts closer together and dissimilar texts further apart. The ultimate goal is to learn how to minimize the distance between similar samples and maximize the distance between dissimilar samples in a high-level semantic feature space. The BERT model relies on contrasting samples within a batch to achieve contrastive learning in the implementation process. For legal cases, the text is encoded by BERT to become sentence embeddings. Each case statement has its unique feature representation for subsequent similarity evaluation. By constructing samples through data augmentation, the BERT weights are trained to extract features closer in the distance for enhanced and similar samples while keeping the features of dissimilar samples farther away.

3.SUBNETWORK BASED ON SIMILARITY BINARY CLASSIFICATION FOR SCM

This subnetwork transforms the text similarity problem into a binary classification problem based on feature-interactive text similarity analysis. In the implementation process, we transform the similarity in the triplets (A, B, C) into a binary classification problem by comparing the similarity between the combination (A, B) and the combination (A, C). It converts a text similarity problem into a binary classification problem. While using the BERT network to extract text features, we take A and B as joint inputs, hoping to obtain their interactive feature representations through the BERT network. Similarly, we take A and C as inputs, and the BERT with shared parameters also expects to obtain their interactive feature representations.

RESULTS



CONCLUSION

This paper proposes an ensemble learning-based method to improve accuracy in assessing legal similarity. Based on feature representation, a deep model is constructed using contrastive learning, and the correlation between text features judges similarity. On the other hand, through the feature interaction matching method, two texts to be evaluated are input into the network together, and information fusion and interaction are carried out at the beginning of the network for the text pair. The two sub-models use different optimization functions and have strong complementarity. Then, the two sub-models are ensembled to improve prediction performance, which can provide ideas for further research.

Improvements have also been made in the model proposed for this task due to the need for more utilization of prior legal knowledge. Utilizing legal knowledge and simulating legal reasoning for the dataset is still challenging. Based on the semi-structured legal texts, legal elements can be extracted from case statements in future work. Judicial elements are one of the important criteria for determining the relevance between cases. Cases with the same elements have similar circumstances and verdicts. By extracting the key elements from case statements, the length of the input text can be reduced, and the model's performance can be improved.

REFERENCES

[1] F. Aman and W. Yanchuan, ‘‘Multi task intelligent legal judgment method based on BERT model,’’ Microelectron. Comput., vol. 39, no. 9, pp. 107–114, 2022.

[2] J. Fang, X. Li, and Y. Liu, ‘‘Low-resource similar case matching in legal domain,’’ in Artificial Neural Networks and Machine Learning—ICANN 2022, PT II (Lecture Notes in Computer Science), vol. 13530, E. Pimenidis, P. Angelov, C. Jayne, A. Papaleonidas, and M. Aydin, Eds. Cham, Switzerland: Springer, Sep. 2022, pp. 570–582, doi: 10.1007/978- 3-031-15931-2_47.

[3] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, ‘‘How does NLP benefit legal system: A summary of legal artificial intelligence,’’ in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, 2020, pp. 5218–5230.

[4] O. A. AlcántaraFrancia, M. Nunez-del-Prado, and H. Alatrista-Salas, ‘‘Survey of text mining techniques applied to judicial decisions prediction,’’ Appl. Sci., vol. 12, no. 20, p. 10200, Oct. 2022.

[5] V. Tran, M. Le Nguyen, S. Tojo, and K. Satoh, ‘‘Encoded summarization: Summarizing documents.