Journal of Applied Informatics and Computing (JAIC)
Vol.10, No.2, April 2026, pp. 1513~1519
e-ISSN : 2548-6861

Dual Encoder Contrastive Similarity for Legal Regulation Retrieval in Case-Based Reasoning

Adil Priman Hati Hulu [1*], Anzas Ibezato Zalukhu [2]
[1] Informatics Engineering, STMIK Methodist Binjai
[2] Information Systems, STMIK Methodist Binjai
adilprimanhatihulu@gmail.com [1], anzaszalukhu@stmikmethobi.ac.id [2]

Article Info

Article history:

Received 2026-02-13
Revised 2026-03-03
Accepted 2026-04-08

Keywords:

Case Based Reasoning,
Contrastive Similarity Learning,
Dual Encoder,
Legal Regulation Retrieval,
Document Similarity.

ABSTRACT

The increasing volume and complexity of legal regulations create challenges in retrieving relevant regulatory documents using conventional keyword-based approaches. Effective similarity assessment requires consideration of structural attributes, contextual content, and temporal characteristics. Case-Based Reasoning (CBR) has been widely applied in legal decision-support systems due to its capability to retrieve similar cases based on prior knowledge; however, conventional CBR approaches commonly rely on manually weighted attributes and static similarity measures, limiting their ability to capture complex semantic relationships within heterogeneous regulatory data. This study proposes the integration of a Dual Encoder architecture with Contrastive Similarity Learning to improve similarity measurement within a CBR framework for legal regulation retrieval. Regulatory documents are represented using categorical, numerical, and textual features and transformed into dense embeddings through neural representation learning. Contrastive learning is performed using 4.000 constructed positive regulation pairs, divided into 3.200 training pairs and 800 validation pairs, enabling optimization of similarity-aware embedding space. Experimental evaluation on Indonesian legal regulation data demonstrates that the proposed model achieves a Precision@5 of 0.7719 and a Mean Average Precision (MAP) of 0.7334, indicating improved ranking consistency and retrieval relevance. The results show that contrastive representation learning enables adaptive and data-driven similarity modeling, providing an effective and explainable retrieval mechanism for regulatory analysis and decision-support applications.

I. INTRODUCTION

Legislation is a formal source of law that develops dynamically along with changes in social needs and public policy, thus forming a complex and interrelated regulatory structure [1]. In legal practice, searching for relevant regulations does not only depend on keyword matching, but also requires measuring similarity based on context, type of regulation, time of publication, and substance of the regulation [2].

Case Based Reasoning (CBR) is an artificial intelligence approach that is suitable for the legal domain because it solves problems by utilizing previous experiences or documents that have a certain level of similarity [3], [4]. However, CBR performance still uses conventional similarity mechanisms based on manual weighting and simple distance metrics, which are less adaptive to complex relationships between regulatory documents [5], [6].

On the other hand, Contrastive Similarity Learning (CSL) has the ability to learn similarity representations adaptively based on data by bringing relevant pairs (Positive Pair) closer and moving irrelevant pairs (Negative Pair) away in the embedding space [7], [8]. Although effective in various document retrieval tasks, the integration of CSL into CBR models for legal regulation similarity search is still limited [9]. Therefore, this study proposes a CSL-based CBR approach to improve the quality of legal regulation similarity measurement in a data-driven and adaptive manner.

Previous studies in legal document retrieval have employed transformer-based embedding models such as BERT and Sentence Transformer to capture semantic similarity. However, most approaches focus solely on textual representations without integrating structured regulatory attributes or adaptive retrieval mechanisms within Case-Based Reasoning frameworks. Unlike prior works, this study integrates contrastive representation learning with dual encoder architecture and CBR retrieval, enabling adaptive similarity learning that combines structural and textual regulation characteristics.

Unlike transformer-only approaches that rely solely on semantic textual embeddings, the proposed CSL-CBR framework integrates structured regulatory metadata with contrastive similarity optimization, enabling adaptive similarity learning within an explainable Case-Based Reasoning retrieval process.

II. METHOD

This study proposes an approach to improve the quality of similarity measurements of legal regulations by integrating Contrastive Similarity Learning (CSL) into the Case Based Reasoning (CBR) model. This approach is used to overcome the limitations of conventional CBR methods which still rely on static similarity measurements based on manual weighting and simple distance metrics [10].

In general, the developed approach utilizes structural and textual embeddings that are concatenated and projected into a shared embedding space through a linear projection layer, which are then studied using a neural network-based dual encoder model [11]. The similarity learning process is carried out contrastively by forming relevant (positive pair) and irrelevant (negative pair) data pairs. The resulting embedding results are then used in the CBR retrieval stage to recommend regulations that have a high level of similarity to the query regulation [12]. Figure 1 shows the general stages of the research flow.

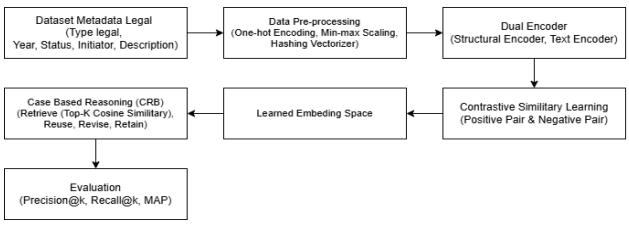

Figure 1 Research Flow Stages

Figure 1 shows the research flowchart, which begins with collecting a dataset of Indonesian laws and regulations. The data used includes the type of law, year of issuance, regulatory status (active, revoked, revised), initiator, and description of the legislation.

Next, a data preprocessing stage is performed to transform the data into a numerical representation that can be processed by the learning model. Categorical attributes are represented using One Hot Encoding (OHE), while textual attributes are represented using a hashing vectorizer. This stage aims to generate initial feature vectors that reflect the structural and textual characteristics of each statutory regulation.

The resulting data vectors are then processed into a dual encoder model, consisting of a structural encoder and a text encoder implemented in a single neural network architecture. This model aims to project the feature vectors into a more representative, low-dimensional embedding space.

In the next stage, the resulting embedding is used in the contrastive similarity learning process. In this stage, the model is trained with positive and negative data pairs, where positive pairs represent cases with high similarity, while negative pairs represent irrelevant cases. The contrastive learning approach aims to reduce the embedding distance between similar cases and increase the embedding distance between dissimilar cases, thus improving the quality of the similarity representation.

The learned embeddings are then utilized within the Case-Based Reasoning (CBR) mechanism. In the retrieval phase, cosine similarity is applied to identify the Top-K regulations most relevant to a given query. The retrieved cases are further processed through the reuse, revise, and retain stages of the CBR cycle, enabling adaptive recommendation based on previously stored regulatory cases.

The final stage is a model performance evaluation to gauge the effectiveness of the search and similarity learning processes. The evaluation is conducted using the Precision@K metrics to measure the accuracy of the search results, Recall@K to measure the completeness of the retrieved relevant results, and Mean Average Precision (MAP) to assess the overall quality of the search results ranking.

The workflow presented in Figure 1 outlines the overall research procedure, while the internal operational mechanism of the proposed similarity learning system is further described through the system architecture illustrated in Figure 2.

To provide a clearer representation of how similarity learning and retrieval processes are implemented, Figure 2 presents the architecture of the proposed CSL-CBR framework. The architecture illustrates the interaction between preprocessing, dual encoder representation learning, contrastive optimization, and the Case-Based Reasoning retrieval mechanism within a unified regulatory similarity learning system.

Figure 2 illustrates the system architecture of the proposed CSL-CBR framework consisting of training and retrieval phases. During the training phase, legal regulation data are preprocessed and encoded using a dual encoder architecture optimized through contrastive learning with NT-Xent loss to construct a similarity-aware embedding space. In the retrieval phase, the learned embeddings are utilized within the Case-Based Reasoning mechanism using cosine similarity to retrieve the Top-K most relevant regulations. System performance is subsequently evaluated using Precision@K, Recall@K, and Mean Average Precision (MAP).

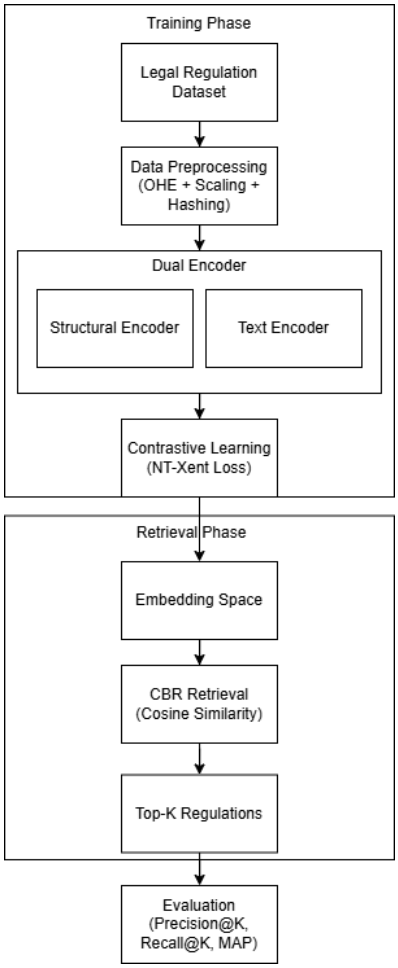

Figure 2 System Architecture of Proposed CSL-CBR

A. Data collection

The dataset used in this study consists of national regulatory documents obtained from Kaggle with the title "law-indonesia-pdf" [13]. The dataset includes various types of regulations, such as laws, government regulations, presidential regulations, ministerial regulations, and regional regulations issued by authorized institutions [14].

The collected dataset consists of multiple regulation categories including laws, government regulations, presidential regulations, ministerial regulations, and regional regulations. The distribution of regulation types is relatively imbalanced, with ministerial regulations forming the dominant portion of the dataset, which may introduce retrieval bias toward frequently occurring regulation categories.

B. Data Preprocessing

The data preprocessing stage aims to transform raw data from legislation into a numerical representation that can be processed by a machine learning model [15]. Data preprocessing is performed on categorical, numerical, and textual attributes to produce feature vectors that represent the structural and substantive characteristics of each regulation.

Numeric attributes in the form of publication years are first converted to numeric types. Handling of missing values is done by filling in the value using the median value of the adjacent attribute [16]. The following is the formula for handling missing values:

$$x_i = \begin{cases} median\ (X), & if\ x_i\ is\ NaN \\ x_i \end{cases} \quad (1)$$

Categorical attributes such as type of regulation, initiator, and legal status cannot be directly processed by the model. Therefore, the One Hot Encoding (OHE) technique is used to represent each category as a binary vector [17]. The following is the OHE formula:

$$b_k = \begin{cases} 1, if\ k = j \\ 0, etc \end{cases} \quad (2)$$

If a categorical attribute has *n* unique categories, then each value is represented as an *n -dimensional vector* , with one element equal to 1 and the rest equal to 0.

All numerical features were normalized using min-max scaling to prevent dominance over high-dimensional textual vectors [18]. The following is the min max scaling formula:

$$x_i^1 = \frac{x_i - x_{min}}{x_{max} - x_{min}} \quad (3)$$

Normalization is applied to ensure that numerical attributes do not dominate high-dimensional textual representations during embedding learning.

Textual attributes in the form of descriptions of rules in the about column are represented using the hashing vectorizer method. This approach maps character n-grams into a fixed-dimensional vector space without building an explicit dictionary [17]. The following is the hashing vectorizer formula:

$$h : t \rightarrow \{0,1, \dots, d-1\} \quad (4)$$

After all attributes are processed, categorical, numeric, and textual features are combined (concatenation) to form a final vector representation of each statutory regulation. The following is the case vector formula:

$$x_{case} = [x_{cat} || x_{num} || x_{text}] \quad (5)$$

To balance heterogeneous feature influence, structural features were scaled by a factor of 0.4 while textual representations were amplified by a factor of 1.2. This weighting strategy prevents sparse structural metadata from dominating dense textual representations during similarity learning.

C. Dual Encoder

The proposed model employs a heterogeneous dual encoder architecture consisting of two independent multilayer perceptron (MLP) encoders designed to process structural and textual regulation features separately. The structural encoder learns representations from categorical and numerical regulatory metadata, while the text encoder captures semantic characteristics derived from textual descriptions. Unlike

Siamese architectures that share parameters, both encoders operate with independent weights, allowing modality-specific representation learning before fusion through embedding concatenation and projection into a shared similarity space. The model was implemented and trained using the PyTorch framework in the Google Colab environment.

Each encoder is implemented as a multilayer perceptron consisting of two fully connected layers followed by ReLU activation and dropout regularization with a probability of 0.5. The encoders independently transform structural and textual feature vectors into 128-dimensional latent representations before projection into a shared embedding space [19]. The following is the embedding vector formula:

$$u = f(x_i) \in R^2 \quad (6)$$

By employing a dual encoder architecture with independent encoders, the model learns complementary similarity representations from structural and textual regulatory features within a unified embedding space. This design enables adaptive similarity measurement while preserving modality-specific information, resulting in efficient and robust regulatory similarity estimation.

D. Contrastive Similarity Learning (CSL)

Contrastive Similarity Learning (CSL) is used in this study to learn the embedding representation of legal regulations by forming similarities between relevant pairs (positive pairs) and irrelevant pairs (negative pairs) [20]. The following is the formula for CSL:

$$sim(z_i, z_j) = \frac{z_i z_j}{||\, z_i||\, ||z_j||} \quad (7)$$

Positive pairs are formed based on similarities in the structural characteristics of regulations, namely the type of regulation, legal status, and proximity of publication years.

A regulation pair is labeled as positive if both regulations share identical regulation type and legal status, and the difference in publication year does not exceed ten years.

Meanwhile, negative pairs are formed by randomly selecting pairs of regulations that do not meet the positive pair criteria. These pairs represent regulatory documents deemed irrelevant to each other in the context of similarity learning.

Negative samples are implicitly generated using in-batch negative sampling, where all non-matching regulations within the same training batch serve as contrastive negative examples.

Model optimization employs the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss with a temperature parameter $\tau = 0.5$ to enhance embedding discrimination during contrastive learning.

E. Case Based Reasoning (CBR)

Case Based Reasoning (CBR) is used as a retrieval model to recommend laws and regulations that have the highest level of similarity to the query regulations based on the CSL embedding results [21]. The following is the CBR model formula:

$$sim(q, x_i) = \frac{z_q z_i}{||\, z_q||\, ||z_i||} \quad (8)$$

The regulation with the highest similarity value is selected as the recommendation result using Top-k Selection [22]. The following is the Top-k selection formula:

$$Top - k = \frac{agr\ \max Sim\ (q, x_i}{x_i} \quad (9)$$

F. Evaluation

The evaluation is performed using three metrics: Precision@K, Recall@K, and Mean Average Precision (MAP) [23]. Precision@K is used to measure the proportion of relevant documents among the top K results: Here is the formula for Precision@K:

$$Precision@K = \frac{Total\ Number\ of\ relevan\ in\ Top - K}{K} \quad (10)$$

Recall@K is used to measure how many relevant regulations were found. The Recall@K formula is as follows:

$$Precision@K = \frac{Total\ regulations\ relevan\ in\ Top - K}{Total\ regulations\ relevan} \quad (11)$$

Mean Average Precision (MAP) is used to measure how well a model finds relevant results. The MAP formula is as follows:

$$MAP = \frac{1}{Q}\sum_{q=1}^{Q} AP_1 \quad (12)$$

III. EXPERIMENTAL SETUP

The experiment utilized 2,000 regulatory documents, from which 4,000 positive regulation pairs were constructed for contrastive learning. These pairs were divided into 3,200 training pairs and 800 validation pairs.

The model was trained using a batch size of 64 for contrastive optimization with a learning rate of 5×10^{-4} using the Adam optimizer. Training was conducted for a maximum of 50 epochs with an early stopping strategy employing a patience value of five epochs based on validation loss monitoring. The best-performing model weights were automatically saved and used during evaluation.

IV. RESULTS AND DISCUSSION

The approach used a contrastive similarity learning (CSL)-based dual encoder integrated into a Case-Based Reasoning (CBR) model to measure regulatory similarity. The dataset was divided into 80% training data and 20% validation data, resulting in 3.200 training data sets and 800 validation data sets from 4.000 positive pairs.

A. Results

Experiments show that the contrastive similarity learning-based dual encoder model is capable of constructing an effective representation of legislative regulation embeddings for retrieval purposes in the Case-Based Reasoning (CBR) model. The training process resulted in a decrease in the training loss value from 4.3240 at epoch 1 to 3.9998 at epoch 10, indicating a stable and convergent representation learning process. Table I shows the experimental results.

TABLE I
EXPERIMENTAL RESULTS

Epoch	Train loss	Validation loss
Epoch 1	4.3240	4.3715
Epoch 2	4.0318	4.3410
Epoch 3	4.0196	4.3372
Epoch 4	4.0135	4.3337
Epoch 5	4.0089	4.3333
Epoch 6	4.0038	4.3338
Epoch 7	4.0027	4.3350
Epoch 8	4.0013	4.3362
Epoch 9	4.0039	4.3403
Epoch 10	3.9998	4.3411

The training process demonstrates a consistent reduction in training loss from 4.3240 at epoch 1 to 3.9998 at epoch 10, indicating stable convergence of the contrastive representation learning process. The validation loss decreased steadily during the early training stages and reached its lowest value of 4.3333 at epoch 5 before showing slight fluctuations in subsequent epochs. This behavior suggests that the model successfully learned discriminative embedding representations while maintaining generalization capability without significant overfitting.

As no further improvement in validation loss was observed after several epochs, an early stopping mechanism was applied to preserve the model parameters corresponding to the best validation performance. The following table presents the evaluation results of the proposed contrastive similarity learning integrated with the Case-Based Reasoning framework.

The observed convergence behavior confirms the effectiveness of contrastive optimization in organizing regulatory embeddings into a stable similarity space.

TABLE II
CSL PERFORMANCE EVALUATION RESULTS WITH CBR

Metric	Value
Precision@5	0.7719
Recall@5	0.0321
MAP	0.7334

The model's performance in the regulatory similarity search task was evaluated using the Precision@5 metric, with a value of 0.7719, indicating that more than 77% of the top five recommended regulations were highly relevant to the query regulation. This indicates that the model is capable of consistently providing accurate regulatory recommendations at the beginning of the search results.

The Mean Average Precision (MAP) metric shows a value of 0.7334, which indicates the model's ability to rank documents accurately based on the level of embedding similarity.

The Recall@5 metric shows a value of 0.0321, indicating that a small portion of the overall regulations are highly relevant and successfully retrieved into the top five results. This value is influenced by the characteristics of the regulatory domain, which has a large number of relevant documents for each query, thus naturally suppressing the recall value at a smaller K limit.

Overall, the high Precision@5 and MAP values confirm that the contrastive similarity learning-based dual encoder approach is able to significantly improve the quality of regulatory similarity measurements, especially in producing accurate and relevant recommendations at the top ranking in the CBR model.

TABLE III
COMPARISON WITH BASELINE METHODS

Method	Precision@5	Recall@5	MAP
TF-IDF Baseline	0.6800	0.0317	0.0250
Dual Encoder without Contrastive Similarity Learning	0.7596	0.0326	0.0283
CSL-CBR Model	0.7719	0.0321	0.7334

Although the improvement in MAP appears substantially larger compared to baseline methods, this behavior is expected due to differences in optimization objectives. Traditional TF-IDF retrieval and the dual encoder without contrastive learning do not explicitly optimize document ranking relationships within the embedding space. Consequently, relevant regulations may appear within top-ranked results but are not consistently positioned across the entire ranking list, resulting in low MAP values.

In contrast, the proposed model employs contrastive learning with NT-Xent loss, which explicitly minimizes embedding distance between semantically related regulation pairs while maximizing separation from irrelevant cases. This optimization directly structures the global similarity space, leading to consistent ranking of relevant regulations across retrieval positions and significantly improving MAP performance.

Therefore, the observed MAP improvement reflects the effectiveness of similarity-aware representation learning rather than evaluation inconsistency.

B. Explainability Analysis

The proposed CSL-CBR framework provides explainable retrieval behavior through embedding similarity analysis derived from contrastive representation learning. Unlike conventional similarity approaches that rely on manually assigned attribute weights, similarity in the proposed system is computed using cosine similarity within a learned embedding space combining structural and textual regulation features.

Explainability is achieved by analyzing similarity contributions originating from two complementary information sources. Structural embeddings capture regulatory metadata such as regulation type, legal status, institutional initiator, and publication year, while textual embeddings represent semantic content derived from regulation descriptions. During retrieval, regulations are recommended based on proximity within this shared embedding space, allowing interpretation of similarity relationships between cases.

For example, a ministerial regulation used as a query tends to retrieve regulations issued within a similar institutional context and temporal range. Structural similarity ensures regulatory consistency, while textual embeddings refine semantic relevance related to policy substance. Consequently, retrieved cases can be interpreted as similar not only lexically but also structurally and contextually.

This embedding-based similarity decomposition enables users to understand why specific regulations are retrieved as relevant cases, thereby supporting transparent and interpretable decision-support processes in legal regulation analysis.

C. Computational Complexity

The computational efficiency of the proposed system is primarily influenced by embedding generation and similarity retrieval stages. During the training phase, contrastive learning requires pairwise similarity computation within mini-batches; however, this process is performed offline and does not affect retrieval efficiency during deployment.

After training, all regulation embeddings are precomputed and stored in the case base. Consequently, the retrieval process only requires cosine similarity computation between the query embedding and stored regulation embeddings. The inference complexity can therefore be expressed as $O(n \cdot d)$, where n represents the number of stored regulations and d denotes embedding dimensionality.

Compared with cross-encoder architectures that require repeated neural inference for each query document pair, the proposed dual encoder framework enables efficient real-time retrieval. This characteristic makes the CSL-CBR approach suitable for large-scale legal regulation repositories and practical decision-support systems requiring fast similarity search.

D. Discussion

Experimental results show that integrating a Contrastive Similarity Learning (CSL)-based dual encoder into a Case-Based Reasoning (CBR) model significantly improves the quality of similarity measurements of legal regulations. A consistent decrease in loss values during training indicates that the model successfully learns informative feature representations compared to conventional similarity approaches based on static features.

A high Precision@5 value indicates that the model is highly effective in recommending relevant laws and regulations at the top of the rankings, which is an important characteristic in legal document retrieval systems. Conversely, a relatively low Recall@5 value can be explained by the nature of the legal regulation domain, which has a large number of relevant documents for each query, so that only a small portion can be covered in the top five results.

Overall, the high MAP value confirms that the model is not only capable of finding relevant regulations but also of ranking them appropriately based on their similarity level. This demonstrates that the CSL approach in the proposed CBR model is more adaptive and representative than conventional similarity models.

However, this research still has limitations, particularly in the heuristic-based positive pair formation and the use of simple text representations. Further development can be done by utilizing transformer-based language models to improve embedding generalization and expand the scope of document relevance.

Error analysis reveals that retrieval errors mainly occur among regulations sharing similar structural metadata but addressing different legal domains. Regulations issued within close temporal ranges tend to produce embedding overlap despite differences in regulatory substance.

Future work may include comparison with probabilistic retrieval models such as BM25 and transformer-based encoders to further evaluate semantic retrieval robustness.

V. CONCLUSION

This study proposes a contrastive similarity learning approach integrated with a Dual Encoder architecture within a Case-Based Reasoning framework for legal regulation retrieval. The proposed method addresses limitations of conventional CBR systems that rely on manually defined similarity measures by introducing data-driven embedding representations capable of capturing structural and semantic relationships among regulatory documents.

Experimental results demonstrate that the proposed CSL-CBR framework achieves a Precision@5 of 0.7719 and a Mean Average Precision (MAP) of 0.7334, indicating improved retrieval accuracy and ranking consistency compared with traditional similarity-based approaches. The integration of contrastive learning enables the embedding space to adaptively model regulatory similarity, improving discrimination between relevant and irrelevant regulations.

Furthermore, the embedding-based retrieval mechanism provides interpretability through similarity relationships learned from both structural metadata and textual content, supporting transparent decision-support processes in legal regulation analysis. The proposed framework also maintains computational efficiency suitable for large-scale regulatory repositories.

Future work may explore transformer-based language models and cross-domain regulatory datasets to further enhance semantic representation and generalization capability in legal information retrieval systems.

REFERENCES

[1] MF Berry, "Formation of the Theory of Legislation," *Muhammadiyah Law Review* , vol. 2, no. 2, pp. 87–91, Jan. 2021, doi: 10.24127/lr.v2i2.1461.

[2] A. Fitryantica, "Harmonization of Indonesian Legislation through the Omnibus Law Concept," *Gema Keadilan* , vol. 6, no. 3, pp. 300–316, Dec. 2019, doi: 10.14710/gk.2019.6751.

[3] B. Bhavishya, A. Shukla, and M. Aggarwal, "Automating Legal Expertise: A Rule-Based Approach to Legal Reasoning Systems," in *2025 2nd International Conference on Computational Intelligence, Communication Technology and Networking (CICTN)* , Feb. 2025, pp. 868–872. doi: 10.1109/CICTN64563.2025.10932494.

[4] W. Yuan, "Design and Implementation of Intelligent Reasoning Engine Based on Legal Framework Network Database," in *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)* , Jan. 2022, pp. 1000–1003. doi: 10.1109/ICSSIT53264.2022.9716419.

[5] A. Yan, H. Yu, and D. Wang, "Case-based reasoning classifier based on learning pseudo metric retrieval," *Expert Systems with Applications* , vol. 89, pp. 91–98, Dec. 2017, doi: 10.1016/j.eswa.2017.07.022.

[6] D. C. Corrales, A. Ledezma, and J. C. Corrales, "A case-based reasoning system for recommendation of data cleaning algorithms in classification and regression tasks," *Applied Soft Computing* , vol. 90, p. 106180, May 2020, doi: 10.1016/j.asoc.2020.106180.

[7] M. Chen, Q.-Y. Liu, S. Chen, Y. Liu, C. Zhang, and R. Liu, "XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power Systems," *IEEE Access* , Jan. 2019, doi: 10.1109/ACCESS.2019.2893448.

[8] G. Izacard *et al.* , "Unsupervised Dense Information Retrieval with Contrastive Learning," Aug. 29, 2022, 2112.09118. doi: 10.48550/arXiv.2112.09118.

[9] E. Amador-Domínguez, E. Serrano, D. Manrique, and J. Bajo, "A Case-Based Reasoning Model Powered by Deep Learning for Radiology Report Recommendation.," *International Journal of Interactive Multimedia and Artificial Intelligence* , vol. 7, pp. 15–26, Dec. 2021, doi: 10.9781/ijimai.2021.08.011.

[10] W. Xu, Y. Huang, S. Song, B. Chen, and X. Qi, "A novel online combustion optimization method for boiler combining dynamic modeling, multi-objective optimization and improved case-based reasoning," *Fuel* , vol. 337, p. 126854, Apr. 2023, doi: 10.1016/j.fuel.2022.126854.

[11] H. Zamani and M. Bendersky, "Multivariate Representation Learning for Information Retrieval," in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval* , in SIGIR '23. New York, NY, USA: Association for Computing Machinery, Jul. 2023, pp. 163–173. doi: 10.1145/3539618.3591740.

[12] K. Amin, G. Lancaster, S. Kapetanakis, K.-D. Althoff, A. Dengel, and M. Petridis, "Advanced Similarity Measures Using Word Embeddings and Siamese Networks in CBR," in *Intelligent Systems and Applications* , Y. Bi, R. Bhatia, and S. Kapoor, Eds., Cham: Springer International Publishing, 2020, pp. 449–462. doi: 10.1007/978-3-030-29513-4_32.

[13] H. Sugiharto, "Law Indonesia Dataset," Kaggle. Accessed: Feb. 10, 2026. [Online]. Available: https://www.kaggle.com/datasets/hermansugiharto/law-indonesia

[14] I. Chalkidis, E. Fergadiotis, P. Malakasiotis, and I. Androutsopoulos, "Large-Scale Multi-Label Text Classification on EU Legislation," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* , A. Korhonen, D. Traum, and L. Màrquez, Eds., Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 6314–6322. doi: 10.18653/v1/P19-1636.

[15] H. Zhong, C. Xiao, C. Tu, T. Zhang, Z. Liu, and M. Sun, "How Does NLP Benefit Legal Systems: A Summary of Legal Artificial Intelligence," May 18, 2020, 2004.12158. doi: 10.48550/arXiv.2004.12158.

[16] GEAPA Batista and MC Monard, "An analysis of four missing data treatment methods for supervised learning," *Applied Artificial Intelligence* , vol. 17, no. 5–6, pp. 519–533, May 2003, doi: 10.1080/713827181.

[17] SG Tendulkar, "Efficient Nearest Neighbor Search in Large-Scale Text Data Using Optimized Locality Sensitive Hashing in Distributed Computing Environments," in *2025 International Conference on Advanced Computing Technologies (ICoACT)* , Mar. 2025, pp. 1–6. doi: 10.1109/ICoACT63339.2025.11005204.

[18] J. Han, J. Pei, and H. Tong, *Data mining: concepts and techniques* , Fourth edition. Cambridge, MA, United States: Morgan Kaufmann Publishers, an imprint of Elsevier, 2023. doi: 10.1016/C2013-0-18660-6.

[19] L. Roḳaḥ and OZ Maimon, *Data mining with decision trees: theory and applications* , 2nd edition. in Series machine perception and artificial intelligence, no. 81. New Jersey London Singapore Beijing: World Scientific, 2015.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A Simple Framework for Contrastive Learning of Visual Representations," in *Proceedings of the 37th International Conference on Machine Learning* , PMLR, Nov. 2020, pp. 1597–1607. doi: 10.48550/arXiv.2002.05709.

[21] L. Fei and Y. Feng, "A novel retrieval strategy for case-based reasoning based on attitudinal Choquet integrals," *Engineering Applications of Artificial Intelligence* , vol. 94, p. 103791, Sep. 2020, doi: 10.1016/j.engappai.2020.103791.

[22] E. Keogh, "Nearest Neighbor," in *Encyclopedia of Machine Learning and Data Mining* , C. Sammut and G.I. Webb, Eds., Boston, MA: Springer US, 2017, pp. 897–897. doi: 10.1007/978-1-4899-7687-1_579.

[23] N. Thakur, N. Reimers, A. Rücklé, A. Srivastava, and I. Gurevych, "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models," Oct. 21, 2021, 2104.08663. doi: 10.48550/arXiv.2104.08663.