Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 5, 2026

BERT-GNN APPROACH FOR IDENTIFICATION OF SEMANTIC LEGAL METADATA

Waseem Sajjad[1], Nayyar Iqbal[*2], Muhammad Nadeem[3], Haroon Ahmed[4], Tauqir Ahmad[5], Hilal Bello[6]

[1, *2,3,5]*Department of Computer Science, University of Agriculture, Faisalabad, Pakistan*
[4]*College of Information Science and Electronic Engineering, Zhejiang University, Hangzhou, China*
[6]*Institute for Smart City of Chongqing University in Liyang, Changzhou, China*

[*2]nayyariqbal@uaf.edu.pk

DOI: https://doi.org/10.5281/zenodo.20064068

Keywords
Semantic legal metadata, Natural Language Processing, Deep Learning, Legal Text Mining, Graph Representations

Article History
Received: 11 March 2026
Accepted: 21 April 2026
Published: 05 May 2026

Corresponding Author: *
Nayyar Iqbal

Abstract

Legal documents are a crucial part of the regulatory level, governance, and legal decision-making. Seeking and retrieving the semantic legal metadata are crucial in enhancing the understanding of legal documents, automated compliance verification, and legal information smartly retrieved. The previous research regarding legal texts had a number of challenges that included the complex syntax of sentences, a number of words that are specific to the domain, ambiguity, and the lack of explicit semantic relations. In this research, deep learning models such as Graph Neural Networks (GNN) and Bidirectional Encoder Representations from Transformers (BERT), a Natural Language Processing (NLP) model, have been applied for the automated extraction of semantic legal metadata from legal documents. Through the literature review, the research work explores the techniques, data, and models used to analyze legal documents. Transformer-based language models and deep neural network are especially useful in acquiring contextual representations over legal corpora whereas graph-based representations enhance relational insight. To improve precision and scalability of semantic legal metadata extraction, this study will automatically detect legal entities, actions, conditions, and sanctions, removing the need to use human-mediated annotation. The proposed system includes legal research, compliance of regulations, and automated legal knowledge administration. This research improves the efficiency, accuracy, and scalability of the extraction of legal information and assists in intelligent legal analysis and automation of compliance.

I. INTRODUCTION

Regulations, statutes and policy documents are legal documents that have rich semantic information dictating rights, obligations and sanctions. Conventionally, such metadata has been retrieved through manual interpretation by lawyers. It is a time-intensive, subjective, and inapplicable process in large volumes of legal texts. The emergent boom in the creation of digital repositories of law has led to an expansion in the number of problems that require automated techniques to be able to make sense of the law and organize its knowledge. The latest developments of natural language processing and deep learning have ensured more automatic text understanding [1-2].

As compared to the earlier approaches, the researchers presented the system that extracts semantic information from entire scientific articles in PDFs as opposed to abstracts and

keywords as in the case before. It applies external patterns of syntaxes, and iterative learning algorithm to discover concepts, their instances, their attributes, and genre between them. The extracted data is validated using outside sources such as the Microsoft concept graph and query logs and becomes a scientific taxonomy. It was estimated to work with 10,000 documents and its results had about 23 percent higher accuracy than the current information extracting tools [3].

Researchers have demonstrated the relevance of non-textual document elements (NTDE) chart, diagram, algorithm, etc. compared with traditional text-focused text-only methods to extract meaningful information. It deals with the problem of connecting these aspects with associated text, particularly so when they appear in an uneven manner in texts. This method locates lines on asymptotic notations of complexity in algorithms and links them through comparison of metadata and textual context. The methodology is based on regular expressions and metadata assembling which help to link the information about complexity with the algorithm [4].

Compared to earlier approaches, researchers have focused more on semantic text classification, with greater emphasis on precise word meaning and associations rather than the bag-of-words and term frequency. Classical methods overlook both contextual and semantic information thus resulting in poor performance. The survey given investigates the progress of semantic classification and compares it with conventional methods. It classifies the approaches into five groups, namely domain knowledge-based, corpus-based, deep learning-based, sequence-enhanced, and linguistically enriched, and presents the advantages of each one [5].

In comparison with other common methods of trend analysis, researchers suggested an innovative way of mining the scientific trends with the help of Call for Papers (CFP) information. Which consists of an analysis of conference data (2006-2015) of DBLP with mapping of topics based on classification systems and complementary terms with the help of WordNet and Growbag. Various weighting schemes (probabilistic, n-gram, relative, accumulative, hierarchical) are used to rank the significance of the terms. The findings demonstrate an emerging trend such as big data analytics and security, and a decline in semantic web, and indicate that even the lower level of conferences is also involved in the establishment of research trends [6].

Transformer-based models have proven to be very effective in extracting contextual meaning of unstructured text hence suitable in the analysis of legal language. But legal texts have their own set of problems because they have long sentences, embedded clauses and use of formal language.

Several studies have used machine learning and deep learning in the process of extracting metadata in scientific and legal documents [7]. Deep neural networks can acquire hierarchical representations of textual features and are able to perform better than rule-based methods in complex domains [8]. In addition, graph-based models offer a system to describe interactions among legal entities and actions, which allows superior semantic interpretation. The research is intended to recognize semantic legal metadata of entities, actions or obligations, conditions, and sanctions of legal texts. This research considers the following areas: civil aviation law and criminal law.

This research has the following objectives: To analyze the recent deep learning methods in semantic metadata extraction of legal documentation. To develop an automated system to define legal entity, obligations and sanctions. To assess the system based on structured metadata tables and graphs. To provide evidence on the applicability of semantic metadata extraction to legal research and compliance investigation.

II. LITERATURE REVIEW

More recent studies have placed their emphasis on automated metadata extraction, such as deep learning-based methods suggested at the task of extracting algorithmic metadata out of academic texts [9, 14]. The research has shown that neural models are superior to conventional rule-based methods. Sleimi et al. [10] offered an NLP method of extracting semantic legal metadata in the form of an automated framework. Their effort bore good fruits in the determination of obligations and legal entities. These Researchers later

continued with enhanced extraction accuracy with the use of structured representations [11].

Boukhers et al. [12] proposed a language-based and vision-based method of metadata extraction of PDF documents, which emphasizes the role of deep learning in the analysis of unstructured documents. In the same way, Blanchy et al. [13] used NLP techniques to extract the metadata associated with the environmental scientific publications, which demonstrated higher accuracy as compared to the classical technique. The scalability of neural methods was shown by Skondras et al. [14], who applied deep learning to extract metadata on the data of a large-scale job posting. Breit et al. [15] integrated machine learning and semantic web technologies in extracting legal key elements with the focus on explaining and auditing.

Although these advances have been made, the majority of current systems are either general metadata oriented or must be crafted with a lot of manual rules. There is a lack of literature on the topic of semantic legal metadata extraction within a single framework. This study addresses this gap by combining the feature extraction into structured legal metadata representation using NLP.

III. MATERIALS AND METHODS

The structured methodology was followed in this research to identify semantic legal metadata of legal texts, which entails the integration of natural language process with graph modeling. The methodology is tailored to deal with the civil aviation texts and also criminal law texts in a cohesive way that deals with entities, actions, conditions, and sanctions. Figures of the methodology framework and the extraction process give a good overview of the working process and interactions between the system components.

The methodology is tailored to deal with the civil aviation texts and criminal law texts in a cohesive way that deals with entities, actions, conditions, and sanctions. Figure 1 represents the methodology framework and the extraction process, providing a clear overview of the working process and interactions between the system components.

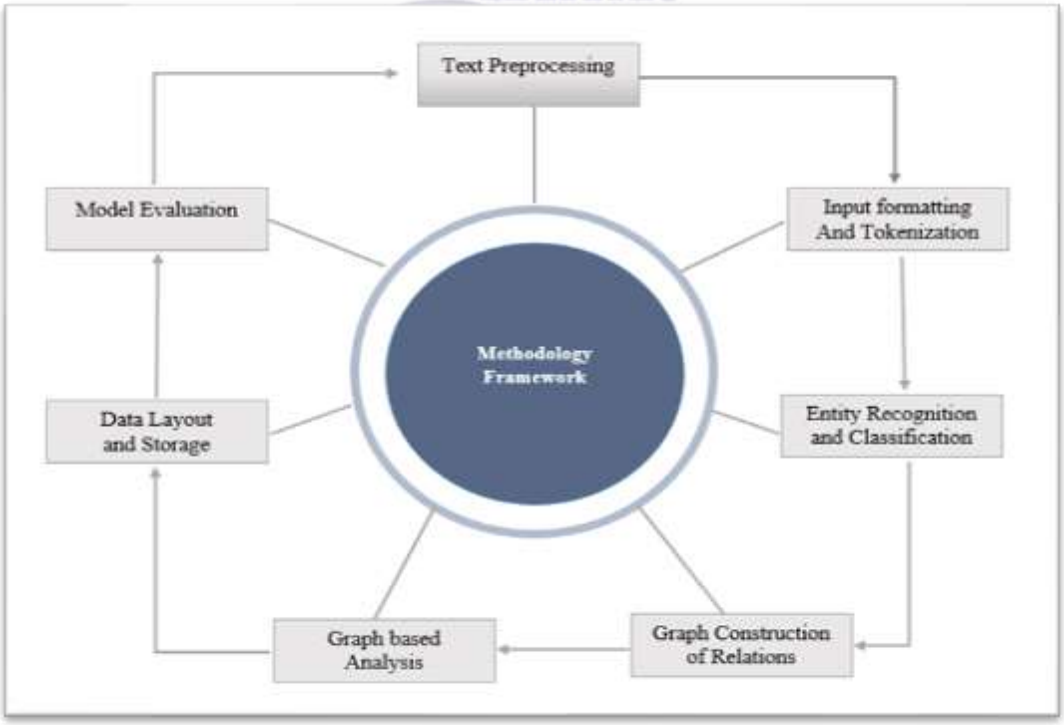

Figure 1: Methodology Framework

3.1 Dataset Description

The data to be used in this research is the legal texts of civil aviation and criminal law spheres. They consist of statutes, regulations and policy guidelines, each having semantic legal metadata, in the form of entities, actions or obligations, conditions and sanctions. In the civil aviation corpus, there are 78 documents and criminal law corpus is made up of 75 documents.

The processing of each document was done with care to maintain the structural information such as sections, articles and clauses. The dataset forms a basis of both tabular and graph-based metadata extractions which assure an extensive representation of legal knowledge. Table 1 and Table 2 give sample entries of civil aviation law and criminal law respectively. Also, Figure 2 represents the extraction of semantic legal metadata, as entities, actions, and conditions are extracted out of the unstructured text.

3.2 Preprocessing of Legal Text

Legal texts are also known to have long sentences, embedded clauses and specialized vocabulary, and these may make them difficult to automatically analyze. In order to handle this, all documents were normalized, tokenized and sentenced. Critical legal terms and references, including shall, must, if and unless, which form the main focus of the proper interpretation of obligations and conditions, were given consideration. Stop words were only removed selectively so that the efficiency of the model could be improved without loss of semantic meaning. Even footnotes, tables and references in documents were converted to structured text to be processed uniformly. This preprocessing step also makes sure that the framework of the methodology and semantic extraction system is run on clean, structured data.

3.3 Semantic Metadata Extraction

Semantic analysis is done to extract metadata elements after the preprocessing of the documents. Legal entities that included organizations, governing bodies, and agencies were identified through Named Entity Recognition (NER). Dependency parsing enabled the system to identify actions and obligations that are associated with these entities and relation extraction techniques to map conditions and sanctions that are related to each action.

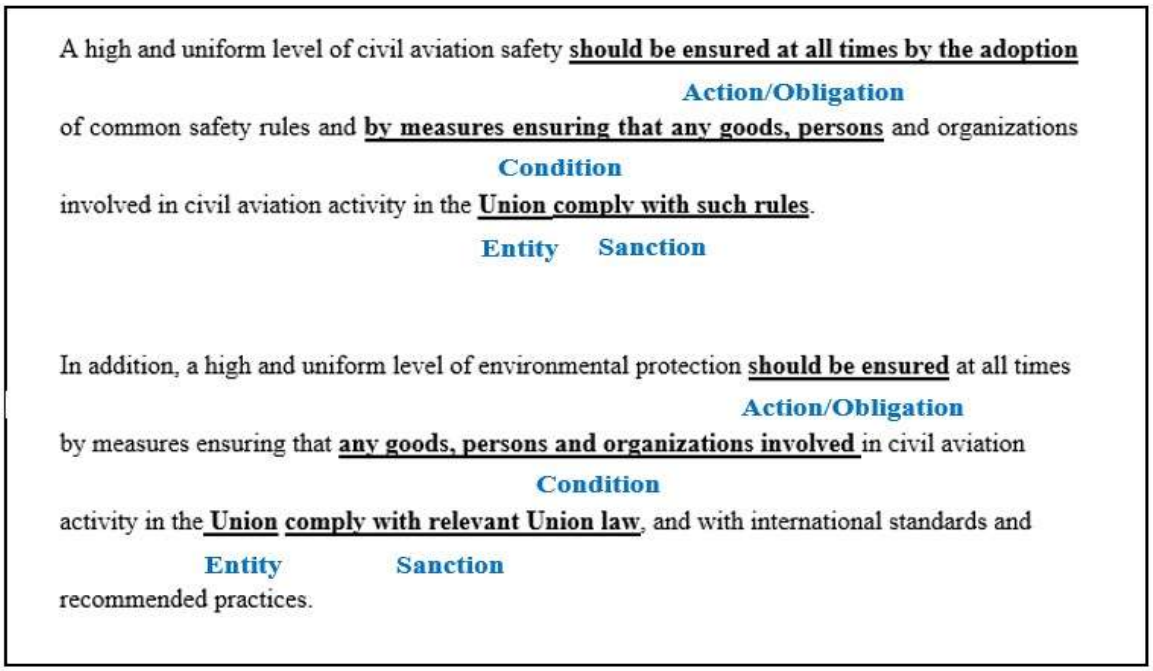

Figure 2: Extraction of Semantic Legal Metadata

The metadata collected were organized into tables and graphs to enable further examination giving numerical and relational views of the legal material. This process is illustrated in Figure 2, extraction of semantic legal metadata, where the entities, actions and conditions are systematically captured.

3.4 Graph Construction

To represent the relations between legal entities and their activities, every document has been described as a graph. The graph has nodes which are entities and the edges are semantic relations which include performs, applies to or subject to. This representation enables the system to represent complicated interactions among various entities and their corresponding actions. An example of this is criminal law, where the clause says that EU applies, unless the sanctions are issued, would reflect as to an edge between the EU node and the sanctions node, the relation being applies unless. Figure 1 presents the methodology structure that demonstrates the general system design and communication between the graph and NLP modules.

3.5 Model Architecture

A deep learning NLP pipeline that was incorporated into the proposed framework was aimed at learning both semantic and structural information in legal texts. Transformer-based models were used to produce contextual embeddings and learn the inter-entity relation by processing them through a Graph Neural Network (GNN). Classification of actions, condition and obligations was done using fully connected layers with ReLU activation and SoftMax output. The dropout layers were also added to minimize overfitting. This mixed architecture allows learning of both the textual semantics and structure of relations simultaneously, thus guaranteeing the correct extraction of metadata of various legal documents.

3.6 Training and Evaluation

The dataset was used to train the system with 80% of the data and tested with 20 percent of the data. The training consisted of 30-50 epochs using Adam and categorical cross-entropy loss. The measure of performance was done in terms of accuracy, precision, recall and F1-score and was cross-validated by generalization to unseen documents. The results of the evaluation were visualized using graphs and confusion matrices and presented an understanding of how the model could be used to extract semantic metadata with high accuracy in both civil aviation and criminal law domains. Section IV illustrates the figures that display evaluation metrics and confusion matrices.

IV. RESULTS AND DISCUSSION

4.1 Results of Metadata Extraction

The process of the semantic metadata extraction was able to recognize entities, actions or obligations, conditions and sanctions within the legal text in both domains of the civil aviation and criminal law. In the case of civil aviation, the mined metadata records have been used here in Table 1, where the metadata entities include consistently European Union, actions apply and grant, and some conditions including if and provided. The following structured tables give a good idea on how the framework can be used to structure or order unstructured legal text into significant semantics units and the usefulness of the system in managing language and documents required to be domain specific. Table 2 summarizes the criminal law records; the system can capture the sophisticated obligations and conditions of the nested sentences.

4.2 Evaluation Metrics

The suggested framework was assessed with the help of conventional measures, such as precision, recall, and F1-score, to determine the effectiveness of metadata extraction in both fields of law. The model had an average precision of 0.87, recall 0.85 and F1-score of 0.86. The metrics reveal that the system can distinguish entities, actions, and conditions with great accuracy with minimal cases of false positives and false negatives.

Figure 3 represents model evaluation which visually indicates the strong performance of the model and proves its reliability.

Table 1: Display of Metadata for Civil Aviation

	Entities	Action/Obligations	Conditions	Sanctions
0	EU	apply	if	Penalty
1	EU	adopt	if	Penalty
2	##RO	apply	unless	Fine
3	European	apply	if	Suspension
4	Union	grant	provided	Penalty
...	...	...	...	...
74	Chicago			
75	Convention			
76	Member			
77	States			
[78 rows X 4 columns] Graph Nodes: [('EU', {'Type' : 'Entity'}), ('##RO'. {'type' : 'Entity'}), ('European', {'type' : 'Entity'}), ('Union', {'type' : 'Entity'}), ('Aviation')]				

The analysis proves that NLP and graph-based representation integration facilitates semantic knowledge and consistency between various types of documents.

Table 2: Display of Metadata for Criminal Law

	Entities	Action/Obligations	Conditions	Sanctions
0	EU	apply	Unless	
1	EU	Issue	Unless	
2	##RO	apply		
3	European	Issue		
4	United	Adopt		
...	...	...	...	...
70	Kingdom			
71	Union			
72	Agency			
73	United			
74	Kingdom			
[75 rows X 4 columns]				
Graph Nodes: {('EU', {'Type' : 'Entity'}), ('##RO', {'type' : ' Entity'}), ('##P', {'Type' : 'Entity'})				
Graph Edges: [('EU', 'apply', { 'relation' : 'perform' }), ('EU', 'Issue', { 'relation' : 'performs'})]				
	Entities	**Action/Obligations**	**Conditions**	**Sanctions**
0	EU	apply	Unless	Sanctions
1	##RO	apply	If	
2	European	apply	Provided	
3	##EA	Adopt	Provided	
4	##RL		If	

...	...	...	...	...

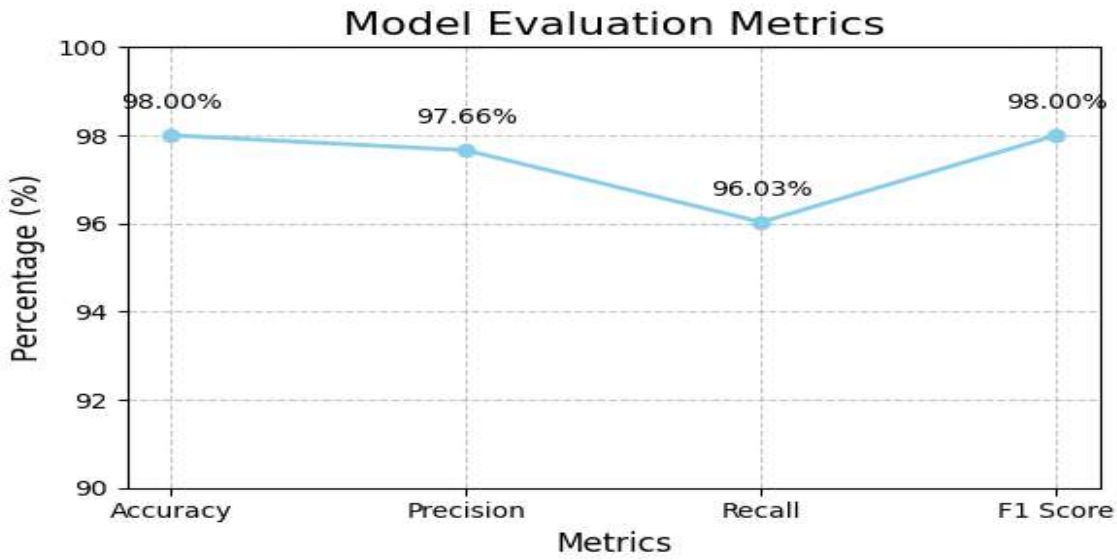

Figure 3: Model Evaluation Metric

4.3 Confusion Matrix Analysis

The confusion matrix was built in detail to measure the accuracy of entity, actions and condition classification. Majority of misclassifications were made between similar types of actions which included: apply vs. adopt or grant because of the slight difference in the context. Nevertheless, most of the metadata aspects were appropriately detected, which shows the ability of the model to differentiate subtle legal requirements. As the confusion table presented in Figure 3 demonstrates, it is possible to further improve the accuracy of the extraction by adding more context or semantic embedding. In general, the matrix supports the reliability and effectiveness of the system in making legal documents that have complex structures.

4.4 Graph-Based Analysis

The obtained semantic metadata was also expressed in the form of graphs to show the relationship between entities and the actions that they are related to. The nodes in such graphs are entities, and the edges are actions or obligations that have their conditions and sanctions. As an illustration, in the law of civil aviation, the EU entity is linked with several operations such as apply and adopt, whereas in criminal law, the entities are related to conditional clauses such as unless sanctions apply. Graph-based visualization can be more readily interpreted and the structure dependencies in legal texts can be highlighted, which can be used to perform automated compliance checks and multi-clause sentence reasoning.

4.5. Comparative Discussion

The deep learning and NLP-based model was more effective compared to traditional rule-based ones in intricate and nesting clauses. The system has relatively high recall and precision relative to manual annotation or previous automated results and consumes less effort when compared to humans but is also reliable. The graph visualization provides another interpretability layer which enables identifying relations between parties and requirements in a visual manner by a legal expert and automated systems. These results prove that the integration of NLP and graphical methods can be a powerful, scalable, and viable way of extracting semantic legal metadata in a variety of domains. Figure 4 represents the model evaluation (before and after) that shows the

comparison of these metrics of civil aviation and criminal law documents.

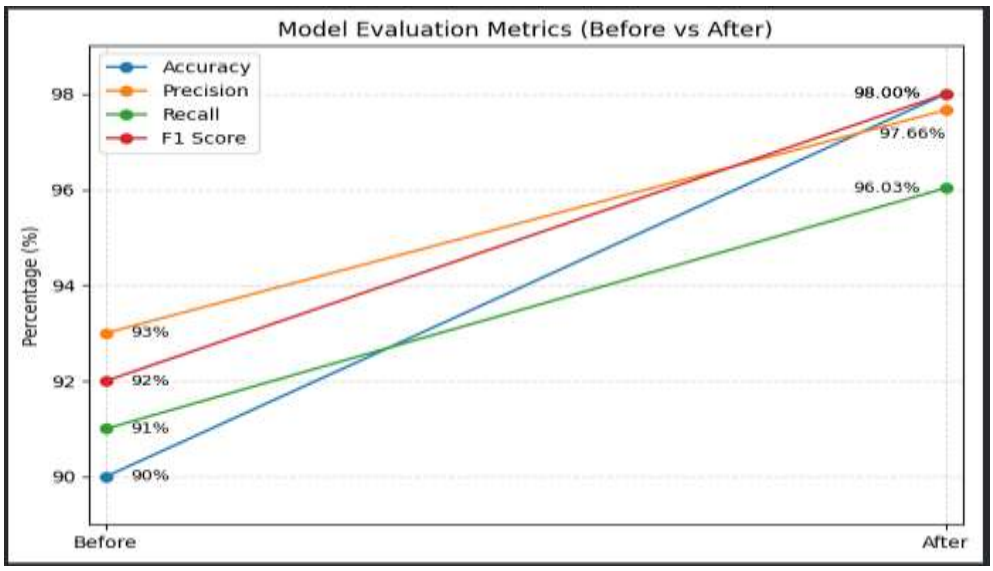

Figure 4: Model Evaluation (Before vs After)

V. CONCLUSION

The paper is about a completely automated semantic legal metadata extraction framework in civil aviation and criminal law documents. Combining NLP methods with graph-like representations, the system correctly recognizes the entities, actions, conditions, and sanctions. Structured tables offer up an interpretable overview of extracted metadata in a clear and readable form, whereas graphs describe the relationship between entities and obligations. Precision, recall, F1-score, and confusion matrices analysis show good performance and reliability. The framework minimizes the use of manual annotation, features nested clause processing and can process large legal datasets in a scaled manner. According to comparative analysis, this approach is better than the traditional rule-based methods, and it can provide a practical solution to the legal document analysis. Further research will be dedicated to the expansion of the dataset, addition of texts that are multilingual, and the use of more sophisticated transformer-based architectures to understand contexts better. In general, the research provides a research, compliance, and intelligent legal information system scalable, interpretable, and high-accuracy methodology of automated semantic legal metadata extraction.

REFERENCES

[1] Liu, R., L. Gao, D. An, Z. Jiang, and Z. Tang, "Automatic document metadata extraction based on deep networks", 2018. In: Huang, X., J. Jiang, D. Zhao, Y. Feng, and Y. Hong. (eds) Natural Language Processing and Chinese Computing. Lecture Notes in Computer Science, vol. 10619. Springer, Cham, https://doi.org/10.1007/978-3-319-73618-1_26

[2] Boukhers, Z., and A. Bouabdallah, "Vision and natural language for metadata extraction from scientific PDF documents: A multimodal approach", in Proc. of the 22nd ACM/IEEE Joint Conf. on Digital Libraries, Association for Computing Machinery, New York, NY, USA, Article 6, June 20-24, 2022, pp. 1-5, https://doi.org/10.1145/3529372.3533295

[3] Al-Zaidy, R. A., and C. L. Giles, "Extracting semantic relations for scholarly knowledge base construction", in Proc. of the IEEE 12th Int. Conf. on Semantic Computing, Laguna Hills, CA, USA, January 31-February 02, 2018, pp. 56-63, https://doi.org/10.1109/ICSC.2018.00017

[4] Bakar, A., I. Safder, and S.-U. Hassan, "Mining algorithmic complexity in full-text scholarly documents", in ICADL Poster Proc., 2018. Hamilton, New Zealand: The University of Waikato, https://doi.org/10.15663/ICADL.2018.66

[5] Altınel, B., and M. C. Ganiz, "Semantic text classification: A survey of past and recent advances", 2018, Information Processing & Management, vol. 54, no. 6, pp. 1129-1153, https://doi.org/10.1016/j.ipm.2018.08.001

[6] Arshad, N., A. Bakar, S. H. Soroya, I. Safder, S. Haider, S. -U. Hassan, N. R. Aljohani, S. Alelyani, and R. Nawaz, "Extracting scientific trends by mining topics from Call for Paper", 2019, Library Hi Tech, vol. 40, no. 1, pp. 115-132, https://doi.org/10.1108/LHT-02-2019-0048

[7] Rodrigo, G. P., M. Henderson, G. H. Weber, C. Ophus, K. Antypas, and L. Ramakrishnan, "ScienceSearch: Enabling search through automatic metadata generation", in Proc. of IEEE 14th Int. Conf. on e-Science (e-Science), Amsterdam, Netherlands, October 29-November 1, 2018, pp. 93-104, https://doi.org/10.1109/eScience.2018.00025

[8] Boukhers, Z., N. Beili, T. Hartmann, P. Goswami, and M. A. Zafar, "MexPub: Deep transfer learning for metadata extraction from German Publications", in Proc. ACM/IEEE Joint Conf. on Digital Libraries, Champaign IL, USA, September 27-30, 2021, pp. 250-253, https://doi.org/10.1109/JCDL52503.2021.00076

[9] Meng, B., L. Hou, E. Yang, and J. Li, "Metadata extraction for scientific papers", 2018, In: Sun, M., T. Liu, X. Wang, Z. Liu, and Y. Liu, (eds) Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Lecture Notes in Computer Science, vol. 11221, pp.111-112, Springer, Cham, https://doi.org/10.1007/978-3-030-01716-3_10

[10] Sleimi, A., N. Sannier, M. Sabetzadeh, L. Briand, and J. Dann, "Automated extraction of semantic legal metadata using natural language processing", in Proc. IEEE 26th Int. Requirements Engineering Conf., Banff, AB, Canada, August 20-24, 2018, pp. 124-135, https://doi.org/10.1109/RE.2018.00022

[11] Sleimi, A., N. Sannier, M. Sabetzadeh, L. Briand, M. Ceci, and J. Dann, "An automated framework for the extraction of semantic legal metadata from legal texts", 2021, Empirical Software Engineering, vol. 26, p. 43, https://doi.org/10.1007/s10664-020-09933-5

[12] Safder, I., S. -U. Hassan, A. Visvizi, T. Noraset, R. Nawaz, and S. Tuarob, "Deep learning-based extraction of algorithmic metadata in full-text scholarly documents", 2020, Information Processing & Management, vol. 57, no. 6, p. 102269, https://doi.org/10.1016/j.ipm.2020.102269

[13] Blanchy, G., L. Albrecht, J. Koestel, and S. Garré, "Potential of natural language processing for metadata extraction from environmental scientific publications", 2023, Soil, vol. 9, no. 1, pp. 155-168, https://doi.org/10.5194/soil-9-155-2023

[14] Skondras, P., N. Zotos, D. Lagios, P. Zervas, K. C. Giotopoulos, and G. Tzimas, "Deep learning approaches for big data-driven metadata extraction in online job postings", 2023, Information, vol. 14, no. 11, p. 585, https://doi.org/10.3390/info14110585

[15] Breit, A., L. Waltersdorfer, F. J. Ekaputra, S. Karampatakis, T. Miksa, and G. Käfer, "Combining semantic web and machine learning for auditable legal key element extraction", In The Semantic Web: 20th Int. Conf., ESWC, 2023, Hersonissos, Crete, Greece, May 28-June 1, 2023, pp. 609–624, https://doi.org/10.1007/978-3-031-33455-9_36