Towards Developing an Automated Chatbot for Predicting Legal Case Outcomes: A Deep Learning Approach

Shafiq Alam[1](✉), Rohit Pande[2], Muhammad Sohaib Ayub[3], and Muhammad Asad Khan[4]

[1] School of Management, Massey University, Auckland, New Zealand
salam1@massey.ac.nz

[2] Department of Computer Science, Lahore University of Management Sciences, Lahore, Pakistan
15030039@lums.edu.pk

[3] Duco Consultancy Limited, Gurgaon, India

[4] Department of Telecommunication, Hazara University, Mansehra, Pakistan
asadkhan@hu.edu.pk

Abstract. The accurate prediction of legal case outcomes is crucial for effective legal advocacy, which relies on a deep understanding of past cases. Our research aims to develop an automated chatbot for predicting the outcomes of employment-related legal cases using deep learning techniques. We compare and significantly improve on mining the New Zealand Employment Relations Authority (NZERA) dataset, using various deep learning models such as Latent Dirichlet Allocation (LDA) with different activation functions of Recurrent Neural Network (RNN) to determine their predictive performance. Our study's findings show that SoftSign-based RNN-LDA models have the highest accuracy and consistency in predicting outcomes.

Keywords: Legal advocacy · Predictive models · Semantic analysis · Deep learning

1 Introduction

Legal advocacy relies heavily on predicting the outcomes of new cases, which requires a deep understanding of the details contained within past cases [9]. In fact, one of the main skills in legal advocacy is the ability to study past cases and make informed decisions for predicting the potential outcome of new legal cases based on that knowledge. Therefore, it is crucial to develop knowledge based on the specifics of past cases in order to forecast future cases accurately. With this in mind, an automated system that can learn from past cases and make predictions for future cases could be incredibly valuable for both the general public and legal practitioners [9]. This kind of system has the potential to offer initial assessments of new cases, taking into account the provided circumstances.

N. T. Nguyen et al. (Eds.): ACIIDS 2023, LNAI 13995, pp. 163–174, 2023.
https://doi.org/10.1007/978-981-99-5834-4_13

The aims of this research included; retrieving and processing past employment case documents from the NZERA; conducting a comprehensive semantic analysis of these documents to identify patterns and relationships within the text; using various deep-learning models to predict the outcomes of these cases and comparing and assess various matrices of these models.

To accomplish these objectives, the study extracted employment case documents from an online dataset, conducted feature selection, applied semantic analysis through Latent Dirichlet Allocation (LDA), and developed deep neural network models to predict employment case outcomes. The accuracy of the models was then evaluated and compared.

This paper extends and significantly improves upon the previous work [13] by a significant increase in data, additional matrices, and improved results. This study's contribution lies in its unique approach to analyzing employment case documents, which has not been widely explored in prior research. Moreover, it is the first known effort to combine feature selection, semantic analysis, and deep learning models to predict employment case outcomes, potentially transforming legal research and fostering future developments in the field. The paper comprises a review of prior research, details of the experimentation, results, and discussion, key findings and their implications, and potential future research.

2 Related Work

The prior research has primarily utilized automated analysis of legal text to extract meaning and predict outcomes for generic cases, as shown in Table 1.

The first application of automated analysis of legal text involves identifying the semantics within case documents using unsupervised machine learning, including LDA on legal documents in China [4], and [5] found that it underperformed against Latent Semantic Analysis (LSA) when analyzing Singaporean Supreme Court judgments. [15] used different models for sequence labeling of Lahore High Court judgments. In contrast, [2] used summarization algorithms such as to analyze Indian Supreme Court case judgments and obtained mixed results.

The second application is to predict legal case outcomes through supervised shallow and deep learning methods, which include support vector machines, random forests, decision trees, and gradient-boosted machines [12,16,18]. Efficient data analysis is essential for informed decision-making and extracting meaningful insights from human-sourced data [1]. Recurrent neural networks [20] have also been used to achieve a low F-measure of 0.36, precision of 0.34, and recall of 0.42 for analyzing Chinese civil court cases. Another notable work [17] compared the performance of Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN) on US Supreme Court decisions and found that CNN outperformed RNN with an accuracy of 72.4% to 68.6%, respectively. [7] used various machine learning and deep learning models to predict Brazilian court decisions. Similarly, [11] employed Support Vector Machine (SVM) to analyze European Court of Human Rights cases achieving an accuracy of 65% with their model.

Table 1. Techniques and their purpose in legal NLP.

Techniques	Purpose
LDA [4]	Identifying semantics in Chinese legal documents
LDA and LSA [5]	Identifying semantics in Singaporean Supreme Court judgments
Hidden Markov Models, Maximum Entropy Markov Models, and Conditional Random Fields [15]	Sequence labeling of Lahore High Court judgments to extract topics
CaseSummarizer, LetSum, and GraphicalModel [2]	Analyzing Indian Supreme Court case judgments using summarization algorithms
Support Vector Machines, Random Forests, Decision Trees, and Gradient-Boosted Machines [12,16,18]	Predicting legal case outcomes using shallow learning methods
Recurrent Neural Networks [20]	Predicting legal case outcomes using deep learning methods
RNN and CNN [17]	Analyzing US Supreme Court decisions
SVM [11]	Analyzing European Court of Human Rights cases
RNN [6,9,14]	Predicting legal case outcomes for Chinese Judgement Online (CJO) cases
SVM, CNN, and RNN [22]	Predicting law articles, charges, and terms of penalty in criminal cases from the Supreme Court of China
RNN-based MANN [8]	Predicting law articles, charges, and prison terms in criminal cases from the Supreme Court of China
CNN [21]	Predicting law articles, charges, and terms of penalty in Chinese criminal cases

In the local domain of China [6] employed RNN on China Judgement Online (CJO) cases and achieved an accuracy of 76.3% while [14] utilized RNN to achieve an accuracy rate of 90.01% and [9] developed an RNN-based model named AutoJudge, achieving an accuracy of 82.2% for the same dataset. Furthermore, [22] studied criminal cases from the Supreme Court of China and used SVM, CNN, and RNN to predict different outcomes, stating that CNN had a higher accuracy than SVM, while their RNN models failed to converge. Their best CNN models achieved an accuracy of 84.7%, 83.6%, and 40.0% for predicting law articles, charges, and terms of penalty, respectively. On the other hand, [8] conducted the same experiment using an RNN-based Multichannel Attentive Neural Network (MANN) and achieved improved accuracy of 91.3%, 95.5%, and 69.3% for predicting law articles, charges, and prison terms, respectively. Additionally, [21] achieved high accuracy rates of 97.6%, 97.6%, and 78.2% for predicting law articles, charges, and terms of penalty using CNN.

3 Experimentation

The section describes the methodology used to analyze New Zealand Employment Relations Authority cases. It covers data retrieval, processing, semantic analysis, model training, and evaluation, including techniques such as LDA for semantic analysis and tokenization for secondary data processing.

3.1 Data Retrieval

To begin, we retrieved NZERA case documents from the online dataset of Employment Law [3]. Our dataset contained $12,389$ case documents spanning from January 1st, 2005 to May 22nd, 2022. However, certain cases were excluded due to inconsistencies in their URL naming.

3.2 Data Preprocessing

Our data preprocessing included the following steps:

Paragraph Extraction: Each paragraph in the case documents was found to contain a unique semantic feature, which required identification for feature extraction. However, the PDF format of the raw data did not preserve the structure of the paragraphs during text extraction. To solve this, regular expressions were used to identify and extract each paragraph as an individual feature. Additionally, all numerical data was retained as it contained meaningful information within the cases. The dataset for the study did not have any annotations or metadata, so individual features of each case had to be manually derived from the raw text. The first step was to identify paragraphs or sections representing the document preamble (P) and the case determinations (D) expressed by the presiding authority. However, as most cases were interim court hearings without final determinations, only 30.66% of the documents ($3,230$ cases) could be identified as having the D feature. Feature selection was performed through keyword searches. Two types of data were derived: (1) full documents (FD) including both P and D features and (2) full documents with determinations redacted ($FD - D$) to enable the independent assessment of case circumstances alone

Manual Document Labelling: The metadata of the collected case documents didn't contain the case outcomes, so binary labels for the cases had to be manually added by reading the documents. Cases dismissed by the authority were counted as losses for the applicant. Out of $3,230$ documents, 260 cases (130 victories and 130 losses) were chosen for classification with an attempt to balance the number of cases between the two labels to achieve fairness in classification [10].

3.3 Semantic Analysis

We analyzed the semantics of $12,311$ case documents using LDA, an unsupervised topic-detection method. The analysis was conducted by testing LDA with different numbers of topics, a maximum iteration of 5, and a learning offset of 50. A text feature extraction method based on term frequency was used to identify distinct features and a variety of top words were selected from each topic to create 10 topic-clusters, each containing several words.

3.4 Secondary Data Processing

The data was processed further after the identification of topic-clusters through semantic analysis. The processing involved measuring document similarity using cosine similarity, reducing document size by only keeping words related to the assigned topic-cluster, tokenizing the data, and converting it into 128-dimensional word embeddings to prepare it for model training.

3.5 Model Training

After conducting semantic analysis to identify topic-clusters, the data was further processed and utilized for supervised learning using three deep neural network models. Each model was tested on both FD and $FD - D$ features. The model employed the Gated Recurrent Unit (GRU) variant of the Recurrent Neural Network (RNN) on a TensorFlow platform and was trained for 26 epochs. The model was tested using two different activation functions: Sigmoid and SoftSign. Additionally, RNN served as the base model to assess the performance of various Latent Dirichlet Allocation (LDA) parameters, such as K, n, and d, as discussed in Sect. 3.3.

3.6 Model Evaluation

We evaluated LDA-RNN models in the single run and cross-validated the models with 10-folds to assess their accuracy, precision, recall, and F1-score.

4 Results and Discussion

The section describes an experiment with two stages: semantic analysis and predictive analysis. In the semantic analysis stage, a recurrent neural network (RNN) was used to classify topics based on Latent Dirichlet Allocation (LDA) models with varying complexity. The best-performing model had five topics, $5,000$ features, and the top 300 words. In the predictive analysis stage, the LDA model was used to train two variations of RNN with different configurations of the number of MultiRNN cells and the inclusion of determinations. The best-performing RNN model had SoftSign activation function, FD-D features, and three cells for single-run accuracy, while the best 10-fold cross-validation accuracy was achieved using RNN with SoftSign activation function, FD-D features, and one cell, which had the highest precision and recall.

4.1 Semantic Analysis

Semantics within the text are analyzed in the first stage using RNN with topic-clusters created by LDA comprising various features, top words, and topics.

Various Number of Features ***(d):*** In the initial experiments, the accuracy of the model was assessed using RNN with topic-clusters created by LDA comprising d features. These LDA models had 5 clusters, each limited to the top 300 words. The results presented in Table 2 revealed that when the model identified $5,000$ features within the NZERA data, the cross-validation accuracy was significantly higher at 0.7299 compared to when it identified $2,000$ features. This indicates that the model performed better with a higher number of features.

Table 2. Accuracy of LDA-RNN with various numbers of features (d), top words (n) and LDA topics (k).

Parameter	Single Run	Cross Validation
d = 2,000	0.5814	0.6652
d = 5,000	**0.8095**	**0.7299**
n = 50	0.7442	0.7203
n = 100	0.6512	0.6656
n = 200	0.7442	**0.7900**
n = 300	**0.8095**	0.7299
n = 400	0.7209	0.6470
k = 4	0.6977	0.7110
k = 5	**0.8095**	**0.7299**
k = 6	0.6047	0.6154

Various Top Words ***(n):*** The next step of the experiments involved testing the performance of LDA-created topic-clusters that were restricted to different values of n, representing the number of top words in each cluster. These models consisted of 5 clusters created from a corpus of $5,000$ features. According to the results presented in Table 2, the top-performing models had n values of either 200 or 300. While the former yielded the highest cross-validation accuracy of 0.79, the latter produced the best single-run accuracy. These findings suggest that LDA models limited to the top 200 or 300 words in each topic-cluster were more accurate than those with higher or lower n values.

Various LDA Topics ***(K):*** We also evaluated LDA models with different values of K, which represents the number of topics, using 5,000 features limited to the top 300 words in each topic-cluster. The results presented in Table 2 indicate that the LDA model with $K = 5$ provided the highest single-run accuracy (0.8095) and cross-validation accuracy (0.7299). This implies that using five topics in LDA models produced better results. Based on the results, which showed that LDA-created topic-clusters with 5 topics, $5,000$ features, and limited to the top 300 words provided favorable outcomes with NZERA data, the subsequent experiments were carried out using LDA models with this configuration.

4.2 Predictive Analysis

In the subsequent phase, the LDA model with the aforementioned parameters was employed to train classification models using two variations of RNN. The results of these experiments with various configurations of the number of MultiRNN cells and inclusion of determinations are provided in Table 3.

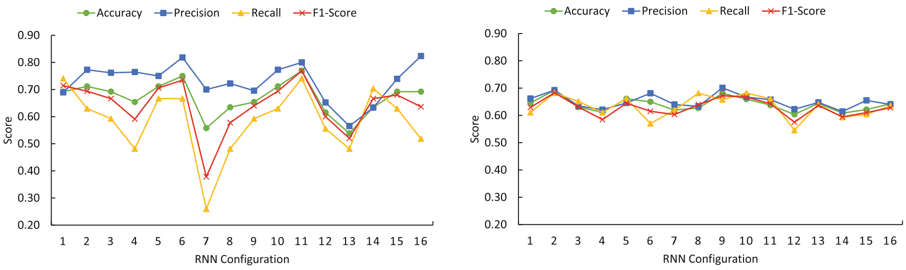
Table 3. RNN Model Performance Using SoftSign and Signmoid Activation Function.

Text Features	Activation Function	RNN Cells	Single Run				10-fold Cross-Validation			
			Accuracy	Precision	Recall	F1-Score	Accuracy	Precision	Recall	F1-Score
FD	SoftSign	1	0.6923	0.6897	**0.7407**	0.7143	0.6462	0.6618	0.6109	0.6295
		2	0.7115	0.7727	0.6296	0.6939	**0.6923**	0.6928	0.6822	**0.6834**
		3	0.6923	0.7619	0.5926	0.6667	0.6308	0.6346	0.6514	0.6310
		4	0.6538	0.7647	0.4815	0.5909	0.6115	0.6212	0.6109	0.5847
	Sigmoid	1	0.7115	0.7500	0.6667	0.7059	0.6608	0.6456	0.6612	0.6450
		2	0.7500	0.8182	0.6667	0.7347	0.6500	0.6812	0.5706	0.6151
		3	0.5577	0.7000	0.2593	0.3784	0.6200	0.6401	0.6192	0.6031
		4	0.6346	0.7222	0.4815	0.5778	0.6269	0.6332	0.6817	0.6392
$FD - D$	SoftSign	1	0.6538	0.6957	0.5926	0.6400	0.6808	**0.7010**	0.6579	0.6722
		2	0.7115	0.7727	0.6296	0.6939	0.6599	0.6670	**0.6824**	0.6672
		3	**0.7692**	0.8000	**0.7407**	**0.7692**	0.6376	0.6578	0.6607	0.6426
		4	0.6154	0.6522	0.5556	0.6000	0.6038	0.6225	0.5457	0.5752
	Sigmoid	1	0.5385	0.5652	0.4815	0.5200	0.6442	0.6478	0.6409	0.6366
		2	0.6346	0.6333	0.7037	0.6667	0.6084	0.6143	0.5931	0.5948
		3	0.6923	0.7391	0.6296	0.6800	0.6215	0.6553	0.6042	0.6104
		4	0.6923	**0.8235**	0.5185	0.6364	0.6423	0.6401	0.6356	0.6281

Analysis of *FD*. The study compared the performance of recurrent neural network models with various configurations in predicting the outcomes of NZERA cases based on the evidence within the case circumstances. Sigmoid activation function with 2 MultiRNN cells demonstrated the best performance with 75% accuracy and 82% precision, whereas RNN-Softsign shows the best performance with 2 cells in 10-fold cross-validation. The 10-fold cross-validation results show that Softsign is most effective in realistically predicting the outcome of NZERA cases.

Analysis of *FD-D*. The performance results show that the RNN-SoftSign based configurations without the evidence within the case circumstances ($FD - D$) were most accurate and consistent in predicting the outcomes of cases. These results demonstrate the potential for using LDA in combination with deep neural networks to predict case outcomes, even before they are officially determined, based solely on the case circumstances.

Analysis of Single Run Vs 10-Fold Cross Validation. Figure 1 shows the performance matrices for the single run and 10-fold cross-validation using various RNN configurations.



(a) Single Run

(b) Cross Validation

Fig. 1. Single run and cross-validation performance matrices using various RNN configurations.

Figure 2 shows the comparison of single run vs cross-validation performance, showing that the single run has slightly higher values for all the metrics compared to cross-validation. This is expected, as the single run uses all the data for training and testing, while cross-validation uses only a subset of the data for testing and the rest for training. Therefore, the single run is more likely to overfit the data, resulting in higher metrics values. Cross-validation, on the other hand, provides a more realistic estimate of the performance.

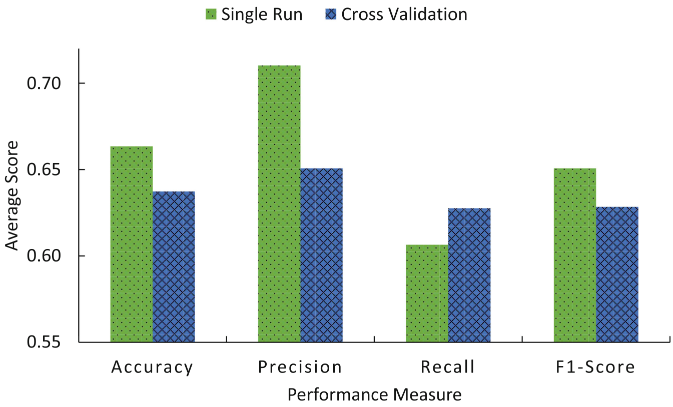


Fig. 2. Comparison of single run and 10-fold cross-validation using an average of the performance matrices.

4.3 CNN with LDA

In this stage of the study, we tested the performance of CNN with both FD and FD - D. The CNN model used a batch size of 64 and a dropout rate of 0.5. These experiments used 5 LDA clusters which were created using 5,000 features and 300 top words. A plot showing the accuracy of this model is provided in Fig. 3. The results are provided in Table 4.

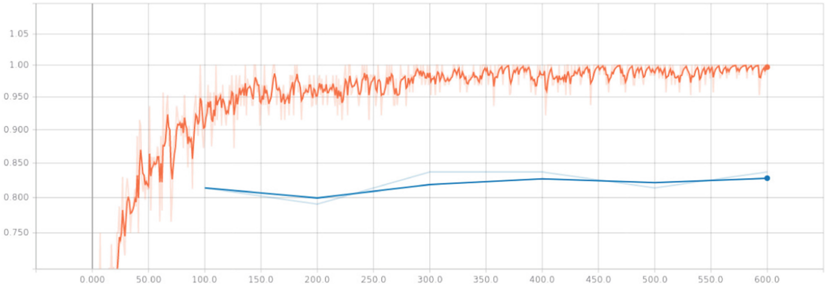


Fig. 3. Train-Test Accuracies with CNN. It shows the accuracy of training data in orange, and test data, in blue. (Color figure online)

Table 4. Accuracy of CNN with LDA.

Features	Single Run	Cross-validation
Full Documents (FD)	**0.8372**	**0.7526**
Full Documents with Redacted Determinations (FD - D)	0.7674	0.7206

The results show that the model analyses FD with consistently higher accuracy than it does for FD - D. Although the cross-validation accuracy of FD-D at 0.7206% can be considered reasonably high, this model may not perform very well when predicting outcomes based on case circumstances alone.

4.4 CAPSULES with LDA

In this section, we have discussed the performance of Capsules with FD and FD-D. The LDA implementation had 5 clusters created with 5,000 features and 300 top words. These experiments used multidimensional GloVe embeddings [19] instead of word2vec. The performance of 50-dimensional, 128-dimensional, and 300-dimensional embeddings was tested. A sample plot of the accuracy is provided in Fig. 4 and the results in Table 5.

The results show that when analyzing FD, 300-dimensional embeddings provided the highest cross-validation accuracy of 0.7758. Meanwhile, 50-dimensional embeddings provided the highest accuracy of 0.8140 with the analysis of FD. Both 50-dimensional and 128-dimensional embeddings performed equally well with Capsules when predicting outcomes from case circumstances alone.

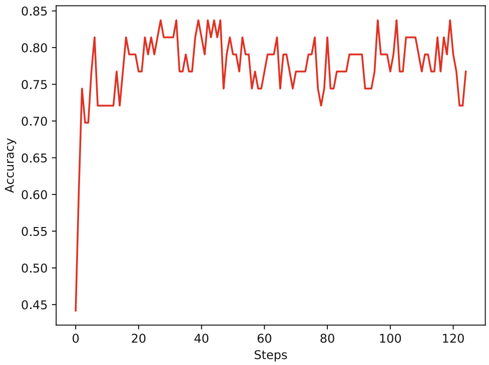


Fig. 4. Accuracy in capsules with 300-dimensional GloVe with redacted determinations.

Table 5. Accuracy of Capsules with LDA.

Embedding Dimensions	Features	Single Run	Cross Validation
50	FD	**0.8140**	0.7571
	FD - D	0.7674	0.7615
128	FD	0.6512	0.7617
	FD - D	0.7674	0.7519
300	FD	0.7907	**0.7758**
	FD - D	0.7476	0.7429

5 Conclusions and Recommendations

The main goal of our research is to develop an automated chatbot that predicts the outcomes of legal cases related to employment relationships, analyze the semantics of legal case documents of the Employment Relations Authority of New Zealand, and compares the performance of various deep learning models in predicting these outcomes.

We retrieve the original data and preprocess it by labeling the extracted text features. We then conduct a comprehensive semantic analysis of the data using LDA to identify patterns and relationships within the text. After that, we implement multiple deep-learning models to forecast the results of these cases. Our research concludes that LDA models with 5 topics and 5,000 features, restricted to the top 300 words, exhibit exceptional performance.

This research contributes a novel approach to the analysis of employment case documents by combining feature selection, semantic analysis, and deep learning models. The results show that LDA models based on RNN-SoftSign demonstrated superior accuracy and consistency and were proficient in making accurate predictions solely based on case circumstances. These findings hold promise for the use of automated chatbots for legal advice and preliminary assessments.

However, to enhance the performance of the model, additional research is necessary, such as testing alternative algorithms and adjusting LDA hyperparameters.

Acknowledgment. The authors acknowledge Citizen AI and Massey University, New Zealand, for providing support and funding the project.

References

1. Ali, S., Ahmad, M., Hassan, U.U., Khan, M.A., Alam, S., Khan, I.: Efficient data analytics on augmented similarity triplets. In: International Conference on Big Data, pp. 5871–5880. IEEE (2022)
2. Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K., Ghosh, S.: A comparative study of summarization algorithms applied to legal case judgments. In: Azzopardi, L., Stein, B., Fuhr, N., Mayr, P., Hauff, C., Hiemstra, D. (eds.) ECIR 2019. LNCS, vol. 11437, pp. 413–428. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-15712-8_27
3. Employment New Zealand: Employment Law Database (2018). https://www.employment.govt.nz/elaw-search
4. Hao, Z., Wei, X., Hu, H.: A comparative method of legal documents based on LDA. In: Abawajy, J., Choo, K.-K.R., Islam, R., Xu, Z., Atiquzzaman, M. (eds.) ATCI 2018. AISC, vol. 842, pp. 271–280. Springer, Cham (2019). https://doi.org/10.1007/978-3-319-98776-7_29
5. Howe, J.S.T., Khang, L.H., Chai, I.E.: Legal area classification: a comparative study of text classifiers on singapore supreme court judgments. CoRR abs/1904.06470 (2019)
6. Jiang, X., Ye, H., Luo, Z., Chao, W., Ma, W.: Interpretable rationale augmented charge prediction system. In: International Conference on Computational Linguistics: System Demonstrations, pp. 146–151 (2018)
7. Lage-Freitas, A., Allende-Cid, H., Santana, O., Oliveira-Lage, L.: Predicting Brazilian court decisions. PeerJ Comput. Sci. **8**, e904 (2022)
8. Li, S., Zhang, H., Ye, L., Guo, X., Fang, B.: MANN: a multichannel attentive neural network for legal judgment prediction. IEEE Access **7**, 151144–151155 (2019)
9. Long, S., Tu, C., Liu, Z., Sun, M.: Automatic judgment prediction via legal reading comprehension. In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 558–572. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_45
10. Mansoor, H., Ali, S., Alam, S., Khan, M.A., Hassan, U.U., Khan, I.: Impact of missing data imputation on the fairness and accuracy of graph node classifiers. In: International Conference on Big Data, pp. 5988–5997. IEEE (2022)
11. Masha, M., Michel, V., Martijn, W.: Using machine learning to predict decisions of the European Court of Human Rights. Artif. Intell. Law **28**(2), 237–266 (2020)
12. Nay, J.J.: Predicting and understanding law-making with word vectors and an ensemble model. PLoS ONE **12**(5), e0176999 (2017)
13. Pande, R., Alam, S.: Predicting the outcome of judicial cases using semantic analysis. In: Symposium Series on Computational Intelligence (SSCI), pp. 1757–1761. IEEE (2020)
14. Shang, L., et al.: Prison term prediction on criminal case description with deep learning. Comput. Mater. Continua **62**(3), 1217–1231 (2020)

15. Sharafat, S., Nasar, Z., Jaffry, S.W.: Legal data mining from civil judgments. In: Bajwa, I.S., Kamareddine, F., Costa, A. (eds.) INTAP 2018. CCIS, vol. 932, pp. 426–436. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-6052-7_37
16. Sulea, O.M., Zampieri, M., Vela, M., van Genabith, J.: Predicting the law area and decisions of french supreme court cases. CoRR abs/1708.01681 (2017)
17. Undavia, S., Meyers, A., Ortega, J.E.: A comparative study of classifying legal documents with neural networks. In: Federated Conference on Computer Science and Information Systems, pp. 511–518. IEEE (2018)
18. Virtucio, M.B.L., Aborot, J.A., Abonita, J.K.C., et al.: Predicting decisions of the philippine supreme court using natural language processing and machine learning. In: Computer Software and Applications Conference (COMPSAC), vol. 2, pp. 130–135. IEEE (2018)
19. Wang, Y., Sun, A., Han, J., Liu, Y., Zhu, X.: Sentiment analysis by capsules. In: World Wide Web Conference, pp. 1165–1174. International World Wide Web Conferences Steering Committee (2018)
20. Xi, R., Zhenxing, K.: Hierarchical RNN for information extraction from lawsuit documents. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1 (2018)
21. Xiao, C., et al.: CAIL 2018: a large-scale legal dataset for judgment prediction. arXiv preprint arXiv:1807.02478 (2018)
22. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning. In: Conference on Empirical Methods in Natural Language Processing, pp. 3540–3549 (2018)