CLASSIFICATION OF LEGAL CONTRACTS BASED ON DEONTIC MODALITIES USING NLP AND BERT

Lalita Neeharika Vajjhala[1], Medam Subrahmanya Sresti[1], Narne Keerthisree Sai[1], H.Anila Glory[2], V.S.Shankar Sriram[3]

[1] Student ,School of Computing, SASTRA Deemed University , Thanjavur
[2,3]Center for Information Super Highway, School of Computing , SASTRA Deemed University , Thanjavur

Abstract

The task of reviewing legal contracts is a very time-consuming job for lawyers and is also a very expensive service for clients who wish to avail of the same. To help this situation and introduce Artificial Intelligence to the world of legal contracts and streamline the process, this research proposes the use of NLP techniques to classify the legal contract sentences based on their deontic modalities, which are very critical in understanding the legal contractual implications. This research aims to find out whether the legal contract reviewing system can be semi-automated, to reduce the time, effort and cost involved in the classification and identification process. The approach incorporates Neural Network (NN) and non-Neural Network models, utilising vector embeddings and BERT models to generate dense representations. The research has achieved an accuracy of 94% proving its effectiveness in identifying and classifying deontic modalities in legal contracts.

Keywords: Natural Language Processing, Neural Networks, Deep Learning, BERT, Deontic Reasoning, Legal Text Classification

Introduction

Statistically, a lawyer takes 7-9 hours on average to analyse a legal contract, which is 50% of the total time spent by a legal professional on reviewing documents. Their clients are charged INR 20,000/- to INR 30,000/- on average per hour. This has made the contract review process inaccessible for clients from various economic backgrounds. This research aims to semi-automate this process with the help of Natural Language Processing (NLP) techniques [1] to aid the lawyers in quickening the process, saving their time. This will also help make the contract review process more economical and easily accessible for people who wish to avail of this service. While reviewing contracts without the help of a legal professional is still possible for a lower-income client, the convoluted terms used in legal contracts make it a difficult task [2].

Legal contracts are written in complex language, consisting of jargons and semantics. It has a varied style of document structure making it difficult to use NLP tools in the legal domain.[3] This research aims to use Neural Network and Non-Neural Network models to categories legal contracts based on deontic reasoning. The study further expands the scope of AI by utilizing pre-trained embeddings using BERT, which revolutionized the usage of NLP tools [4].

The part of a sentence that denotes an expression is called "Modality". The modalities that emphasise rules are classified as "Deontic Modalities". Most legal contracts revolve around the grounds of Deontic modalities [5]. These are further categorised as "Obligation", "Permission" and "Prohibition". For example, a clause in a contract that states "My land will belong to my brother after my death" contains a deontic modality - "obligation" that denotes what is expected to happen to the ownership of the piece of land after the passing of the current owner. The identification of the deontic modality is hence a critical segment of the contract review process that aids legal professionals in understanding the implications of the contents of the contract, in this case, ensuring the owner's brother is given the land's authority.

NLP methods like text classification offer a streamlined approach to contract review, facilitating the identification of contractual norms. Text classification involves assigning labels to text based on content, making it highly efficient for categorizing norms within contracts. Integrating non-neural network models with NLP techniques reduces the time and costs associated with contract review, thereby ensuring affordable access to justice for individuals from diverse backgrounds.

However, the complex language used in legal contracts challenges the NLP tools. The jargons, semantics, structure and style used in legal language are intricate, which makes it very difficult to use these techniques effectively. Ensuring the intactness of the meaning of the text while transforming it into a numerical format needed for computational methods is very crucial. Despite these hurdles, an effective contract review system has been designed that uses NLP, NN, non-NN and BERT models to semi-automate the process [5].

The methodology of this work involves using pre-annotated datasets and word embeddings by pre-processing them to make them suitable for training the various non-NN (Logistic Regression (LR) [6], Support Vector Machine (SVM) [7], Naive Bayes (NB) [8]) and NN models (Convolutional Neural Network (CNN) [9] and Long Short-Term Memory (LSTM)). The research also made use of BERT models that create rich dense vector embeddings which are more efficient in analysing textual data.

This paper consists of the following sections: Chapter 1 introduces the problem statement; Chapter 2 delves into the background work and the theory of deontic reasoning; Chapter 3 discusses the work done in this domain to date; Chapter 4 explains the working methodology opted in the research; Chapter 5 details the results, limitations and scope for future work; Chapter 6 concludes the study.

Related Works

Various researchers have leveraged cutting-edge techniques and models in contract analysis to tackle different aspects of the contract review process. Federico Ruggeri used memory-augmented neural networks to detect discrimination in consumer contracts in 2022 and proposed using BERT models for future study [10]. Gokul Rejithkuma proposed SOTA approaches and used models like T5-large to identify deontic modalities in software engineering contracts in 2023 [11]. Don Tuggener created LEDGAR, a multilabel corpus of legal provisions in contracts that had over 60,000 annotated contracts using ML and NLP techniques in 2020 [12]. Ali Bedii Candaş used ML models to automate reviewing construction contracts in 2022 and prepared a document classification prototype to smoothen the classification process [13]. Vivek Joshi in his study used BERT models to automate classification in software engineering contracts in 2021, which helped achieve a higher F1 score [14]. Jingyun Sun implemented BERT models in his research, measured BERT's performance for the DMC (Deontic Modality Classification) task, and proposed DeonticBERT in 2023 [15]. Abhilasha Sancheti in her research carried out agent-specific multi-label deontic modality classification in 2022 using transfer learning and introduced a corpus of English contracts called LEXDEMOD, annotated with their deontic modalities [16]. Gang Tian proposed using bi-LSTM for a smart contract classification approach in 2020 and introduced an attention mechanism to improve the focus on the most relevant features [17]. Teng Hu used LSTM to perform transaction-based classification of Ethereum smart contracts in 2021 and implemented an approach that can be used for anomaly detection and identifying malicious contracts [18]. Chaochen Shi proposed the use of byte code rather than source code to solve vulnerabilities existing in the automation of contract classification in 2022. These are mainly caused because most smart contracts are not open-source making it difficult to use the NLP tools for classification tasks [19].The significance of NLP techniques for preprocessing and the importance of embeddings for converting text to vectors is crucial in advancing contract analysis. These techniques enable researchers to extract meaningful insights from legal documents, paving the way for more efficient and accurate contract analysis processes.

Methodology

The methodology section of this paper outlines the approach employed to classify legal contracts utilizing both traditional machine learning techniques and neural network models, including BERT in Fig.1,Fig.2,Fig.3.This comprehensive methodology encompasses a diverse range of strategies to achieve accurate classification results, integrating both non-neural network and neural network methodologies to address the complexities inherent in legal document analysis. The proposed methodology for the research involves the following processes and steps:

Datasets

This research has used the annotated datasets made available [20]. The dataset consists of a Norm and Gold Standard Dataset. The Norm dataset consisted of an equal proportion of norm and non-norm sentences made specifically to train the models to identify the norm and non-norm statements from legal contracts. The Gold Standard dataset consisted of real-life proportions of the deontic modalities in legal contracts, for practical application.

Binary-Label Classification

Binary Label Classification was performed using the Norm Dataset, which classifies the input sentence as a norm and non-norm. As a part of the NLP pre-processing in Fig.1, tokenization and lemmatization were performed on the Norm Dataset. Then, the binary label classification was done using non-neural network models (Stochastic Gradient Classifier and Logistic Regression) and neural network models (CNN) in Fig1.TF-IDF (Term Frequency-Inverse Document Frequency) was performed to transform the text data into the numerical format, and the results were measured using accuracy, precision, recall and F1 score.

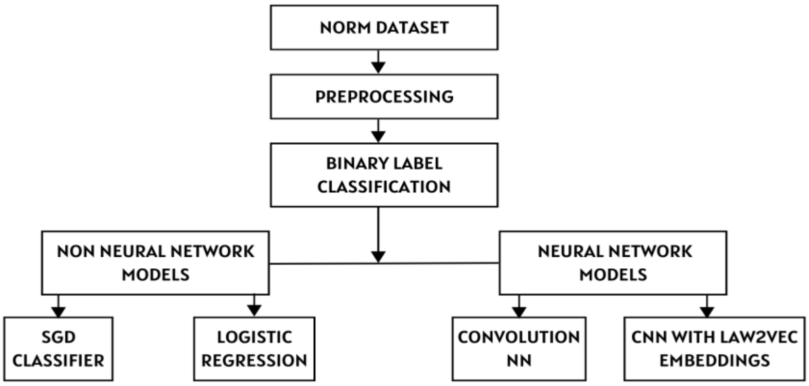


Fig.1:Binary-label classification for the classification of the norm and non-norm sentences

While that was done for the non-neural network models, dense vector embeddings, specifically law2vec file was used for the neural network model [21]. It is a file that contains pre-trained vector embeddings designed specifically for legal text. The research extracts words and their corresponding vectors from the embeddings, which are later filled in an embedding matrix. Since this CNN model worked better than the non-neural network models, cross-validation was performed to concretise the results. This approach reduced the loss value, as compared to the pure CNN model.

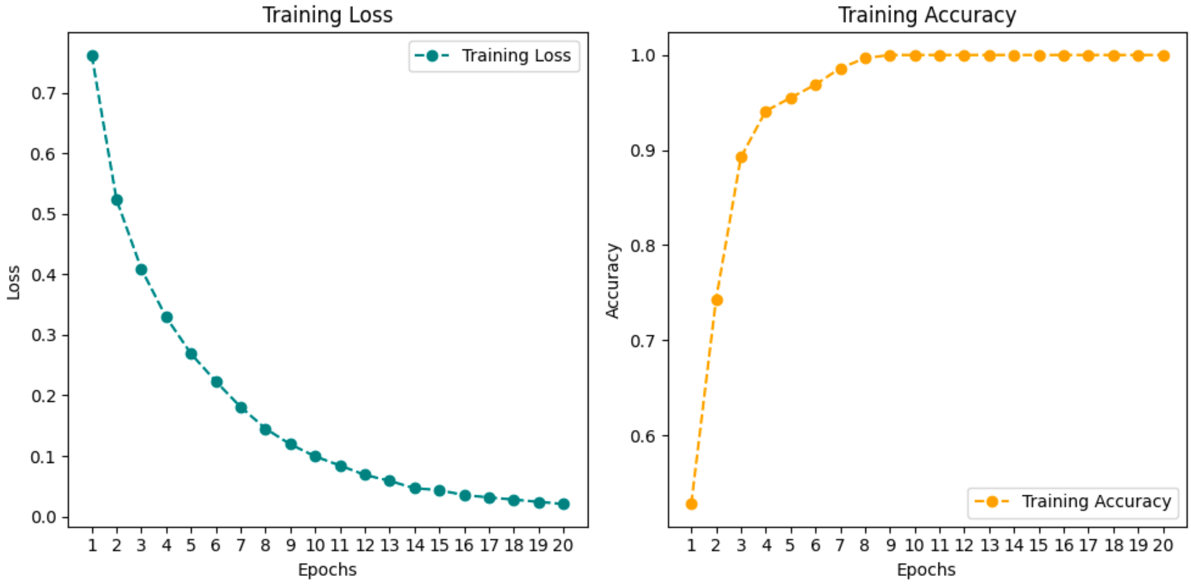


Fig.2: Performance graphs for the CNN model for Training Loss and Training Accuracy

Pre-processing and creation of dataset for Multi-Label Classification

Here, the annotated and reviewed gold standard dataset is pre-processed to create a final dataset that is used to train the various models to classify the input text sentences based on deontic modalities as permission, obligation and prohibition. The process involved dropping the unnamed columns, and the columns that do not contribute to the required text classification. The duplicates are also removed. The next step involved checking the unique values present in the target variable, which turned out to be 10 as given below (Fig.4).

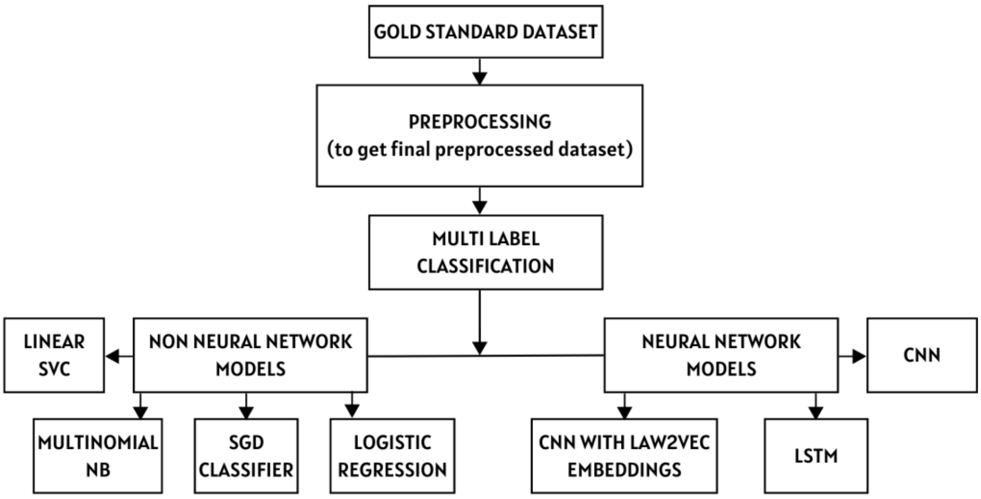
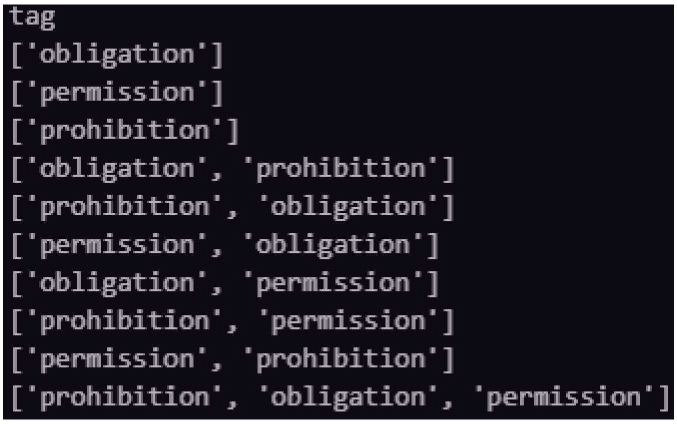


Fig.3: Multi-label classification for the classification of sentences into permission, obligation and prohibition



```
tag
['obligation']
['permission']
['prohibition']
['obligation', 'prohibition']
['prohibition', 'obligation']
['permission', 'obligation']
['obligation', 'permission']
['prohibition', 'permission']
['permission', 'prohibition']
['prohibition', 'obligation', 'permission']
```

Fig 4: The various values present in the target variable

The count plot of the tags showed that the "Obligation" tag was three times more frequently occurring than the other tags in Fig.5.

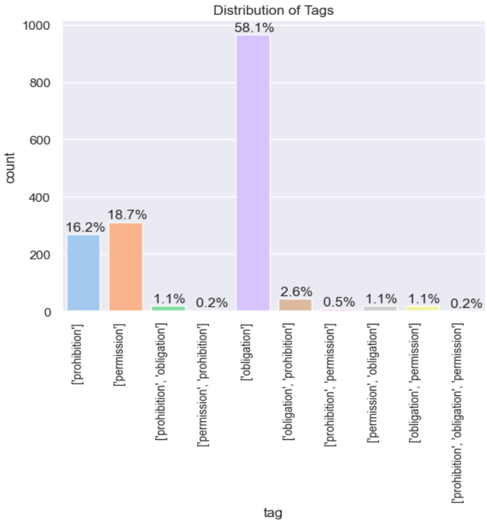


Fig 5: Imbalance in the dataset

To balance out the dataset in Fig.5, it was split into a majority dataset and a minority dataset where the majority dataset consisted of the obligation tag and the minority dataset consisted of the rest of the tags. Down sampling was performed on the majority dataset and concatenated the resulting dataset with the minority dataset to create a resulting dataset that was highly balanced. This ensured that during the training process, there was no opportunity for bias. The count plot in Fig.6 of the balanced dataset showed the following results proving that the resulting dataset is completely balanced.

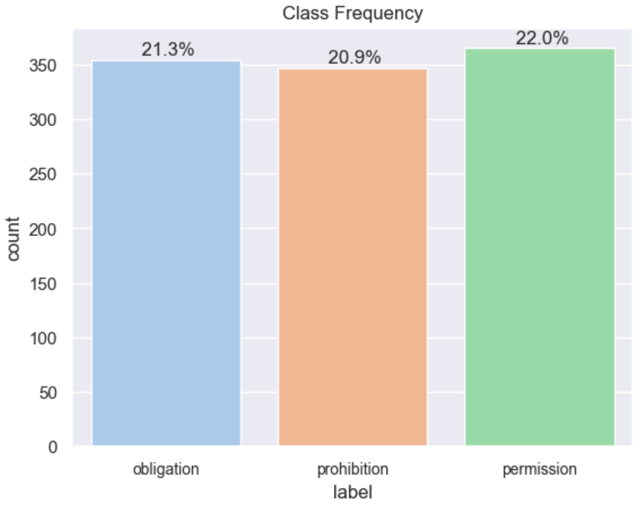


Fig 6: Balanced dataset

NLP pre-processing techniques like lemmatization and TF-IDF Vectorization were used to create the final balanced dataset which was used to train all the multi-label classification models in Fig.3.

Multi-Label Classification

The final pre-processed dataset created earlier was used for training non-NN and NN models to classify the input sentences into the deontic tags Fig.3. Among the preliminary non-NN, Linear SVC, Multinomial Naive Bayes, Logistic Regression and SGD Classifier were put to use. Out of the NN models available, CNN was used. Tokenization, keras pre-processing and transformation of text into feature vectors were a few of the steps in making the input suitable to train the CNN model.

To test the performance of the model on pre-trained word embeddings, the CNN model was run along with the law2vec dense vector embeddings. The law2vec dense vector embeddings are contained in a file which is first loaded into the dataframe. Then, the model iterates through every row of the file and considers the first word to be the actual word, takes the rest of the row contents as an array of the vector dimensions and finally stores them as word-vector pairs in a dictionary. These are populated as word vectors from the embeddings into an embedding matrix which is initially at zero. This model resulted in a higher accuracy than the earlier model.

To further assess the model's performance and reduce any overfitting, cross-validation along with CNN was performed on the final preprocessed dataset. The 5-fold model achieved the highest accuracy among all the NN models.

RNN is a preferred model when considering text classification. Here, the study made use of an RNN model with a single LSTM (Long Short-Term Memory) layer, along with callbacks. However, this model performed poorly as compared to the other NN models, whereas it surpassed the highest accuracy obtained from the non-NN models.

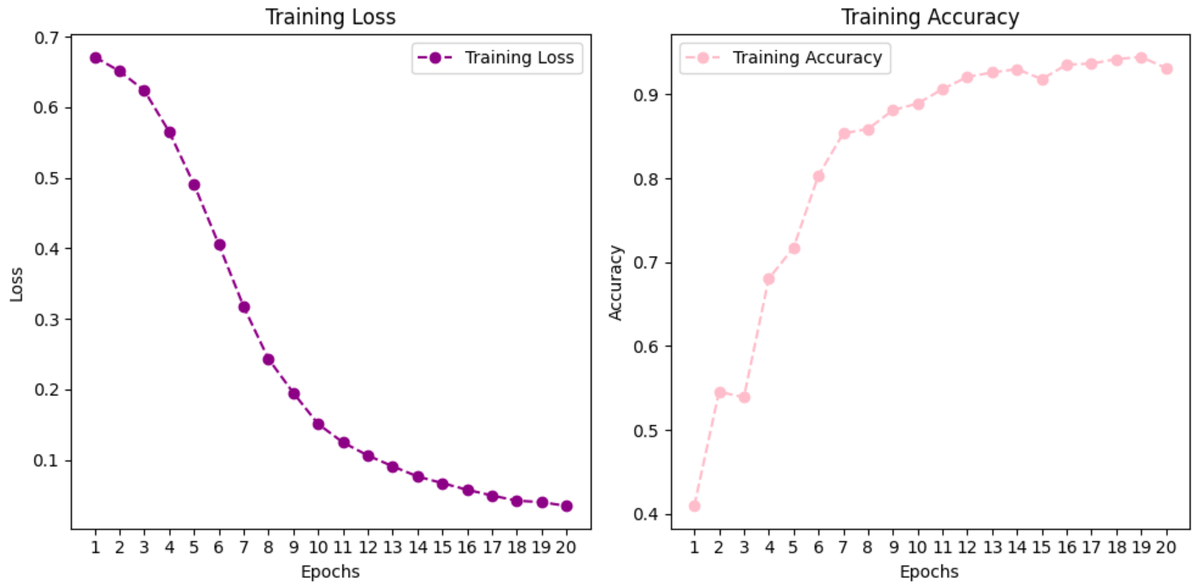


Fig 7: performance of the CNN model and include the plots for Training Loss and Training Accuracy.

Use of Transformers for multi-label classification

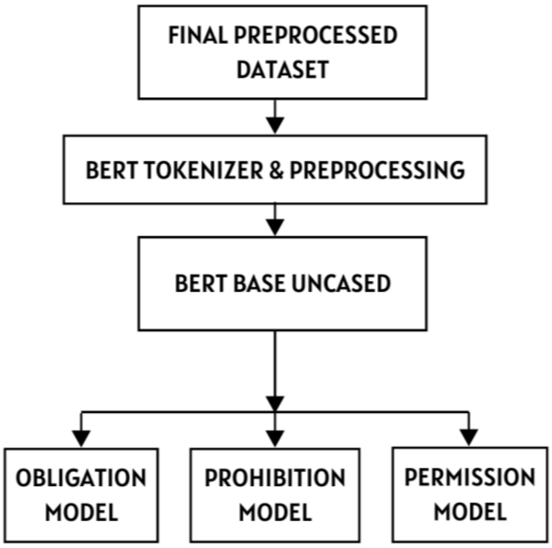


Fig.7: Multi label classification using the BERT models

One step ahead, the research employed state-of-the-art deep learning methodologies to develop a model that performed better than all the other models used so far in the research. The BERT (Bidirectional Encoder Representations from Transformers) model was used in Fig.8, which has a bidirectional architecture that enables a more comprehensive study of the contextual relationships within textual data. This feature of BERT made it a pioneer in contemporary NLP research endeavours.

Methodology

The methodology involved preprocessing using the BERT Tokenizer in Fig.8, ensuring alignment of the dataset for the BERT model. The data is segregated into training, validation and testing subsets through meticulous partitioning to help evaluate the model. After preprocessing the data, the BERT models have been initialized. In this research, three BERT models - BERT-base-uncased, CONTRACTS-BERT-base and LEGAL-BERT-base were used. Then, the BERT tokenizer is used to break the sentences into individual tokens. The tokenized input sentences were formatted into the appropriate input tensors that the BERT model expected, which included multiple processes like creating attention masks to indicate which tokens are actual words and which are padding tokens. While training the model, backpropagation and gradient descent have been used. Backpropagation helps in understanding how each parameter affects the loss, facilitating the optimization process and Gradient descent works on minimising the loss function, iteratively. The BERT model takes the input tensors as the input and predicts the corresponding labels. The loss is calculated using the actual and predicted labels. The loss is backpropagated through the network and the model's weights are updated to minimise the loss. Among all the BERT models used, BERT-base-uncased proved to be better than the CONTRACTS-BERT-base and LEGAL-BERT-base, due to the larger size of the corpus used to train the models. Three instances of the same BERT model were created to identify the three deontic modalities - permission, obligation and prohibition. BERT models capture intricate linguistic nuances, enabling better semantic comprehension and achieving enhanced classification performance, even with the input that makes use of complex language like legal contracts.

Results, limitations and future study

Results

To evaluate the selected models, the research made use of metrics like Accuracy, Precision, Recall, F1 Score and Loss [20] in equ (1-4).

- $Accuracy = TP + TN /(TP + TN + FP + FN)$ (1)
- $Precision = TP(TP + FP)$ (2)
- $Recall = TP(TP + FN)$ (3)
- $F1Score = 2 * Precision * Recall/(Precision + Recall)$ (4)

Accuracy provides the measure of the proportion of the correctly classified sentences among all the sentences of the dataset, indicating how often the model is correct while predicting a classification. Precision measures the proportion of the correctly identified sentences belonging to a particular label. It shows how precise the model is while predicting the sentences of a particular label. On the other hand, Recall represents the proportion of sentences correctly identified by the model, as belonging to a certain label, among all the other sentences that actually belong to that label, giving a measure of the model's ability to identify all relevant instances of a label. The F1 Score is the harmonic mean of Precision and Recall and provides a balance between both. It also helps in cases where there is an imbalance between labels, like the dataset used in this research. The Loss functions are used to study the discrepancy between the predicted labels and the actual labels of the sentences. The lower the loss value, the better the predicting performance of the model. In the binary classification in Table 1, the Linear SGD classifier performed the best, with an accuracy of 90% among all the non-NN models. Among the select NN models, the CNN model that made use of the pre-trained Law2vec embeddings performed the best with an accuracy of 97% and a precision of 97%.

For the multi-label classification in Table 2, both NN and non-NN models were used. Among the non-NN models, Linear SVC outperformed the rest of the models with an accuracy of 75% and better results in all the other metrics as well. Among the NN models, CNN performed the best with an accuracy of 88%, which further improved to 91% upon cross-validation. CNN along with the pre-trained Law2vec embeddings achieved an accuracy of 87%, which was lesser than the CNN model, but better than the LSTM results which were the least, with an accuracy of 76%. This was higher than the non-NN models, but still low, compared to the other NN models. The BERT model achieved an unbeaten average accuracy of 94% and an average precision of 94% in Table 3. This is attributed to the contextualized embeddings generated by BERT, which identifies the different meanings of the same word when used in different contexts. Also, BERT considers both the left and right contexts of each word to capture Bidirectional contextual information. In contrast, the Law2vec embeddings are generated only using the left or right context alone. BERT is also trained on a larger corpus of data as compared to a very specific and limited corpus that is used to train the Law2vec embeddings. The results of all the models used are given below along with the evaluation metrics used. While the non-NN models performed efficiently, with SVC reaching an accuracy of 75% in Table 2, the NN models, particularly the CNN model proved to be the best, at an accuracy of 91%. The dense vector embeddings generated using BERT proved to have the highest accuracy among all the models used in the research, with an average accuracy of 94% in Table 3. These results highlight the potential of AI, NN and BERT models in the semi-automation of the contract review process, aiding legal professionals to decrease the time taken to analyse a legal contract and also decrease the potential cost for those who wish to avail of this service.

Table 1:Binary Classification

Model	Accuracy	Precision	Recall	F1 Score	Loss
SGD CLassifier	0.90	0.91	0.92	0.91	-
Logistic Regression	0.82	0.84	0.86	0.83	-
CNN	0.87	0.86	0.88	0.87	0.29

Table 2: Multiclass classification

Model	Accuracy	Precision	Recall	F1 Score	Loss
SGD Classifier	0.75	0.79	0.81	0.84	-
Logistic Regression	0.68	0.76	0.68	0.79	-
Linear SVC	0.75	0.80	0.80	0.85	-

Multinomial NB	0.65	0.73	0.65	0.76	-
CNN	0.88	0.97	0.91	0.90	0.16
CNN cross-validation	0.91	0.97	0.89	0.89	0.18
CNN with law2vec embeddings	0.87	0.90	0.86	0.88	0.20
LSTM	0.76	0.73	0.60	0.65	0.52

Table 3: BERT for Multiclass classification

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**Loss**
BERT_base_uncased for obligation	0.94	0.93	0.93	0.93	0.28
BERT-base_uncased for permission	0.92	0.95	0.92	0.91	0.11
BERT_base_uncased for prohibition	0.97	0.96	0.92	0.97	0.09
Average for the models	0.94	0.94	0.92	0.92	0.16

Limitations

The main limitation that was faced while carrying out this study was the size of the dataset. While the dataset used was enough to carry out the research and get reasonable and working models, the lack of more instances in the dataset led to less accurate models. Also, the models were trained on datasets that had more occurrences of a single deontic label, which led to poor performance while predicting sentences that had all three labels present. While executing the models, a high amount of Computation Power was required, which is available only in specifically designed systems.

Conclusion

This research has answered the question of whether the legal contract review system can be semi-automated. The research used annotated datasets that contains English-based contracts and performed multiple NLP preprocessing techniques to prepare a final dataset. This dataset was used to train various ML and NN models to classify legal contracts based on their deontic modalities. The research also implemented BERT models which further preprocessed the dataset and created dense vector embeddings. These were then fed into the BERT-base-uncased model, which achieved commendable and concrete results, showcasing the effectiveness of the research in classifying legal contracts based on their deontic modalities. The proposed model has achieved an accuracy of 94% and a precision of 94% in Table 3, which concludes that the research proved its efficiency in classifying contract sentences based on the deontic modalities (permission, prohibition and obligation).

References

[1] Maree M, Al-Qasem R, Tantour B. Transforming legal text interactions: leveraging natural language processing and large language models for legal support in Palestinian cooperatives. International Journal of Information Technology. 2024 Jan;16(1):551-8.

[2] BOWCOTT O (2016) Legal fees investigation reveals huge disparities between law firms. The Guardian, April 5, 2016 [viewed on 06 July 2022]. Available from: https://www.theguardian.com/law/2016/apr/05/legal-fees-nvestigation-reveals-huge-disparities-between-law-firms

[3] Krasadakis P, Sakkopoulos E, Verykios VS. A Survey on Challenges and Advances in Natural Language Processing with a Focus on Legal Informatics and Low-Resource Languages. Electronics. 2024 Feb 4;13(3):648.

[4] Aires JP, Pinheiro D, Lima VS, Meneguzzi F. Norm conflict identification in contracts. Artificial Intelligence and Law. 2017 Dec;25(4):397-428

[5] Licari D, Comandè G. ITALIAN-LEGAL-BERT models for improving natural language processing tasks in the Italian legal domain. Computer Law & Security Review. 2024 Apr 1;52:105908.

[6] Aseervatham S, Antoniadis A, Gaussier É, Burlet M, Denneulin Y. A sparse version of the ridge logistic regression for large-scale text categorization. Pattern Recognition Letters. 2011 Jan 15;32(2):101-6.

[7] Joachims T. Text categorization with support vector machines: Learning with many relevant features. InEuropean conference on machine learning 1998 Apr 21 (pp. 137-142). Berlin, Heidelberg: Springer Berlin Heidelberg.

[8] Rasjid ZE, Setiawan R. Performance comparison and optimization of text document classification using k-NN and naïve bayes classification techniques. Procedia computer science. 2017 Jan 1;116:107-12.

[9] Chen Y. Convolutional neural network for sentence classification (Master's thesis, University of Waterloo).

[10] Boginskaya O. A Corpus-Based Study of Deontic Modality in Legal Discourse. Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje. 2022 Jul 29;48(1):1-26.

[11] Neill JO, Buitelaar P, Robin C, Brien LO. Classifying sentential modality in legal language: a use case in financial regulations, acts and directives. InProceedings of the 16th edition of the International Conference on Articial Intelligence and Law 2017 Jun 12 (pp. 159-168).

[12] Matulewska A. Deontic modality and modals in the language of contracts. Comparative Legilinguistics. 2017 Feb 8;2(1):75-92.

[13] Bulcke PV. Dealing with deontic modality in a termbase: the case of Dutch and Spanish legal language. Linguistica Antverpiensia, New Series–Themes in Translation Studies. 2013 Dec 31;12.

[14] Mahoney CJ, Zhang J, Huber-Fliflet N, Gronvall P, Zhao H. A framework for explainable text classification in legal document review. In2019 IEEE International Conference on Big Data (Big Data) 2019 Dec 9 (pp. 1858-1867). IEEE.

[15] Samuel G. Classification of contracts: A view from a common lawyer. Santos, Francisco Javier Andres/Baldus, Christian/Dedek, Helge (Hg.), Vertragstypen in Europa Historische Entwicklung und europäische Perspektiven, München. 2011:117.

[16] Bondi M, Diani G. Conveying deontic values in English and Italian contracts: A crosscultural analysis. ESP across Cultures. 2010 Sep;7:7-24.

[17] Pasetto L, Cristani M, Olivieri F, Governatori G. Automated Translation of Contract Texts into Defeasible Deontic Logic. Logics for New Generation Artificial Intelligence (LNGAI 2021).:81-92.

[18] Abdelmoneim D. Semantic deontic modeling and text classification for supporting automated environmental compliance checking in construction (Doctoral dissertation, University of Illinois at Urbana-Champaign).

[19] Wan L, Papageorgiou G, Seddon M, Bernardoni M. Long-length legal document classification. arXiv preprint arXiv:1912.06905. 2019 Dec 14

[20] Graham SG, Soltani H, Isiaq O. Natural language processing for legal document review: categorising deontic modalities in contracts. Artificial Intelligence and Law. 2023 Nov 11:1-22.

[21] Chalkidis I, Kampas D. Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law. 2019 Jun 15;27(2):171-98.

[22] Vujović Ž. Classification model evaluation metrics. International Journal of Advanced Computer Science and Applications. 2021;12(6):599-606.