Convolutional Neural Network-based Automatic Prediction of Judgments of the European Court of Human Rights

Arshdeep Kaur and Bojan Božić

School of Computer Science, Technological University Dublin
arshdeep.kaur@myTUDublin.ie, bojan.bozic@TUDublin.ie

Abstract. In the past few years, predictive modelling has brought revolutionary changes in the way various industries function. Advancements in the areas of Deep Learning (DL) and Natural Language Processing (NLP) have made their application to different problem areas highly promising. In the legal domain, positive results have been obtained in predicting the judgements of various Courts of different countries using DL and NLP. However, not much research has been carried out in the area of legal judgement forecasting for the European Court of Human Rights (ECHR). The models designed in the previous research employ only one Machine Learning algorithm namely a Support Vector Machine (SVM) to solve such problem.

This study applies DL and NLP to the problem of automatic prediction of judgements for ECHR. Extensive experiments are conducted which compare the performance of models trained on SVM with linear kernel as part of previous research (Medvedeva, Vols & Wieling, 2018) with the models trained on Convolutional Neural Networks (CNN) as proposed in this study. To implement this, state-of-the-art NLP techniques are applied to the text data. Moreover, pre-trained and custom trained Word Embedding text representations are considered. Statistical tests are performed to gather sufficient statistical evidence to determine which algorithm performs better at providing a solution to this problem. Based on the results obtained, it is established that overall, CNN models outperform SVM models as the former achieves an average accuracy of 82% whereas the latter achieves 75%. Specifically, CNN models for four Articles out of nine achieve statistically significant higher accuracy than SVM models.

Keywords: *Convolutional Neural Network, Support Vector Machine, Natural Language Processing, Word Embedding, European Court of Human Rights*

1 Introduction

Law and Order is one of the integral components of society. It is a well-known fact that in any legal court, case proceedings take a lot of time before a final judicial decision is declared. This leads to piling of cases and long waiting time which is an ordeal for innocent people. Lawlor once surmised that one day, computers would be able to analyze and predict the outcomes of judicial decisions (Lawlor, 1963). Automating the process of forming legal judgments on different court cases is one of the promising areas of application of artificial intelligence that can revolutionize the legal domain. Building an automated system that can accurately predict the outcome of a case will help in reducing the error of judgment that might happen by a human judge. It will allow the judges to focus on more complex tasks by prioritizing the cases. This will also improve the delay caused in handling legal cases and will provide justice to individuals faster. Hence, the efficiency of the judicial system will be improved.

Automatic prediction of ECHR case judgments has been previously carried out by developing systems that employ Machine Learning (ML) algorithms like SVM. Although such systems have achieved a decent prediction accuracy, there is a possibility of building more accurate prediction models by leveraging the power of DL algorithms. Models built using such algorithms are capable of learning the complexities of large text datasets which other ML models may not be able to do. Based on the literature review performed in the next section, CNN models have been observed to give good results for many text classification problems and hence, they seem to be a potential modeling technique for ECHR cases.

In the previous work, N-grams were used as the text representation. Although they can capture some of the context of the text data, they can lead to high-dimensional sparse matrices. Text representation such as Word embeddings can capture the semantics of data well and has been observed to give good results with models, based on literature review. This aspect of NLP has not been explored previously which needs to be evaluated for ECHR problem. If a legal assistant system is built using DL algorithms based on Word Embeddings and succeeds at achieving higher accuracy, it can be useful in assisting judges at ECHR with unforeseen real word legal cases once it is deployed.

This study has implemented automated decision making for ECHR by leveraging the power of Semantic Analysis of text and the ability of CNN to learn local features in text without manually engineering the features. The research methodologies used were a mix of qualitative and quantitative methods where a Sequential Exploratory design was followed. Secondary research has been conducted to collect an existing dataset for ECHR judicial cases, perform a systematic literature review and summarise the findings. Using Constructive form of research, two different models were built, a general ML model and a DL model, which were analytically compared to determine whether the new model outperforms the baseline model. Deductive reasoning has been employed in validating the hypothesis and concluding whether the model built as part of this study was better at predicting the judicial decisions of ECHR cases than the existing model based on statistical test results.

2 Related Work

2.1 Implementation of Artificial Neural Networks (ANNs) for NLP applications in various domains

This section describes the research work carried out in various text classification problems where models trained on DL algorithms outperformed models trained on ML algorithms.

Classification of medical documents at the sentence level into 26 clinical subject categories has been performed using various modeling and data processing techniques. It has been observed that CNN models with Word2Vec achieved 15% higher accuracy compared to Logistic Regression with Sentence Embedding, Mean Word Embedding, and Bag of Words (Hughes, Li, Kotoulas & Suzumura, 2017). A system was designed to predict Colorectal Cancer among patients by capturing temporal nature of medical records dataset obtained from Julius Centre, Netherlands. It was found that Recurrent Neural Network models (RNNs) performed at par with the state-of-the-art ML algorithms namely SVM, Random Forests, Logistic Regression, and Decision Trees and achieved an Area Under the Curve (AUC) of 0.811 (Amirkhan, Hoogendoorn, Noomans & Moons, 2017). All such research work signifies that ANNs like CNN and variants of RNN can give better performance than traditional ML models without any need of manually handcrafting the features for text classification problems in various domains.

2.2 Implementation of Artificial Intelligence(AI) in Legal Industry

The wide spectrum of research carried out in the past years have constantly reiterated the potential of AI in the legal domain. It has been suggested that if an accurate legal expert system is deployed, it can have a profound effect on the legal sector (Bermen & Hafner, 1989).

Research related to the application of AI in the legal domain has been ongoing for the past couple of years. ML algorithms have been designed for predictive modeling for different problems and promising results have been obtained. For French Supreme Court, a model designed using SVM based on unigram and bigram text representation achieved 98% average F1 score in predicting the case judgment (Sulea, Zampieri, Vela, Dinu & Genabith, 2017). A Random Forest model trained on the asylum cases dataset predicted with an accuracy of 80% whether an applicant would be granted asylum by the US Courts (Dunn, Sagun, Sirin & Chen, 2017). ANN architectures like CNN and RNN variants and different NLP techniques have also been employed for various legal problems. For US Supreme Court, ANNs have been designed for classification of documents for determining the outcome of legal cases. It has been observed that CNN models with Word2Vec text representation achieved 72.4% accuracy in classifying the court decision into 15 categories. Various modeling methods were implemented like Latent Dirichlet Allocation with logistic regression, Doc2vec with logistic regression, SVM with bag of words, CNN with Word2Vec, GloVe and fasttext, Long short-term memory (LSTM) with Word2Vec and Gated Recurrent Unit (GRU) with Word2Vec. CNN with Word2vec was found to achieve the best accuracy (Undavia, Meyers & Ortega, 2018). In another research work, DNN using ReLU activation function, momentum optimization and dropout technique trained on data containing 7700 cases outperformed SVM models and achieved 70.4% accuracy in predicting the judgments of US Supreme Court (Sharma, Mittal, Tripathi & Acharya, 2015). All such research work implies that ML and DL algorithms give positive results in solving many problems of legal domain.

2.3 Implementation of ANNs in ECHR

A small body of research has been previously carried out for designing a legal forecasting system for ECHR. The previous work on judicial cases developed an SVM model having a linear kernel and applied on data represented using N-Gram and topics which predicted whether an Article of ECHR Convention (Article 3, 6 and 8 considered) has been violated or not with 79% accuracy (Aletras, Tsarapatsanis, Preotiuc-Pietro & Lampos, 2016). The scope of this study was limited in the way that it did not train SVM models for all the Articles of ECHR. This was considered in further research conducted as an extension of this work. SVM models based on data represented using Bag of Words were trained for each of the Articles 2, 3, 5, 6, 8, 10, 11, 13 and 14 which achieved an average training accuracy of 75% (Medvedeva, Vols & Wieling, 2018).

Based on the related work, it has been observed that ANNs like CNN and RNN variants can provide promising results for various text classification problems. It depends on the data, whether capturing local key phrases or long-term semantic dependencies by a model would lead to better classification performance. Hence, it cannot be said beforehand which model out of CNN and RNN variants would perform better for any problem. For similar problem in ECHR, only general ML algorithm like SVM has been implemented and good performance results have been obtained. But there is a possibility that if ANNs are trained and evaluated for ECHR, they can outperform the results obtained from SVM models. This is the motivation which led to further investigation in this study.

3 Methods

The implementation of the research work was performed as per the CRISP-DM model. Accordingly, major steps incorporated in this study are Data Understanding, Data Preparation, Modelling, and Evaluation as depicted in Figure 3.1. In the first phase, existing dataset was obtained and understood as part of the Data Collection and Data Understanding unit. In the second phase, the experiment conducted previously in the referred research paper (Medvedeva, Vols & Wieling, 2018) was replicated. Data preparation was performed, and a model trained on SVM was designed. In the third phase, new experiment was conducted where data preparation was done differently followed by building a CNN model for each article. These two phases were part of Data preparation and Modelling units. In the last phase, performance of designed models was evaluated, and hypothesis testing was carried out using statistical tests. Detailed description of the experimental process has been provided in the following sections.

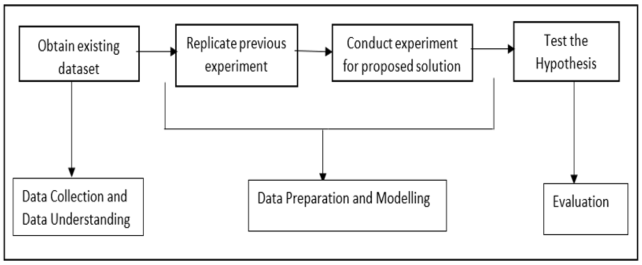
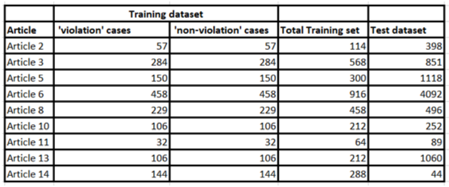

Obtain existing dataset
Replicate previous experiment
Conduct experiment for proposed solution
Test the Hypothesis
Data Collection and Data Understanding
Data Preparation and Modelling
Evaluation

Fig. 3.1. High Level Description of Experiments

3.1 Data Understanding

Dataset collection was a primary step performed for achieving the research objectives of this study. As data is the foundation on which all the models are built, it is important to ensure that the collected data is of high quality. Since the aim of this study was to compare the performance of previously built SVM models with the proposed CNN models, it was important to collect the same dataset to carry out such an analysis. For this reason, Crystal ball data was the dataset that was obtained from the previous research work (Medvedeva, Vols & Wieling, 2018) carried out in predicting the judgments of ECHR legal cases. It contained training and test dataset for each of the articles namely Article 2, 3, 5, 6, 8, 10, 11, 13 and 14, which correspond to different Human Rights. For each article, these datasets contained text files of different published cases along with their declared judgments by ECHR. They were obtained from HUDOC website on September 11, 2017[1]. All the text files present in the dataset were in a structured format where each file of any article had the same subsections. These subsections were Procedures, Circumstances, Relevant Laws and Facts,



	Training dataset			
Article	'violation' cases	'non-violation' cases	Total Training set	Test dataset
Article 2	57	57	114	398
Article 3	284	284	568	851
Article 5	150	150	300	1118
Article 6	458	458	916	4092
Article 8	229	229	458	496
Article 10	106	106	212	252
Article 11	32	32	64	89
Article 13	106	106	212	1060
Article 14	144	144	288	44

Table 3.1. Count of cases in Training and Test dataset for ECHR

[1] https://www.dropbox.com/s/lxpvvqdwby30157/crystal_ball_data.tar.gz

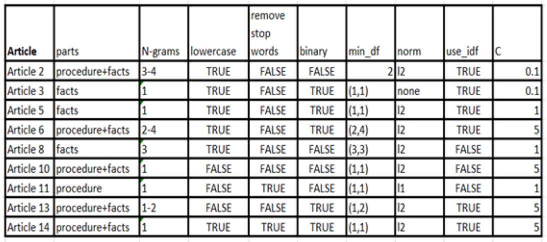
which were treated as features of the dataset. The target value of any case was of type binary and could take any of the two-class values namely 'violation' or 'non-violation', implying violation or non-violation of an article. The total number of judicial cases for all the articles in training and test dataset were 3132 and 8400 respectively. The distribution of cases for each article of ECHR for training and test dataset can be understood from Table 3.1. The training dataset was a balanced dataset and the test dataset contained cases with only 'violation' target for all the articles except Article 14 which contained cases with only 'non-violation' target. Other articles namely Article 4, 7, 9, 12 and 18 have been dropped as they contained very a smaller number of cases.

3.2 Data Preparation

The existing dataset was loaded and visualized for better understanding. It was observed that the data needed to be cleaned to improve its quality and text preprocessing was required to convert it into a form suitable for modeling. Since this study involved a comparison of previous work with the new solution, the previous work has been reproduced. Data preparation for SVM model was very different from the data preparation performed for CNN model.

Data Preparation for SVM Models

The text data was organized in the form of various features namely Procedures, Circumstances, Relevant Laws and Facts. All these features were not equally important in predicting the decision of a judicial case of ECHR. In the previous work, the best predictors were identified for each of the articles. In order to reproduce the previous work, only the data corresponding to the best predictor was retained. The text data was represented using N-grams as this was the representation used in the prior work and Tf-Idf Vectorizer was used to remove irrelevant N-grams from text. A grid search operation was conducted over different parameters for SVM model for each article and the best parameters found are described in Table 3.2. The data pre-processing for SVM model was performed as per the results of Grid Search operation.



Article	parts	N-grams	lowercase	remove stop words	binary	min_df	norm	use_idf	C
Article 2	procedure+facts	3-4	TRUE	FALSE	FALSE	2	l2	TRUE	0.1
Article 3	facts	1	TRUE	FALSE	TRUE	(1,1)	none	TRUE	0.1
Article 5	facts	1	TRUE	FALSE	TRUE	(1,1)	l2	TRUE	1
Article 6	procedure+facts	2-4	TRUE	FALSE	TRUE	(2,4)	l2	TRUE	5
Article 8	facts	3	TRUE	FALSE	FALSE	(3,3)	l2	FALSE	1
Article 10	procedure+facts	1	FALSE	FALSE	FALSE	(1,1)	l2	FALSE	5
Article 11	procedure	1	FALSE	TRUE	FALSE	(1,1)	l1	FALSE	1
Article 13	procedure+facts	1-2	FALSE	FALSE	TRUE	(1,2)	l2	TRUE	5
Article 14	procedure+facts	1	TRUE	TRUE	TRUE	(1,1)	l2	TRUE	5

Table 3.2. Best parameters for SVM models

Data Preparation for CNN Models

As discussed in the literature review, the advantage of ANNs lies in the fact that they can automatically extract important features from text without having to manually engineer them. Because of this reasoning, whole text data was considered for building predictive CNN models.

It is a well-known fact that word embeddings are good at capturing the context of text data (Rudkowsky et al., 2018), therefore this text representation has been chosen for building CNN model. They were assumed to provide better prediction accuracy as they captured relevant information from data. Initially, all the data from different subsections were extracted and merged together for each case as no feature selection was to be performed. Then, data cleaning operation was performed where unwanted symbols, white spaces, and digits were removed. Various techniques of NLP were employed for text data pre-processing. Punctuation marks were removed using regular expressions, all the text was converted to lowercase, pre-defined stop word list was used to remove stop words and all the words were lemmatized. Lemmatization was implemented instead of Stemming as it is considered more effective in areas like legal domain where language plays an important role (Plisson, Lavrac & Mladenic, 2004). Since there were too many words in the text data which could add noise, ten percent of the words with lowest tf-idf scores were removed for all the cases. This helped in removing less important document specific words which might not have been present in the pre-defined stop word list containing only most commonly used words. Tf-Idf scores were used instead of any other method like Zipf's law, etc. for generating list of irrelevant words because it determines the significance of words not only based on the document in which it is present but also across the entire collection of documents. The words with the lowest tf-idf scores have the least relevance across all documents, hence they were eliminated.

Various word embedding text representations have been used for preparing data for CNN models. The pre-trained word embeddings used are fasttext, GloVe and Word2vec. Since these embeddings have a vocabulary which is defined beforehand using data from different domains, a customized word embedding has also been learned in order to build the legal domain-specific vocabulary. This takes care of the fact that the pre-trained embeddings may not contain the words of legal domain and hence, they may or may not be good representations for ECHR data. All these representations have been implemented on the pre-processed text data. Different CNN models were built in the later stages with each of these word embeddings and a comparison was made to determine their effectiveness.

3.3 Modeling

Once the data pre-processing was carried out and text data was converted into vectors of real numbers using appropriate text representation, a model trained on SVM with linear kernel and a model trained on CNN were designed for each article to answer the research question.

SVM Modelling

For performing a comparative analysis, it was important to reproduce the previous work. To achieve this, an experiment was conducted to build an SVM model with linear kernel for each article of ECHR. All the parameters found using grid search as presented in Table 3.1 were employed in building SVM models. In the previous work, the performance of the models was evaluated using accuracy as the performance metric, considering the dataset was balanced in nature. To evaluate the performance of SVM model, 10-fold cross-validation was performed to determine the training accuracy of the model. A total of 9 SVM models, one for each article, were fitted on the training data. To check the testing accuracy, each of the models trained on SVM was applied on the test dataset and testing accuracies were reported. Confusion matrices and Classification reports for training and test datasets were generated to better understand the performance of the models. ***The results (training accuracies) obtained indicated that the experiment was successful at replicating the original work.***

CNN Modelling

The CNN modelling was carried out in two different experiments. In the first experiment, a comparison was made between different word embedding representations to identify the best text representation for training CNN model for ECHR problem. In the second experiment, the best-found text representations were used to determine the training accuracy of CNN models to compare their performance with the previous work based on SVM models.

After data preparation, different word embeddings like fasttext, Word2Vec, GloVe, and custom trained embedding were used for text representation. A conservative CNN architecture was considered as a starting point for all articles for the ease of comparison. Since DL models usually take more time to build than ML algorithms, holdout method (75:25 train-test split) was considered for evaluating the performance of the CNN models. The CNN models were fitted on the data and architecture changes were performed to find the best models. ***The criteria chosen to select the best model was that if the training accuracy of the CNN model for any of the considered text representations was more than that of the SVM model, the architecture was selected.*** If the model architecture was considered unsuitable for data, either the hyperparameters or the number of layers in the architecture were modified manually. Early Stopping and Checkpointing were used as the callbacks to monitor the model performance. Since accuracy was used as a performance measure for SVM models, the same metric has been used for this study to ensure consistency and ease of comparison. ***The best performing CNN model as per the selection criteria considered was found similar for all Articles except 5 and 11. The best-found word embedding was Word2Vec for articles 2, 8 and 10, GloVe for article 3, fasttext for articles 5 and 11, and custom for articles 6, 13 and 14.***

To get a common ground for such a comparison, same CNN architectures for corresponding articles with the best-found text representations were used.

CNN models were then trained and evaluated on the data using 10-fold cross-validation, which was used in the previous work. This helped in attaining training and testing accuracy which could be compared with the corresponding values of SVM models. ***Based on the results, it was found that the best-found CNN models performed better at predicting the judgements of ECHR for all articles as compared to reproduced SVM models.***

3.4 Evaluation

After the modeling phase, it was found that CNN models achieved higher training accuracy than SVM models. To determine if such differences were actual and not due to random chance, statistical significance difference tests were performed. For such tests to be performed, several samples of training accuracies for both SVM and CNN models for each article were required. To achieve this, 2 times repeated 10-fold cross-validation was performed to evaluate the performance of models. For each article, 20 (2*10) such training accuracies were obtained for SVM models and the best-found CNN models respectively. To perform difference tests, it was important to understand if the distribution of data samples was Gaussian. Visual normality inspections like histograms, probability distributions and QQ plots were performed. To further validate normality, normality test such as Shapiro-Wilk test was used on the distribution of training accuracies of both SVM and CNN models as this test is considered to be a highly powerful normality test (Mendes & Pala, 2003). If the p-value was less than 0.05, then one could reject normality hypothesis else one failed to reject it. The distribution was not normal for SVM models for Article 2 and 5, and CNN models for Article 5 and 11. For SVM and CNN models of all other articles, the distribution was found normal. If the distribution was found Gaussian for both the models for each article, then parametric difference test was conducted otherwise non-parametric difference test was conducted. Although the same dataset was used for SVM and CNN models, the data for both were processed in different ways and was shuffled randomly before training, hence independent samples difference test was considered. Student's t-Test was considered as parametric difference test and Mann-Whitney U Test was considered as non-parametric difference test for such case. The p-value obtained from the difference test was indicative of whether the difference in the training accuracies was statistically significant. If the p-value was less than 0.05, then it was stated with confidence that there was enough evidence to reject the null hypothesis and there was statistically significant difference in the training accuracy of both the models. If value was more than 0.05, then there was not enough evidence to reject the null hypothesis.

The findings could be related to the research question as the statistical tests helped in comparing both SVM and CNN models and determining whether there was a statistically significant difference in their training accuracies. If there was such a difference, then it could be concluded whether the solution in the referred paper or the solution proposed in this study was better at predicting the decisions of ECHR cases.

4 Results

This section provides an analysis of the results obtained and uses them to validate the hypothesis and answer the research question. It was observed that the best-found CNN models achieved a higher training accuracy than SVM models for all articles when their performance was evaluated using 10-fold cross-validation technique. To validate whether such difference in mean accuracies was statistically significant, difference tests were performed for all the articles. Based on the results obtained, the hypothesis assumed for different articles can be confirmed or refuted as below:

Article 2, 10, 11 and 14: Experimental results provided sufficient statistical evidence to reject the null hypothesis. A model trained with CNN, using Word embedding as text representation, achieved a statistically significant higher training accuracy than a model trained with SVM having the linear kernel, using N-gram as text representation, for predicting whether an Article of ECHR has been violated or not.

Article 3, 5, 6, 8, and 13: Experimental results failed to provide sufficient statistical evidence to reject the null hypothesis. A model trained with CNN, using Word embedding as text representation, did not achieve a statistically significant higher training accuracy than a model trained with SVM having the linear kernel, using N-gram as text representation, for predicting whether an Article of ECHR has been violated or not.

For Articles 3, 5, 6, 8 and 13, there exists a possibility that a statistically significant higher training accuracy could have been achieved if a different architecture of CNN model was considered. This requires further experiments and observations to be validated, which can be carried out as part of future work.

The results of hypothesis testing indicate that overall, the models trained on CNN achieved either similar or higher training accuracy than models trained on SVM with linear kernel for predicting the judgments of ECHR. ***The average training accuracy of CNN models was 82% which has been found higher than the average training accuracy of SVM models which was 75%. Thus, the CNN models trained as part of this study prove to be better predictive modeling solutions compared to the SVM models proposed in previous research (Medvedeva, Vols & Wieling, 2018) for the problem of legal forecasting for ECHR.***

5 Conclusion

5.1 Contributions and impact

Various experiments have been conducted in this study to develop a highly accurate legal forecasting system for ECHR. Based on the statistical test results, it has been found that the CNN models achieved either similar training accuracies (Articles 3, 5, 6, 8 and 13) or higher training accuracies (Articles 2, 10, 11 and 14) than SVM models. Overall, CNN models built as part of this study achieved an average training accuracy of 82% which is higher than 75% as achieved by SVM models designed in the previous research. This is a significant improvement in the accuracy with which the outcomes of the legal cases are predicted. This study makes significant contribution to the existing body of research on application of AI in the legal domain in terms of the building a legal forecasting system for ECHR using DL algorithm and word embeddings which have not been done before. The results act as proofs of the positive impact that the improved prediction accuracy will make in the legal domain. If such CNN models are deployed as legal assistant systems in real-time, they can fasten the litigation process and help in improving the current judicial state of ECHR. Also, the study opens various avenues of future research which can make great strides in automating many tasks of legal domain.

5.2 Future Work and recommendations

As part of future work, some recommendations can be suggested for further exploration. First, the CNN models were not trained for Articles 4, 7, 9, 12 and 18 due to less data samples available in the dataset considered for this study. Therefore, more published case judgments can be collected from ECHR website in order to design a comprehensive legal assistant system. Second, the part 'Circumstances' was identified as the most important part of text for determining the decision of a legal case (Aletras, Tsarapatsanis, Preoţiuc-Pietro & Lampos, 2016) using SVM models for Articles 3, 6 and 8. A similar experiment can be conducted to identify the best part(s)/predictor(s) to train CNN model for each article, instead of using the whole data. It can be verified whether CNN models trained on the best predictor(s) achieve higher training accuracy compared to the results obtained in this study. Third, since the dataset considered was collected in September 2017, further data augmentation can be performed by collecting data from the ECHR website post this date. The large size of dataset may improve the learning of CNN models and it can be investigated whether they yield better accuracy results. Fourth, the best-found CNN models in the study were the ones which achieved higher training accuracy than SVM models, for specific word embeddings. There exists a possibility that even better results can be obtained if further architecture changes are made to the CNN models for each article. Also, it might be possible that with custom trained word embeddings, such architectural changes can significantly improve the performance of the models. The experimental study can be conducted on this in future and the training accuracies obtained can be compared with the results of this study. Fifth, since GRU recurrent models usually perform well with text data, further exploration can be performed to find appropriate ways of fitting such models on training data. It can be identified whether learning long term dependencies in the text is more useful than learning key-phrases for predicting the judgments of ECHR. Sixth, the accuracy with which CNN models make decisions can be compared with that of an actual human judge on various cases of all articles. This will be useful in determining whether CNN models can be deployed as standalone systems rather than as legal assistant systems.

References

Lawlor, R. (1963). What Computers Can Do: Analysis and Prediction of Judicial Decisions. *American Bar Association Journal*, 49(4), 337-344.

Hughes, M., Li, I., Kotoulas, S., & Suzumura, T. (2017). Medical text classification using convolutional neural networks. *Studies in health technology and informatics*, 235, 246-250.

Amirkhan, R., Hoogendoorn, M., Numans, M. E. & Moons, L. (2017). Using Recurrent Neural Networks to Predict Colorectal Cancer among Patients. *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, 1-8.

Bermen, D. H. & Hafner, C. D. (1989). The Potential of Artificial Intelligence to Help Solve the Crisis in Our Legal System. *Communications of the ACM*, 32(8), 928-938.

Sulea, O., Zampieri, M., Vela, M., Dinu, L. P. & Genabith, J. (2017). Predicting the Law Area and Decisions of French Supreme Court Cases. *Proceedings of Recent Advances in Natural Language Processing*, 716–722.

Dunn, M., Sagun, L., Şirin, H., & Chen, D. (2017, June). Early predictability of asylum court decisions. *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, 233-236.

Undavia, S., Meyers, A., & Ortega, J. E. (2018). A comparative study of classifying legal documents with neural networks. *In 2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, 515-522.IEEE.

Sharma, R. D., Mittal, S., Tripathi, S., & Acharya, S. (2015). Using Modern Neural Networks to Predict the Decisions of Supreme Court of the United States with State-of-the-Art Accuracy. *International Conference on Neural Information Processing*, 475-483. Springer, Cham.

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., & Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A Natural Language Processing perspective. *PeerJ Computer Science*, 2(e93), 1-19.

Medvedeva, M., Vols, M., & Wieling, M. (2018). Judicial decisions of the European Court of Human Rights: Looking into the crystal ball. *Proceedings of the Conference on Empirical Legal Studies in Europe 2018*, 1-24.

Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., & Sedlmair, M. (2018). More than bags of words: Sentiment analysis with word embeddings. *Communication Methods and Measures*, 12(2-3), 140-157.

Plisson, J., Lavrac, N., & Mladenic, D. (2004). A Rule based Approach to Word Lemmatization. *Proceedings of IS-2004*, 83-86.

Mendes, M., & Pala, A. (2003). Type I error rate and power of three normality tests. *Pakistan Journal of Information and Technology*, 2(2), 135-139.