Predicting decisions of the European Patent Office's Boards of Appeal using Machine Learning

David Bareham[0009−0000−7542−8966]

Department of Computer Science, University of Liverpool, UK
d.bareham@liverpool.ac.uk

Abstract. This research assesses the feasibility of applying machine learning (ML) methods to the problem of case outcome prediction for appeals from the European Patent Office's (EPO) Boards of Appeal, concerning the grant of a patent application. The task is conceptualised as binary classification in which an appeal can *affirm* or *reverse* the prior judgement. Using a range of ML classifiers and textual representations, including custom-trained word and document embeddings, two experiments were conducted on appeal cases from both the Examining and Opposition Divisions of the EPO. The first experiment uses randomly-sampled data and the second uses year-stratified data, to perform prediction. The F1-scores achieved for the future prediction task across both divisions are 86.64% and 78.55% respectively. The results demonstrate the viability of applying ML techniques to predict the outcome of appeals concerning the patent grant procedure, and help to identify patents as a promising legal domain for future research. Furthermore, explainability analysis conducted with SHAP helps to identify a direction for future work concerning more robust explainability.

Keywords: European Patent Office · Machine Learning · Patents.

1 Introduction and Related Work

Patents, can be loosely defined as "a legal title granting its holder the right – in a particular country and for a certain period of time – to prevent third parties from exploiting an invention for commercial purposes without authorisation." [22]. The patent system acts as a mechanism to grant a limited commercial monopoly on an invention in exchange for technical disclosure of such an invention for a period of 20 years [4]. This is designed to create a mutually beneficial relationship between the patentee and the state, with the patentee able to legally enforce potential infringement from competitors, and the state able to proliferate the technical details of the invention to the public, which may otherwise have remained a trade secret. Economically the patent system is perceived to provide an incentive for new inventions and increased R&D spending [4], though there is ongoing debate within the literature as to the empirical validity of these

claims [12]. Regardless, patent applications have increased in volume internationally, more than tripling between 1985-2018 [11], and in 2022 there were 193,640 patent applications filed to the EPO[1] making it the patent jurisdiction with the fifth most filings internationally[2].

Filing patents across multiple jurisdictions can be costly and time-consuming so to mitigate this, the European Patent Convention (EPC) 1973 [3] was ratified creating a mechanism for the grant of multiple national patents within a single application [4]. Litigation and infringement are still dealt with by the respective national legal systems but the grant is handled by the EPO whose role is to administer the EPC. For a European patent application to be granted, the Examining Division of the EPO will assess the substantive content of the application according to a criteria including whether the invention is novel and there was an inventive step in its realisation.

After being granted by the Examining Division, a third party, i.e. a commercial competitor, has 9 months to object to the granting of the patent, which is heard before the Opposition Division. Any party who has been adversely affected at any stage, by the Examining Division or Opposition Division, may file an appeal against the decision. The Technical Boards of Appeal are responsible for appeals concerning refusal of a patent application from the Examining Division or appeals against decisions of the Opposition Division. The decisions granted in appeal proceedings are generally delivered at oral proceedings and have the force of *res judicata*, making the decisions subject to no further legal action [3].

The aim of this work is to assess the feasibility of engaging in case outcome prediction of EPO appeal decisions by applying ML techniques, in order to understand whether this previously understudied data source may be fruitful for further research. This work also aims to contribute to a growing body of literature on patent grant prediction from a different perspective, as prior work focuses on predicting whether a patent will be granted or refused at a department of first instance rather than after an appeal. A range of experiments will be reported considering different classification models, input representations and hyper-parameters across both randomly-sampled and year-stratified data. The evaluation uses standard performance metrics such as F1-Score, as well as analysing the explainability and interpretability offered by the best models. This evaluation will be used to identify scope for improvement in future work.

The task of case outcome prediction is the branch of Artificial Intelligence (AI) and Law referring to automatically predicting the outcome of a court decision given some input relevant to the decision. Most work focuses on the European Court of Human Rights (ECtHR) [1,19], the Supreme Court of the United States [15,16] and the Chinese Legal System [34,32].

Comparatively little research has been performed in case outcome prediction for the legal domain of Intellectual Property Law, encompassing sub-domains

[1] https://report-archive.epo.org/about-us/annual-reports-statistics/statistics/2022.html

[2] https://www.wipo.int/en/ipfactsandfigures/patents

[3] https://www.epo.org/en/legal/epc-1973/2006/convention.html

such as trademarks and patents. Whilst there has been an increased interest in applying computational analysis techniques to the field of Intellectual Property law, in particular patents [4], related work has primarily focused on the task of patent grant prediction [8,33,13] or predicting whether a specific patent is likely to be litigated or not [10,14]. This gap motivated the selection of the EPO Boards of Appeal as the focus domain for this work since, to the best of my knowledge, no prior work has computationally analysed the appeals process after a decision has been rendered on a patent application from a department of first instance.

2 Methods

2.1 Data & Pre-Processing

There are two distinct datasets from the EPO which form the basis of this work: Decisions of the EPO Boards of Appeal[5] and European Patent Full-Text Data for Text Analytics[6]. The European Patent Full-Text Data consists of XML-tagged titles, abstracts, descriptions, claims and search reports from 1978 onwards for patent applications. Text from approximately 500k publications was used for training the word and document embeddings described in Section 2.2. The Decisions of the EPO Boards of Appeal data formed the basis of the experiments conducted and consist of textual decisions from all subsidiary courts of the EPO Boards of Appeal from 1978-2022, with more than 40,000 decisions in XML format. Due to changes to the law over time as well as to the nature of patented inventions, the decision was made to constrain the time period of this work to decisions rendered after, and including, the year 2000. The language was also constrained to English, duplicate cases were removed and only cases falling within the remit of the Technical Board of Appeal were included. This left 21,426 unique appeals across both the Examining and Opposition Division.

Simple keyword matching using the SpaCy [25] library was constructed to identify the type of appeal (Examining Division, Opposition Division, Admissibility or Other), who brought the appeal (for the Opposition Division) and the decision outcome (Affirmed, Reversed or Other). The patterns used were manually created based on subsets of the data and were validated to ensure a sufficient accuracy[7]. Only cases which could be identified to have affirmed or reversed outcomes were used from appeals against the Examining Division or Opposition Division. In addition, for the Opposition Division cases, a dummy variable was added to capture the party bringing the appeal (though appeals in which both parties were appellants were excluded for this work).

[4] The term 'Patinformatics' coined in [30] refers to this process of patent data mining and using automated tools to extract insights and intelligence from patents [27].

[5] https://www.epo.org/searching-for-patents/data/bulk-data-sets/boards-of-appeal-decisions.html

[6] https://m.epo.org/searching-for-patents/data/bulk-data-sets/text-analytics.html

[7] Due to space limitations, further details on pre-processing, pattern matching, nested cross-validation results and hyperparameters are given at https://github.com/dahrb/EPO-Project/blob/main/JURIX_2023_DC__Supplemental_Information.pdf

The appeal texts were split into their constituent parts: Summary of Facts, Reasons for Decision and Order. The Summary of Facts outlines the facts of what happened in the prior decision, the core arguments the appeal is based on and the desired outcome for the appellants and/or opponents. The Reasons for Decision summarise the rationale from the board for coming to a particular outcome, which is given in the Order section. To predict the outcome of appeal cases *ex ante* we must use only data which was available before the verdict was given. For EPO appeals the only data currently available concerns decisions which have already been rendered, thus to test the possibility of predicting the outcome *ex ante* we must make a similar assumption as [1], that there is enough similarity between parts of the text of the published judgements and the information available prior to the proceedings. To justify this assumption, we exclude the Reasons for Decision section, as this is written in hindsight to justify an already decided appeal, but we use the Summary of Facts, as that purports to information available before the appeal proceedings.

Additionally, standard textual pre-processing steps were undertaken: removing non-alphanumerical characters and lowercasing all text.

2.2 Feature Engineering and Models

For input to the ML algorithms, the words within the Summary of Facts sections need to be represented numerically. A variety of traditional feature-based representations are used such as a bag-of-words (BOW) approach using n-grams, and an extension of this approach using the term frequency-inverse document frequency measure (TF-IDF) (as defined in SciKit-Learn [24]). Also explored are the use of pre-trained word embedding models, such as 300-d Word2Vec [20] trained by Google[8] and 200-d Law2Vec [6].

Two other embeddings were pre-trained specifically: Patent2Vec and PatentDoc2Vec. Both use the respective GenSim [26] package's implementations of Word2Vec and Doc2Vec with a corpus primarily consisting of data from the patent publications, with a sub-section of EPO Decisions data. The parameters used for both Patent2Vec and PatentDoc2Vec are the same as those used for Law2Vec [6] other than the dimensionality of the final embeddings (300-d).

The feature-based ML algorithms used were chosen due to their popularity in the task of legal case outcome prediction: Support Vector Machines [19], Random Forests [16], Logistic Regression [29] and XGBoost [2]. All algorithms were implemented using Scikit-Learn [24], except XGBoost which uses [7].

Furthermore, the models built with these algorithms and input representations are evaluated against a variety of classification metrics to assess performance. Initially a baseline classifier is defined using a very simple strategy. For our purposes, the baseline always guesses the majority class present in the dataset. The other evaluation metrics are Accuracy, F1-Score[9] and Matthew's Correlation Coefficient (MCC) [9].

[8] https://code.google.com/archive/p/word2vec/

[9] As defined in https://deepai.org/machine-learning-glossary-and-terms/f-score

2.3 Experimental Setup

The aim of the first experiment[10] is to understand what level of predictive performance is possible with this data using random sampling to generate the train and test sets. For this, the dataset is balanced regarding the outcome distribution before performing a stratified split into the training (90%) and the test (10%) sets. A balanced training set ensures the dataset does not have a bias towards the majority class. However, if we leave the test set balanced, when the real data is imbalanced, this could cause us to overestimate our predictive performance on an unrealistic test set. Following work such as [23,6], a realistic test set was created mimicking the original outcome distribution observed in the data.

To determine the best input representation, ML algorithm and hyperparameters for each domain, a k-fold nested cross-validation procedure is undertaken to mitigate overfitting in the model, and input, selection process [5]. Nested cross-validation differs from traditional, or flat, k-fold cross-validation procedures since it consists of a stratified outer cross-validation procedure, where each fold is used as the test data for an inner cross-validation procedure to tune the hyperparameters of the model [31]. This eliminates the positive performance bias introduced by flat cross-validation methods, in which the hyperparameters are tuned on the same data used to assess model performance and provides a more reliable way of assessing the model fitting procedure's performance.

After choosing the best combination of ML algorithm and input representation, a random search (100 iterations) is conducted over the entire training set to tune the hyperparameters, before test set evaluation.

The aim of the second experiment is to test the prediction of only future cases, following [19], by creating more realistic train and test sets. The issue with experiment 1 is that a case from 2019 may form part of the training set, and be used to predict the outcome, in the test or cross validation procedure, of a case from 2001, which fails to mirror the nature of the application of legal process happening in linear time. To mitigate against this issue, the training set and test sets are divided into cases from different years: training set from 2000-2018, and test set from 2019-2022, with further sub-divisions created in the test set to monitor whether there is an observable degradation in performance the further away the test set is in time from the training set, as observed in [19].

The procedure for selecting the best ML algorithm, hyperparameters and input representation is largely the same as experiment 1 other than both inner and outer cross-validation procedures use a method which splits the data according to the order they are given to the model (ascending date order), to ensure that the validation stage tests for the models that deliver the best future prediction capabilities[11].

[10] Experiments were undertaken on Barkla, part of the High-Performance Computing facilities at the University of Liverpool, UK. All code can be found at https://github.com/dahrb/EPO-Project.

[11] TimeSeriesSplit() in Sci-Kit Learn https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.TimeSeriesSplit.html

3 Results & Discussion

3.1 Experiments

After performing the nested cross-validation, an emergent pattern across both experiments was the dominance of n-gram approaches over the word and document embeddings. For example, in experiment 1 for the Examining Division, the highest F1-score for an embedding approach was Patent2Vec (73.16%) compared to TF-IDF (86.31%). Similar differences in performance are found across both experiments in all domains. One possible explanation for this disparity may be the way word embeddings were used as a word embedding only represents an individual word, thus to represent an entire appeal document the word embeddings must be combined into a single embedding. The choice was made in this work to simply average the word embeddings, but this could lose crucial information in representing the semantic space, decreasing performance.

To mitigate this limitation, a document embedding, PatentDoc2Vec, was also trained but did not achieve significantly better results, marginally outperforming Patent2Vec in the Opposition Division task yet underperforming against Patent2Vec for the Examining Division.[12] This may be due to the nature of the texts PatentDoc2Vec was trained on, as the vast majority of documents fed to PatentDoc2Vec were sections of patent applications, such as the claims, with only a small minority actually being the 'summary of facts' sections from decision documents.

Consequently BOW and TF-IDF representations were chosen for the final models with XGBoost as the ML algorithm because it reported the best results for all experiments and divisions.

Table 1 shows the results for the final models on the test data. We can observe a strong performance across all experiments, universally exceeding the baseline accuracy of only predicting the majority outcome (Affirmed). Despite this, a significant drop in performance ($\approx 10\%$) can be noticed between the Opposition and Examining Division scores. This is an unsurprising result as connecting the binary outcomes to the 'summary of facts' for the Opposition Division is a substantively more complex task.

For Examining Division cases it is always the patentee who is bringing the appeal with an 'Affirmed' outcome meaning that the patent application is rejected (as the prior decision is upheld) and 'Reversed' resulting in the patent's grant. For Opposition Division cases either the patentee or the opposition may bring the appeal (or both simultaneously[13]), and this is captured in the data with a dummy variable. However, the relationship between the grant of the patent application and the outcome is more complex than in the Examining Division case as, for example, an appeal can be lodged by the patentee against the patent being maintained in an amended form. In this instance, a 'Reversed' outcome

[12] Both the embeddings trained for this work outperform Law2Vec, but not Word2Vec, with statistical signifance $p < 0.05$ (Mann-Whitney Test).

[13] Though these appeals have been excluded for this work due to inaccuracy in identifying them using the pattern matching techniques.

may only result in the granted patent being amended further or granted as first proposed; while an 'Affirmed' outcome may only result in the patent maintaining its amended form from the prior decision. Either way, the substantive issue of whether the patent application ought to be granted or refused is a non-issue in these cases, whereas it is crucial in others. The dummy variable indicating who brought the appeal is unable to capture this legally substantive nuance, in combination with the textual representation, therefore a degradation in performance using these methods, over the Examining Division cases, was expected.

Another observation from Table 1 is that the results of experiment 2, across both divisions, on average outperform the experiment 1 results. These results are contrary to the observed degradation in performance with the future prediction task, compared to random sampling, as in the ECtHR domain [19]. One reason the future prediction results may be so strong is due to the use of the time series split for cross validation, meaning we are only picking models which are good at predicting future outcomes. Another may be the stability of a domain like patents and the criteria for their grant by the EPO Boards of Appeal, which have received little substantive change since its inception.

However, a small degradation is observed between the 2019-20 test set and the 2021-22 test set across both divisions. This may support the observation in [19] that the further away the test set is from the training data, the more performance decreases, though this observation could also be an artefact of the smaller test set in 2019-20 compared to 2021-22 ($\approx$ 200 less cases).

Table 1: Test data results for Experiment 1 and 2

	Patent Refusal				**Opposition Division**				**Average**			
	Acc	**F1**	**MCC**	**Baseline**	**Acc**	**F1**	**MCC**	**Baseline**	**Acc**	**F1**	**MCC**	**Baseline**
Experiment 1	86.39	86.24	72.58	56.36	78.22	78.20	56.89	54.33	82.30	82.22	64.74	55.35
2019-2020	**88.42**	**87.63**	**75.27**	62.92	**78.97**	**78.82**	**57.86**	56.01	**83.70**	**83.23**	**66.57**	59.47
2021-2022	86.33	85.33	70.71	**63.90**	78.61	78.27	56.71	**57.82**	82.47	81.80	63.71	**60.86**
2019-2022	87.52	86.64	73.30	63.35	78.78	78.55	57.29	56.94	83.15	82.60	65.30	60.15
Overall									82.91	82.46	65.08	58.96

3.2 Interpretation

To interpret the results of the best models we can use a method called SHAP (SHapley Additive exPlanations), proposed by [18] and derived from Shapley Values in game theory [28] to calculate the marginal contribution of different features by fairly allocating the contributions among them. For applying SHAP to XGBoost we use a model-specific method proposed called treeSHAP [17]. From this we can derive local interpretations to understand the marginal contribution of the features in a specific case, and we can derive global interpretations by aggregating these local interpretations.

An example of a global interpretation can be seen in Fig. 1 for experiment 2 within the 2019-2022 test set of Examining Division appeals. Here the positive

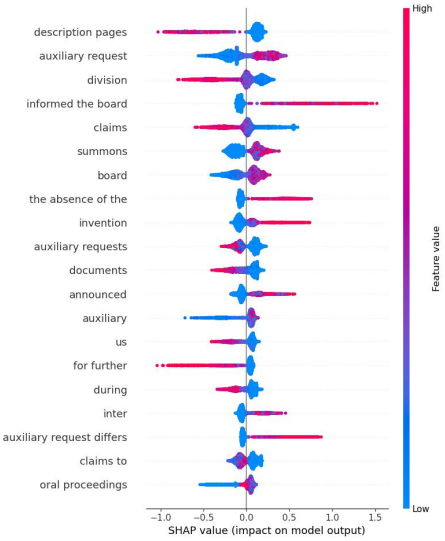

Fig. 1: Global SHAP

SHAP values indicate a feature which contributes to an 'Affirmed' outcome and negative SHAP values to a 'Reversed' outcome. The points in red denote that the n-gram is common within the given appeal and the points in blue denote that the n-gram is rare or not found within the given appeal.

Examining the features with the highest marginal contribution at either a global or local level, one can observe how they are mostly administrative or procedural in nature, e.g. 'description pages', 'informed the board' and 'auxiliary request differs'. Few, if any, of the features correspond with the kinds of legal factors one may expect to be relevant in deciding whether an appeal regarding a patent application's grant or refusal is successful, such as novelty or an inventive step. This demonstrates the limitations of post-hoc explanatory methods such as SHAP for legal case outcome prediction because despite offering interpretability into how the contribution of different features affects the outcome, the values themselves provide very little explainability useful in comprehending the reasons for the outcome. This is a problem which has previously been identified for NLP approaches in case outcome prediction [35].

4 Conclusion

The aim of this work was to assess the feasibility of engaging in case outcome prediction on appeal cases from the EPO's Boards of Appeal concerning patents. The results achieved by the models across both divisions and experiments (overall average of 82.46% F1-Score) are very encouraging for establishing both the domain and the application of case outcome prediction methods to patents as promising avenues for further research.

Despite the encouraging results, the lack of explainability available with the methods presented is a major limitation and future work as part of this project will focus on the correspondence between explicit legally relevant factors and the reasoning undertaken by a proposed model. One potential direction is the incorporation of Case-Based Reasoning systems with ML models into a hybrid system, with the benefit of explicitly representing the relevant legal factors [21]. Additionally, the results achieved are using sections of documents created after a decision has been rendered, therefore to truly assess predictive performance, only documents available before a decision is taken ought to be used.

References

1. Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D., Lampos, V.: Predicting judicial decisions of the european court of human rights: A natural language processing perspective. PeerJ Computer Science (2016)
2. Almuslim, I., Inkpen, D.: Legal judgment prediction for canadian appeal cases. In: CDMA. pp. 163–168 (2022)
3. Andrea, V.: Appeal procedure before the european patent office. In: Takenka, T. (ed.) Patent Law and Theory. Edward Elgar Publishing (2009)
4. Bently, L., Sherman, B., Gangjee, D., Johnson, P.: Intellectual Property Law. Oxford University Press, 5 edn. (2018)
5. Cawley, G.C., Talbot, N.L.C.: On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of Machine Learning Research **11**, 2079–2107 (2010)
6. Chalkidis, I., Kampas, D.: Deep learning in law: early adaptation and legal word embeddings trained on large corpora. Artificial Intelligence and Law **27**, 171–198 (2019)
7. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD. pp. 785–794. KDD '16, ACM (2016)
8. Chi, Y.C., Wang, H.C.: Establish a patent risk prediction model for emerging technologies using deep learning and data augmentation. Adv. Eng. Inform. **52**(C) (2022)
9. Chicco, D., Warrens, M.J., Jurman, G.: The matthews correlation coefficient (mcc) is more informative than cohen's kappa and brier score in binary classification assessment. IEEE Access **9**, 78368–78381 (2021)
10. Chien, C.V.: Predicting patent litigation. Law & Society: Private Law eJournal p. 48 pages (2011)
11. Hall, B.H.: Patents, innovation, and development. International Review of Applied Economics (2022)
12. Hall, B.H., Harhoff, D.: Recent research on the economics of patents. Working Paper 17773, National Bureau of Economic Research (2012)
13. Hido, S., Suzuki, S., Nishiyama, R., Imamichi, T., Takahashi, R., Nasukawa, T., Idé, T., Kanehira, Y., Yohda, R., Ueno, T., Tajima, A., Watanabe, T.: Modeling patent quality: A system for large-scale patentability analysis using text mining. Information and Media Technologies **7**, 655–666 (2012)
14. Juranek, S., Otneim, H.: Using machine learning to predict patent lawsuits. SSRN Electronic Journal (2021)

15. Katz, D.M., Bommarito, II, M.J., Blackman, J.: A general approach for predicting the behavior of the supreme court of the united states. PLOS ONE **12**(4), 1–18 (04 2017)
16. Kaufman, A.R., Kraft, P., Sen, M.: Improving supreme court forecasting using boosted decision trees. Political Analysis **27**(3), 381–387 (2019)
17. Lundberg, S.M., Erion, G.G., Chen, H., DeGrave, A.J., Prutkin, J.M., Nair, B.G., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I.: Explainable ai for trees: From local explanations to global understanding. ArXiv **abs/1905.04610** (2019)
18. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st NIPS. p. 4768–4777. NIPS'17 (2017)
19. Medvedeva, M., Vols, M., Wieling, M.: Using machine learning to predict decisions of the european court of human rights. Artif. Intell. Law **28**(2), 237–266 (jun 2020)
20. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space (2013)
21. Mumford, J., Atkinson, K., Bench-Capon, T.: Reasoning with legal cases: A hybrid adf-ml approach. vol. 362, pp. 93–102. IOS Press BV (12 2022)
22. Office, E.P.: European patent guide: How to get a european patent (2022), https://link.epo.org/web/how_to_get_a_european_patent_2022_en.pdf
23. O'Sullivan, C., Beel, J.: Predicting the outcome of judicial decisions made by the european court of human rights (2019)
24. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
25. Rehurek, R., Sojka, P.: Software framework for topic modelling with large corpora (2010), https://api.semanticscholar.org/CorpusID:18593743
26. Rehurek, R., Sojka, P.: Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic **3**(2) (2011)
27. Rezende, J.M.D., Rodrigues, I.M.D.C., Resendo, L.C., Komati, K.S.: Combining natural language processing techniques and algorithms lsa, word2vec and wmd for technological forecasting and similarity analysis in patent documents. Technology Analysis & Strategic Management pp. 1–22 (2022)
28. Shapley, L.: 7. A Value for n-Person Games. Contributions to the Theory of Games II (1953) 307-317., pp. 69–79. Princeton University Press (1997)
29. Strickson, B., De La Iglesia, B.: Legal judgement prediction for uk courts. In: Proceedings of 3rd ICISS. p. 204–209. Association for Computing Machinery (2020)
30. Trippe, A.J.: Patinformatics: Tasks to tools. World Patent Information **25**(3), 211–221 (2003)
31. Wainer, J., Cawley, G.: Nested cross-validation when selecting classifiers is overzealous for most practical applications
32. Yang, W., Jia, W., Zhou, X., Luo, Y.: Legal judgment prediction via multi-perspective bi-feedback network (2019)
33. Yao, L., Ni, H.: Prediction of patent grant and interpreting the key determinants: an application of interpretable machine learning approach. Scientometrics (2023)
34. Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., Sun, M.: Legal judgment prediction via topological learning pp. 3540–3549
35. Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z., Sun, M.: How does nlp benefit legal system: A summary of legal artificial intelligence (2020)