Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4717–4722
Marseille, 20-25 June 2022

Text Classification and Prediction in the Legal Domain

Minh-Quoc Nghiem[1,3,5], **Paul Baylis**[3], **André Freitas**[1,4], **Sophia Ananiadou**[1,2,5]
[1] Department of Computer Science, The University of Manchester, United Kingdom
[2] Alan Turing Institute, [3] Bott and Co Solicitors, [4] Idiap Research Institute, [5] National Centre for Text Mining
{minh-quoc.nghiem, andre.freitas, sophia.ananiadou}@manchester.ac.uk, p.baylis@bottonline.co.uk

Abstract

We present a case study on the application of text classification and legal judgment prediction for flight compensation. We combine transformer-based classification models to classify responses from airlines and incorporate text data with other data types to predict a legal claim being successful. Our experimental evaluations show that our models achieve consistent and significant improvements over baselines and even outperformed human prediction when predicting a claim being successful. These models were integrated into an existing claim management system, providing substantial productivity gains for handling the case lifecycle, currently supporting several thousands of monthly processes.

Keywords: text classification, legal judgment prediction, flight compensation

1. Introduction

Natural language processing (NLP) technologies can play an important role in supporting legal processes. In certain circumstances, NLP technologies can be used to aid lawyers as they investigate or review facts or details of a case. Additionally, it can also provide insight or automate the operational processes in legal work. There is increasing interest in the application of NLP technologies to the legal domain, and the field is slowly transforming as a result.

Accessing legal services is often expensive and requires paying a high fee to traditional lawyers. Many law firms offer "no-win, no-fee" agreements to make access to legal services easier and more affordable, where you pay a percentage of any compensation you receive if the case is successful. It is common for personal injury and flight compensation claims to be "no-win, no-fee", and the amount recovered is usually relatively small. While trying to maintain a quality service, many law firms have to reject smaller cases without a significant margin. In order to operate profitably, law firms have to reduce operating costs as well as assess the likelihood of a claim being successful before entering into a "no-win, no-fee" agreement with clients.

In this paper, we present our methodology of developing, deploying, and maintaining language processing technologies to classify airlines' responses to the claim and predict litigation outcomes of flight compensation claims. We were able to embed the results into our existing case management system, which dramatically improved email and document management for the team of claims handlers. By combining a transformer-based model with human-in-the-loop, our methodologies allow us to classify airlines' responses and forecast litigation outcomes. We also provide the details of the framework to allow us to continuously improve the deployed systems through retraining the models with additional data which is systematically collected and labeled.

Table 1 shows samples of airlines' responses and the labels we use to classify the responses. The responses can be classified into five classes: `Settlement`, `Denial`, `Paid Direct`, `Dealing Direct`, and `Need Further Infomation`. The problem can be formalised as a multi-class text classification where one or more labels from the predefined list can be assigned to a response (a response can be both `Settlement` and `Need Further Infomation` as seen on the second example of table 1).

Furthermore, when the response is a `Denial` one, we need to assess the likelihood of the claim being successful. Since the party who loses the case will be ordered to pay legal costs to the party who wins the case, we only make a formal claim at court when the likelihood of the claim being successful is high. The problem can be formalised as a regression problem where a prediction score close to 1.0 means the case has a high chance of success and a prediction score close to 0.0 means the case has a low chance of success.

This paper makes the following three key contributions:

- A practical method for legal text classification where we combine different transformer-based models to automatically classify different responses. We also present the methods to continuously improve classification results after system deployment.
- A model to predict the claim being successful for flight compensation where we combine text and non-text features to improve prediction performance. This is the first work done for flight compensation and the prediction is made at the very early stage before a claim is made at court.
- The description of the automation of a real-world legal workflow, using contemporary transformer-based methods.

Response	Label
Response text: ... Your clients are entitled to compensation for the delay to [FLIGHT NUMBER] on 18 August 2019. The distance of the disrupted journey was between 1,500km and 3,500km as calculated in accordance with EU legislation. Therefore, your clients are each entitled to €400.00 in compensation ...	`Settlement`
Response text: ... In order to settle this case, please find the General Release and Settlement of Claim forms and the Personal Data Collection and Processing Agreement forms attached for passenger's perusal and completion. Please return the completed forms to this email along with passenger's passport copy. Upon receipt we will action accordingly. ...	`Settlement, Need Further Infomation`
Response text: ... We've refused your claim for compensation because we told your clients about the cancellation at least 14 days before they were due to depart, using the contact details in their booking. Under EU legislation, we aren't liable to pay compensation for this kind of situation....	`Denial`
Response text: ... Hereby you will find enclosed the response to your complaint. ... **Attachment text**: ... We would like to inform you that, in accordance with Article 5.3 of Regulation (EC) 261/04, the carrier shall not be obliged to pay compensation if it can prove that the flight disruption is caused by extraordinary circumstances which ...	`Denial`

Table 1: Samples of airlines' responses and labels

The remainder of this paper is organized as follows. Section 2 provides a brief overview of the related work. Section 3 presents details of our methodology. Section 4 describes the experimental results and discussions. Section 5 concludes the paper and points to avenues for future work.

2. Related Work

In recent years, there has been an increasing amount of research on using machine learning for legal text interpretation. Common models for legal text classification include the construction of support-vector machine-based models (Octavia-Maria et al., 2017), or deep learning-based frameworks such as convolutional neural network (Hammami et al., 2021). Since 2019, with the introduction of the Bidirectional Encoder Representations from Transformers (BERT) language representation model (Devlin et al., 2019), researchers have been able to further increase the performance of legal text classification (Limsopatham, 2021; Sarkar et al., 2021). The ability of these models to capture large-scale linguistic phenomena and be fine-tuned to a specific domain suits the classification and inference requirements within the legal domain.

Besides legal text classification, several studies have attempted to predict the judicial decisions of the court. These approaches rely on different methods, such as rule-based (Ruger et al., 2004), decision trees (Ruger et al., 2004), random forest (Katz et al., 2016), support vector machines (Aletras et al., 2016) or deep learning models (Luo et al., 2017; Branting et al., 2021; Chalkidis et al., 2019; Zhong et al., 2018; Xu et al., 2020; Long et al., 2019). Most deep learning-based models use text as the main feature and do not integrate with categorical or numerical features. Previous work aimed to combine transformer-based models using text data with other non-textual features (Zhang et al., 2019; Gu and Budhkar, 2021; Ostendorff et al., 2019) but these have not been validated in the context of legal classification. To the best of our knowledge, this paper is the first work that reports an end-to-end legal processing workflow, supported by transformer-based classifiers, aiming to automate the textual interpretation and the prediction of a claim being successful in court (supported by other non-textual evidence).

3. Methodolody

Traditionally, in a law firm, documents are processed by human staff and saved to the Case Management System (CMS). In our proposed framework, the constructed models classify the documents to pre-defined classes before sending them to the staff for manual processing. The staff now need to verify the classification results before continuing their usual business workflow and save the document to the CMS. The verified labels can then be used to re-train the models to improve their performance. An outline of the workflow is depicted in Figure 1.

While integrating the classification models can substantially improve productivity, in order to maintain accountability and to provide legal safeguards, humans need to be kept in the loop during decision-making.

3.1. Data Preparation

We use two methods for collecting and preparing the data needed for model training. For historical data, we use heuristic labelling to obtain the labels. For new coming data, the labels are collected when the staff verify the classification results.

Prior to the system implementation, the staff had been using the CMS to save every document related to any legal case (fig. 1(a)) but they did not have any formal label nor save it to the CMS. These documents were, however, annotated with a comment after each document was analysed. For example "settlement email from airline" or "from airline, further info needed".

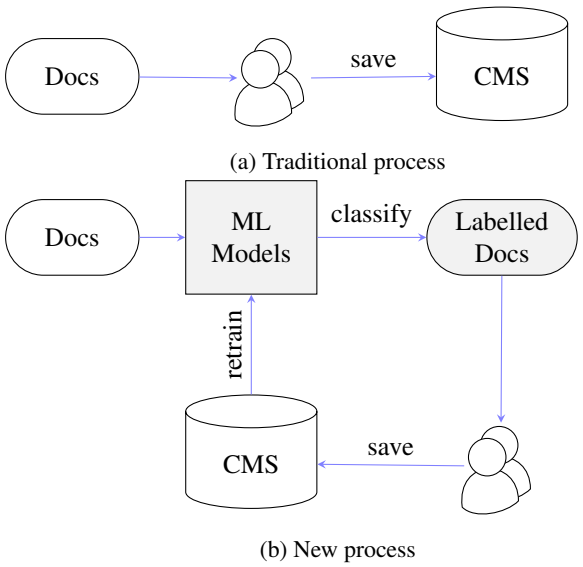

Figure 1: Human-in-the-loop workflow for claim processing.

Taking advantage of the comments, we use a rule-based string matching heuristic to get the labels for the training data. For example, if the comment contains the phrase "settle" or "settlement" and the document is from the airline, we assign the label `Settlement` to the document. A small portion of the data is then manually verified by staff for quality assurance. Test data is manually annotated by annotators from the law firm with legal background using the Paladin annotation toolkit (Nghiem et al., 2021).

For assessing the likelihood of the claim being successful, we use the "status" of the case to extract the needed labels. The status of a case is simply the indication of where the case stands at a particular time during a process. We use a mapping function to map the status of a case to get the final outcome of a process. A case now takes one of the four classes: `Success`, `Failed`, `Inconclusive` (client withdrew), or `Pending` (we have not yet known the result). For example, cases with the status "File Closed Valid Defence" or "Unsuccessful Court Claim" will have the label `Failed` while cases with status "Client Paid Claim Successful" or "Enforcement Proceeding Started" will have the label `Success` assigned. Currently, we only use `Success` and `Failed` cases for building the prediction model.

After the models are built, legal staff now use a new process when dealing with any new coming document (fig. 1(b)). Using the new process, the staff need to verify the classification outputs from the models before saving the documents to the CMS. This allows for a systematic and continuous improvement of the quality of the model. Schulz et al. found that verifying the suggestions have positive effects on the speed and performance of human annotator, while not introducing noteworthy biases (Schulz et al., 2019). The new labelled data is then combined with the current data to retrain the models periodically.

3.2. Airlines' Responses Classification

The model to classify airlines' response is depicted in Figure 2. There are three main components in the model: an input layer, a BERT (Bidirectional Encoder Representations from Transformers) model and a classification layer.

The input layer takes in the airline response text and the attachment text. The input representation is constructed by summing the corresponding token, segment, and position embeddings according to the procedure used by Devlin et al. (Devlin et al., 2019). The token `CLS` is put at the beginning of the sequence and the token `SEP` is used to separate the response text and the attachment text. When training, the BERT layer is first initialized with the pre-trained parameters ($BERT_{BASE}$ in this case), and all of the parameters are fine-tuned using labelled data from the classification task.

Because the model must learn representations for both responses with and without attachment, accuracy and prediction results may be affected (around 45 per cent of the responses have attachments, according to our data). Hence, we separate the responses into two groups: responses with attachment and responses without attachment, and train two separate models to handle each type of response. If there is no attachment to the response, the input layer is the same as the one depicted in figure 2 without the component after the first `SEP` token. Responses without attachments will be classified by this model, while responses with attachments will be classified by the full model.

The classification layer is added to the BERT output to predict the score for the labels. There are five output neurons in this layer, corresponding to the five classes. The alternative is to have only one output neuron for each class but to classify for all classes, we need five different models. In section 4 we report the results of the experiment with both methods.

3.3. Prediction of a claim being successful

The model to predict the claim being successful is depicted in Figure 3. There are four main components in the model: an input layer, a BERT model, a multilayer perceptron (MLP), and a classification layer.

Since the prediction of a claim being successful needs to be made before we make a formal claim at court, we are not able to access the features that are commonly used by other outcome prediction models such as "term", "court", "month of argument". Instead, we make use of the following features:

- Airline's response text: the denial response from the airline. An explanation of why the claim was rejected is usually included in the response. For example, the rejected reasons of the denial responses in table 1 are " we told your clients about the cancellation at least 14 days before they

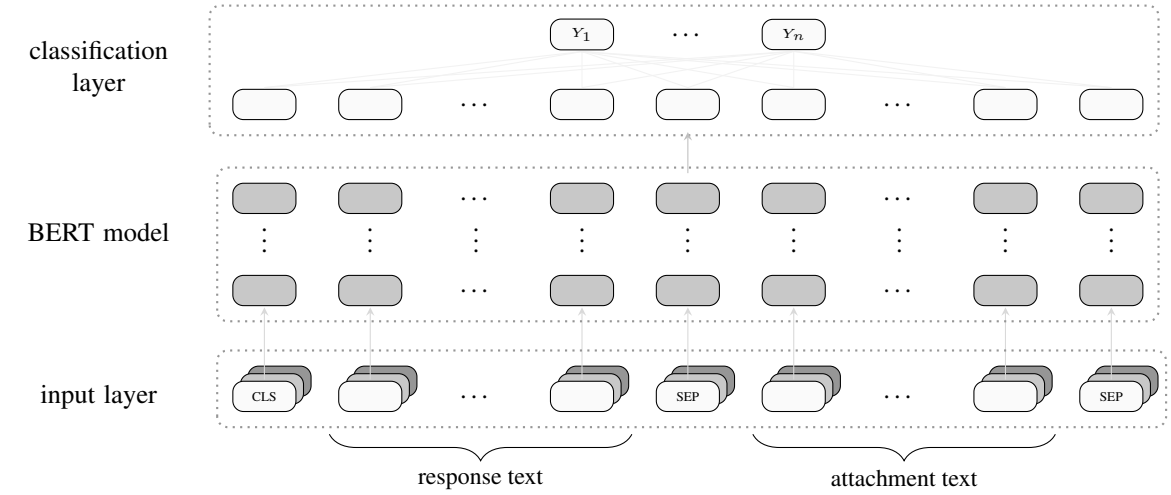

Figure 2: Airlines' response classification model

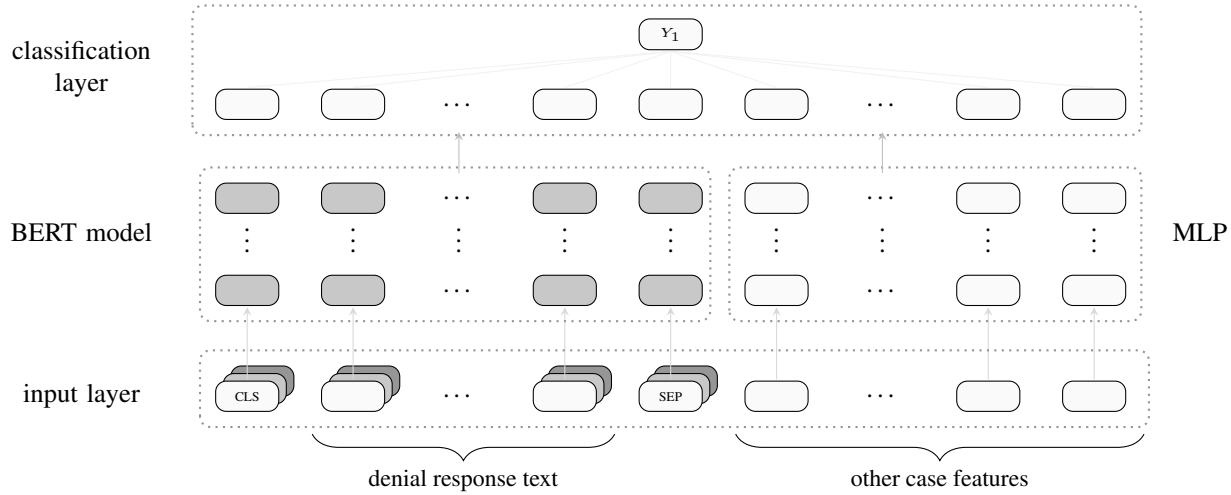

Figure 3: Prediction model for a claim being successful.

were due to depart" and "an extraordinary circumstance".

- Flight information: airline's name, departure airport, arrival airport, scheduled departure time, flight duration.
- Observed weather data: a 3-hour window of observed weather data at the departure airport from the scheduled departure time. The weather data includes the temperature (e.g. 14), the visibility (e.g. 9.9), the wind speed (e.g. 17.1), the weather condition (e.g. mostly cloudy, partly cloudy, clear). We can evaluate the validity of a flight cancellation due to adverse weather condition based on the weather data.

The response text and attachment text (if any) are concatenated and then fed through BERT while the flight information and observed weather data features are fed through MLP. We then concatenate the output from BERT model and MLP and feed the concatenated result to the final regression layer.

4. Results and Discussions

4.1. Experimental setup

For airlines' response classification, we use 4,472 responses for training and 1,000 for testing[1]. We compare four different approaches: 1. one multi-class classification model for all responses (the baseline, BERT_I) 2. two multi-class classification models for responses with and without attachment (BERT_{II}) 3. five single-class classification models for all responses, one for each class (BERT_V) 4. ten single-class classification models for responses with and without attachment, two for each class (BERT_X). For evaluation, we report micro precision, recall, and F1 scores.

For the model which predicts claim success, we use 2,759 cases for training and 500 for testing. We com-

[1]Sample data github.com/anonymous3007/datasample

pare two different approaches: PRED$_{TEXT}$ uses only the text feature (airline's response text and attachment text if any) and PRED$_{ALL}$ uses all available features (text and non-text). For evaluation, we report mean squared error (MSE), root mean square error (RMSE), and mean absolute error (MAE).

We used the pre-trained BERT base uncased model with outputs embeddings of size 768. We used three layers for MLP, the dropout rate was set to 0.1 for all fully-connected layers, and Adam (Kingma and Ba, 2015) was used as the optimizer with learning rate of 0.0001. We report results based on an average of three random seeds. The models were developed using the PyTorch library.

4.2. Results

Table 2 shows the precision, recall, and F1 scores of different approaches for airlines' responses classification. Though BERT$_X$ produced the highest scores, it requires ten different models to be trained and stored, and it takes five times longer to classify the data than the other methods. We can notice that models separating responses with and without attachment (BERT$_{II}$, BERT$_X$) outperforms other models consistently (BERT$_I$, BERT$_V$). We prefer BERT$_{II}$ and use it in practice because compared to other methods, it requires only two models to train, inference takes less time, and the results are comparable.

Method	Precision	Recall	F1
BERT$_I$	91.79	96.20	93.95
BERT$_{II}$	92.27	96.31	94.25
BERT$_V$	91.98	96.30	94.09
BERT$_X$	**92.37**	**96.80**	**94.53**

Table 2: Airlines' responses classification results

Table 3 shows per-class precision, recall, and F1 scores of BERT$_{II}$ model. The scores of `Settlement` and `Denial` classes are better than the scores of `Paid Direct`, `Dealing Direct`, and `Need Further Infomation`. This result may be explained by the fact that we have more data on `Settlement` and `Denial` than other classes (almost two-thirds of the data). Due to the nature of the flight compensation business, most airline responses are either accept or reject.

Class	Precision	Recall	F1
Further Info	91.84	88.24	90.00
Paid Direct	66.67	88.89	76.19
Denial	**96.91**	**98.74**	**97.82**
Settlement	92.64	97.81	95.16
Deal Direct	77.78	87.50	82.35

Table 3: Per-class scores of BERT$_{II}$ model

Table 4 shows the MSE, RMSE, and MAE of different approaches for the claim being successful prediction. The lowest error is achieved by PRED$_{ALL}$, even better than human performance which is measured by the ratio of cases won after litigation, approximately 62.78 per cent on the training data. This is only a rough estimation for human performance since we do not know whether non-litigation cases would win or not. According to these results, flight information and weather data are important evidence to support the claim success prediction.

Method	MSE	RMSE	MAE
PRED$_{TEXT}$	0.2221	0.4713	0.4495
PRED$_{ALL}$	**0.1684**	**0.4103**	**0.3191**
Human	0.3695	0.6079	0.3695

Table 4: Claim being successful: prediction results

4.3. Discussion

The results demonstrated that transformer-based classifiers can deliver results that can support the automation of legal workflows. All classifiers were considered to be functional within the supporting workflow, and within a human-in-the-loop setting, provide the balance between safety and productivity. Providing the supporting staff with the prediction from AI models could improve the quality and consistency of work. In this case study, the classifiers are closely integrated with the Case Management System. For example, in our case, when a human staff confirms a `Settlement` response, informing emails could be sent automatically.

A human-in-the-loop approach is also an effective way to systematically and continuously improve the model. While this point is self-evident, this paper provides a production-level account of a close integration of transformer-based models within an existing Case Management System workflow. The newly collected data has allowed us to achieve a 97% average F1 score with `Settlement` and `Denial` classes close to 99% (BERT$_{II}$ model with additional 4,000 responses as training data).

5. Conclusion

We introduced a case study of a legal analysis workflow, supported by transformer-based models for legal text interpretation and case prediction in the domain of flight compensation. By using transformer-based models to classify responses from airlines and by integrating text data with other types of data to predict a successful claim, this study has demonstrated the development of production-level quality legal text classification, contextualised within a Case Management System. Integrated text classification can support the consistency of the legal workflow, and support the economic feasibility of cases which otherwise could not be supported. Future work includes the integration of active and pro-active learning within the case management system and the analysis of the transferability of the proposed model to other legal domains.

Acknowledgments

This research has been carried out with funding from KTP11612 and NaCTeM. We would like to thank the anonymous reviewers for their helpful comments.

6. Bibliographical References

Aletras, N., Tsarapatsanis, D., Preotiuc-Pietro, D., and Lampos, V. (2016). Predicting judicial decisions of the european court of human rights: a natural language processing perspective. *PeerJ Computer Science*, 2:e93, 10.

Branting, L. K., Pfeifer, C., Brown, B., Ferro, L., Aberdeen, J., Weiss, B., Pfaff, M., and Liao, B. (2021). Scalable and explainable legal prediction. *Artificial Intelligence and Law*, 29(2):213–238.

Chalkidis, I., Androutsopoulos, I., and Aletras, N. (2019). Neural legal judgment prediction in English. In *Proceedings of the 57th ACL*, pages 4317–4323, Florence, Italy, July. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, pages 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Gu, K. and Budhkar, A. (2021). A package for learning on tabular and text data with transformers. In *Proceedings of the Third Workshop on Multimodal Artificial Intelligence*, pages 69–73, Mexico City, Mexico, June. Association for Computational Linguistics.

Hammami, E., Faiz, R., and Akermi, I. (2021). A dynamic convolutional neural network approach for legal text classification. In *International Conference on Information and Knowledge Systems*, pages 71–84. Springer.

Katz, D., Bommarito, I., and Blackman, J. (2016). A general approach for predicting the behavior of the supreme court of the United States. *PLOS ONE*, 12.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In Yoshua Bengio et al., editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Limsopatham, N. (2021). Effectively leveraging BERT for legal document classification. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 210–216, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Long, S., Tu, C., Liu, Z., and Sun, M. (2019). Automatic judgment prediction via legal reading comprehension. In *Chinese Computational Linguistics*, pages 558–572, Cham. Springer International Publishing.

Luo, B., Feng, Y., Xu, J., Zhang, X., and Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. In *Proceedings of the 2017 Conference on EMNLP*, pages 2727–2736, Copenhagen, Denmark, September. Association for Computational Linguistics.

Nghiem, M.-Q., Baylis, P., and Ananiadou, S. (2021). Paladin: an annotation tool based on active and proactive learning. In *Proceedings of the 16th Conference of the EACL: System Demonstrations*, pages 238–243, Online, April. Association for Computational Linguistics.

Octavia-Maria, Zampieri, M., Malmasi, S., Vela, M., P. Dinu, L., and van Genabith, J. (2017). Exploring the use of text classification in the legal domain. In *Proceedings of 2nd Workshop on Automated Semantic Analysis of Information in Legal Texts (ASAIL)*, London, United Kingdom.

Ostendorff, M., Bourgonje, P., Berger, M., Schneider, J. M., Rehm, G., and Gipp, B. (2019). Enriching BERT with knowledge graph embeddings for document classification. In *Proceedings of the 15th Conference on NLP, KONVENS 2019, Erlangen, Germany, October 9-11, 2019*.

Ruger, T. W., Kim, P. T., Martin, A. D., and Quinn, K. M. (2004). The supreme court forecasting project: Legal and political science approaches to predicting supreme court decisionmaking. *Columbia Law Review*, 104(4):1150–1210.

Sarkar, R., Ojha, A. K., Megaro, J., Mariano, J., Herard, V., and McCrae, J. P. (2021). Few-shot and zero-shot approaches to legal text classification: A case study in the financial sector. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 102–106, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Schulz, C., Meyer, C. M., Kiesewetter, J., Sailer, M., Bauer, E., Fischer, M. R., Fischer, F., and Gurevych, I. (2019). Analysis of automatic annotation suggestions for hard discourse-level tasks in expert domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2761–2772, Florence, Italy, July. Association for Computational Linguistics.

Xu, N., Wang, P., Chen, L., Pan, L., Wang, X., and Zhao, J. (2020). Distinguish confusing law articles for legal judgment prediction. In *Proceedings of the 58th Annual Meeting of the ACL*, pages 3086–3095, Online, July. Association for Computational Linguistics.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., and Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the ACL*, pages 1441–1451, Florence, Italy, July. Association for Computational Linguistics.

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z., and Sun, M. (2018). Legal judgment prediction via topological learning. In *Proceedings of the 2018 Conference on EMNLP*, pages 3540–3549, Brussels, Belgium, October-November. ACL.