applied sciences

MDPI

Article

Predicting Institution Outcomes for Inter Partes Review (IPR) Proceedings at the United States Patent Trial & Appeal Board by Deep Learning of Patent Owner Preliminary Response Briefs

Bahrad A. Sokhansanj * and Gail L. Rosen

Department of Electrical & Computer Engineering, College of Engineering, Drexel University, Philadelphia, PA 19104, USA; glr26@drexel.edu
* Correspondence: bahrad@bahradlaw.com

Abstract: A key challenge for artificial intelligence in the legal field is to determine from the text of a party's litigation brief whether, and why, it will succeed or fail. This paper shows a proof-of-concept test case from the United States: predicting outcomes of post-grant inter partes review (IPR) proceedings for invalidating patents. The objectives are to compare decision-tree and deep learning methods, validate interpretability methods, and demonstrate outcome prediction based on party briefs. Specifically, this study compares and validates two distinct approaches: (1) representing documents with term frequency inverse document frequency (TF-IDF), training XGBoost gradient-boosted decision-tree models, and using SHAP for interpretation. (2) Deep learning of document text in context, using convolutional neural networks (CNN) with attention, and comparing LIME and attention visualization for interpretability. The methods are validated on the task of automatically determining case outcomes from unstructured written decision opinions, and then used to predict trial institution or denial based on the patent owner's preliminary response brief. The results show how interpretable deep learning architecture classifies successful/unsuccessful response briefs on temporally separated training and test sets. More accurate prediction remains challenging, likely due to the fact-specific, technical nature of patent cases and changes in applicable law and jurisprudence over time.

Keywords: law; litigation; natural language processing; explainable artificial intelligence; interpretable machine learning; patents; post-grant reviews

Citation: Sokhansanj, B.A.; Rosen, G.L. Predicting Institution Outcomes for Inter Partes Review (IPR) Proceedings at the United States Patent Trial & Appeal Board by Deep Learning of Patent Owner Preliminary Response Briefs. *Appl. Sci.* **2022**, *12*, 3656. https://doi.org/10.3390/app12073656

Academic Editor: Valentino Santucci

Received: 21 February 2022
Accepted: 3 April 2022
Published: 5 April 2022

1. Introduction

Law is fertile ground for applications of artificial intelligence (AI) [1]. Litigation and transactional legal advice often involves analyzing a large amount of information, and legal reasoning has properties amenable to computation. The law is made up of a closed universe of legal statutes and cases that can be used as precedents, and arguments and decisions are formulated using logical reasoning. AI applications in law have been extensive and growing in such areas as:

- contract preparation and analysis [2,3];
- electronic document analysis for fact investigation and discovery in litigation [4–6];
- automatic preparation of documents, analyzing patents and patent portfolios [7–10];
- enhancing legal research, i.e., identifying relevant cases and statutes [6,11];
- legal analytics to predict litigation outcome based on the identity of judges, common fact patterns, and other discrete characteristics [12–14].

While legal documents have certain common formal and structural elements, they are written and read by humans. Accordingly, Natural Language Processing (NLP) methods that use machine learning to derive meaning from natural, unstructured, and informal text are particularly applicable to law. NLP has been commercialized for searching case law

and electronic documents, as well as in the area of contracts, patents, and form documents in litigation such as Complaint and Answer litigation pleadings [2,6,15]. These applications either involve short lengths of text (e.g., search query and target phases), structured text that follows standard formats (contracts and pleadings), or technical and specialized subject matter characterized by jargon words and phrases that can readily characterize documents (patents).

NLP is still highly experimental in the area of litigation briefs and opinions. Briefs are documents written by attorneys to argue their case on a preliminary issue (generally called a "motion") or for judgment. Opinions are written by judges to explain their reasoning to justify their decision, guide future litigators, and persuade judges on a higher court who may hear an appeal. Because of the persuasive element, briefs and opinions are written in a more loose style and with less formal language than precise technical documents like contracts and patents. Analyzing briefs and opinions is therefore a very difficult task for computers [16]. Machine learning in litigation has thus primarily focused on non-textual features, such as the identity of the judge, and encoding the legal and factual issues and the type of argument. For example, to predict the outcome of a dispositive summary judgment motion in a patent lawsuit, a machine learning model can be trained to predict the chance of success based on such features as (i) the subject matter coding of the patent, (ii) the age of the patent, (iii) licensing history of the patent, (iv) the kind of arguments that the defendant has raised (patent ineligibility, prior art, infringement), and (v) the court where the dispute will be heard (e.g., in the United States, Texas courts tend to be more favorable for plaintiffs and California courts to defendants) [17,18].

A feature subset-based model of the text of litigation briefs will miss additional information that would be found in the full text. By contrast, a model based on the natural language in full context can be more flexible, efficient, and less biased than using feature subsets. Specifically, a predictive NLP model based on the text of briefs will naturally derive patterns of factual and legal issues that are in dispute, and how the parties have argued their cases. There is no model designer who must make subjective choices of what keywords or issue codes to use to characterize the text, which are choices that can reflect a bias or miss certain issues. Feature encoding (or transformation) methods also require a human reader who will subjectively set the value of features, such as whether a brief raises an argument of patent obviousness or not. NLP models, by contrast, learn semantic relationships and word-meaning in the course of training, alleviating time-consuming review to subjectively encode documents. Moreover, if a predictive NLP model can also be *interpretable* or *explainable*, then the model can derive additional useful insight [19]. For example, an interpretable model may reveal what linguistic patterns are particularly persuasive, or changes over time in how judges evaluate particular fact patterns and legal argument.

In this paper, we demonstrate a proof-of-concept of analyzing litigation briefs and opinions using two diverse interpretable machine learning methodologies: (1) Extreme gradient boosting for ensemble decision tree modeling (XGBoost) based on the term frequency-inverse document frequency (TF-IDF) of individual words, and (2) a whole-document contextual convolutional neural network (CNN)-based model that includes attention layers for highlighting potentially significant features. We analyze administrative litigation before the United States Patent Trial & Appeal Board (PTAB) to resolve post-grant patent challenges called post-grant review (PGR), inter partes review (IPR), and covered business method patent reviews (CBM). IPR, PGR, and CBM cases are similar proceedings in which a party petitions the PTAB to challenge the validity of a patent held by a patent owner [20,21]. This paper focuses on the first stage of these proceedings, when a panel of PTAB judges decides whether to institute a trial with evidence on the petitioner's brief supporting their challenge, usually taking into account a preliminary response brief submitted by the patent owner. In these proceedings, the petitioner (i.e., challenger) is almost always the defendant in a patent lawsuit brought in U.S. federal court by the patent owner. As such, these are generally high stakes cases litigated by sophisticated lawyers.

The legal and factual issues are specific to particular patent validity challenges—but the text of the parties' briefs is generally unstructured and has a lot of stylistic variation

as lawyers attempt to craft the briefs to be as persuasive to the decision-makers as possible. Therefore, while PTAB represents a closed universe of a particular kind of litigation, it nonetheless poses a highly challenging textual analysis problem. Adding to the complexity of the challenge is that while all the relevant documents are publicly available, they are available only as PDF files which must be converted to text, a frequently noisy process (while clean text is available through commercial databases, those databases generally do not offer bulk download options and have restrictive terms of use.) Soon, all documents in U.S. litigation proceedings will be freely available as well; however, they will likely be only available in a similar format. In sum, our PTAB analysis represents a realistic, practically relevant problem.

To validate and demonstrate as a proof-of-concept interpretable machine learning to provide insight in the analysis of litigation documents, this paper focuses on two analytical tasks: (1) Deriving the outcome of the proceeding from the written opinion prepared by PTAB judges, and (2) predicting the outcome of the case based on the patent owner's preliminary response brief. The goal of the studies shown here is to demonstrate that interpretable machine learning has the potential to obtain practically useful insight from litigation brief text as a component of litigation data analytics.

The paper proceeds as follows. Section 2 describes the legal problem, reviews relevant literature on machine learning and NLP in the legal field, and specifically reviews related work applying deep learning methods and interpretation methods to litigation outcome prediction. Section 3 describes the methods used to collect data, pre-process documents, classify documents to predict outcomes (including the detailed structure of the deep learning model), and interpret trained machine learning models. Section 4 shows the results of deriving the outcome from the written PTAB decision, comparing results using decision tree-based and deep learning, and the results for outcome prediction based on the patent owner's preliminary response brief. Section 5 discusses the qualitative comparison between the methods used in the study, challenges that were found in predicting the outcome of litigation briefs, important caveats, and directions for future work. The paper concludes with a brief summary of key findings.

2. Background and Related Work

2.1. Legal Context: Post-Grant Review Proceedings

This paper studies a particular kind of litigation: post-grant patent review proceedings in the United States before a specialized tribunal: the Patent Trial & Appeal Board. Briefly, in 2011, the United States Congress passed a law called the Leahy–Smith America Invents Act (AIA), which went into effect on 16 September 2012. The AIA created three new review mechanisms for U.S. patents: inter partes review (IPR), post-grant review (PGR), and covered business methods review (CBM) [20]. These three mechanisms are very similar. In all three, a party files a petition with the U.S. Patent & Trademark Office (PTO) challenging the validity of a U.S. patent based on legal and factual bases, called "Grounds", set forth in the petition. The PTO Director then has the discretion by statute to decide whether to institute a trial on the challenge. If instituted, the trial is generally decided within 12 months by a panel of three Administrative Patent Judges (APJs, or for the purposes of this paper we will refer to them as "judges") within the Patent Trial & Appeal Board (PTAB) unit of the PTO. In practice, the Director further delegates the authority to initially determine whether to institute a trial to the panel of judges who would later hear and decide the trial if instituted.

Figure 1 shows a schematic of the post-grant review proceeding. The patent owner is granted an opportunity to file a preliminary response to the petition, often called a patent owner's preliminary response (POPR). While the preliminary response is optional, it is filed in most cases, and one study showed that, empirically, just filing it increases by 31% the probability of the patent owner ultimately succeeding in the proceeding, in part by increasing the chance of the petition being denied at the initial stage [22]. Within 6 months of the petition being filed, the APJs issue an initial determination. The initial determination

states whether to grant or deny trial on the petition, and thereby proceed to the next phase of the proceeding. APJs must grant the petition and institute trial if there is a reasonable likelihood of success for any ground of invalidity raised in the petition.

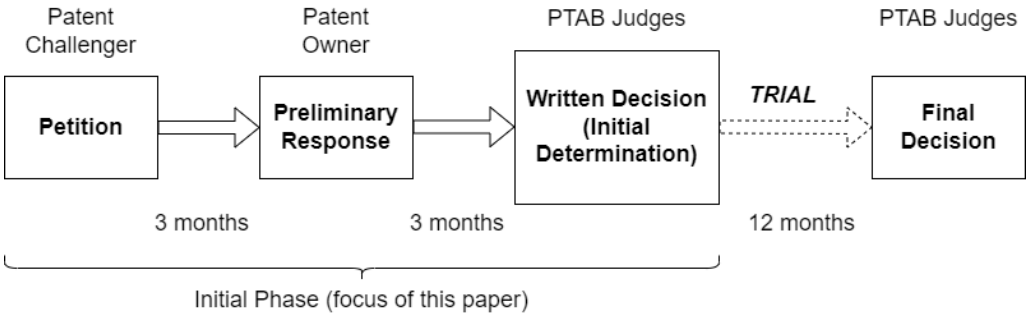


Figure 1. Timeline for IPR, PGR, and CBM post-grant patent review proceedings before the U.S. Patent and Trial Appeals Board (adapted from [23]). The proceeding begins with a petition, and then within 3 months, the patent owner may file a preliminary response brief. No later than 6 months after the petition is filed, a panel of PTAB judges determines whether to institute trial. If they deny trial, the proceeding ends. If they grant a trial institution, then after 12 months of additional briefing, expert depositions, and oral hearing before the judges, a final verdict on patent validity will issue. This paper focuses on the first phase of the proceedings, up to the decision to institute trial, which is also called the initial determination.

The primary differences between IPR, PGR, and CBM proceedings is what patents are eligible, and what invalidity grounds can be asserted. An IPR petition can be filed against any patent by anyone. An IPR, however, must be based solely on grounds that have as their factual basis printed prior art (i.e., patents or publications). Usually IPRs are brought by the challenger that was sued on the patent in a separate federal lawsuit brought by the patent owner. PGR and CBM petitions may, by contrast, be based on additional grounds of invalidity available under U.S. patent law (such as a lack of support for claims in the patent specification). A PGR can be filed against patents within 9 months of it issuing and only on inventions more recent than 16 September 2012. CBM has since been phased out; when it was available, it could only target a "business method" patent, and it has been phased out. In this paper, we consider all petitions together. In all cases, usually the challenger filing the petition is an entity that was sued for infringing on the patent in federal court.

The rules for post-grant review proceedings have been established by the AIA statute, regulations promulgated by the PTO and precedents set through PTAB decisions, and a number of U.S. Supreme Court decisions since the proceedings began in 2012. Importantly, the rules and judicial tendencies have changed over time—an aspect of the data (a.k.a. concept drift) that has previously been identified as a major challenge for applying machine learning to all kinds of litigation [24]. For example, in April 2018, the U.S. Supreme Court decided in *SAS Institute, Inc. v. Iancu* that if any ground in a petition is deemed to have a reasonable likelihood of success, then trial should be instituted on all grounds. Before that decision, PTO rules provided that judges could grant trial on one or some grounds and deny trial on the others, leading to a "mixed" result, i.e., partial victory for a patent owner if certain claims escaped review. Another change that occurred in October 2018 is that the PTO changed the legal standard for how broadly or narrowly PTAB judges should interpret claims [25].

Yet another significant change, examined in this paper, is the increased importance of arguing that a petition should be denied when there was parallel litigation that could be resolved before a PTAB trial would terminate. This change was based in PTAB decision in *NHK Spring Co., Ltd. v. Intri-Plex Techs., Inc.* (case number IPR2018-00752) in September 2018 that was designated as precedential by the PTO Director (and thus binding on all APJs) in May 2019, further reinforced by its decision in *Apple, Inc. v. Fintiv, Inc.* (IPR2020-

00019) in March 2020 (designated as precedential in May 2020). These cases established a controversial policy in which PTAB would more likely deny trial institution in the event of a parallel proceeding that could be resolved along the same or shorter timeline than a PTAB proceeding on the grounds of avoiding redundancy and improving efficiency [26,27]. The net result was a reduction in petition grant success rates in 2019 through 2020; however, it appears as though in 2021, this drop has reversed, coincident with new leadership at the PTO [28]. Among the learning tasks shown in this paper is training a model to identify whether a patent owner raises an argument based on parallel proceedings in the preliminary response brief.

2.2. Machine Learning and Natural Language Processing in Legal Applications

Applying machine learning (ML) and AI methods to law have been a topic for speculation and experimentation since at least the 1960s [29]. As computational power and algorithmic sophistication have rapidly improved, exemplified by the emergence of natural language processing (NLP) as a practical tool, ML and AI applications in the legal domain have become more prevalent [30]. As such, the background that we provide here will not be exhaustive, and will focus particularly on the most relevant work, particularly the analysis of patent review proceedings and predicting litigation outcomes.

The earliest work in modeling judicial decisions was based on Boolean equations and logical rule-based approaches [31,32]. Prior to the development of more computational resources and capacity for machine learning, much quantitative legal analytics work was based on analyzing citations to case law and statutes [33]. In the United States, the use of text and other features to predict the outcome of judicial proceedings has often focused on the federal Supreme Court, for which there exist readily (and freely) accessible opinions, notable interest issues of legal importance, quantifiable divisions in legal philosophy and ideology, and consistency in the judges (Justices) who make decisions [34,35]. Recent outcome prediction modeling has been able to achieve up to 70% accuracy at the level of predicting Supreme Court Justices' votes based on various features [36]. Other work has looked at other closed legal universes, where there is a known body of statutes and a constrained body of facts, such as tax law [37,38]. Notably, besides simple outcome prediction, other work has used computational text analysis or NLP methods, such as document vectorization, to examine measures of judicial sentiment and modes of reasoning in U.S. appellate courts, where again there are often jurisprudential (philosophical) and ideological divisions [39]. Researchers have also sought to analyze the writing style of jurists using NLP, such as U.S. Supreme Court Justice Neil Gorsuch [40].

A popular way to summarize and classify documents has been to represent them as a vector of the TF-IDF values of document terms [41]. In this approach, a training set of documents defines a vocabulary of terms that appear in those documents. In each document, each term in the vocabulary is assigned the ratio of the frequency of that term in the document to the frequency of the term in all documents. This method and its variants have been extensively used in many fields, including legal document analysis [42–44]. Another approach that has emerged for translating documents to linguistic features is representating terms by pretrained embedding vectors. These pretrained word embeddings, which are generated by self-supervised machine learning on a large corpus of documents, resulting in vectors in a space where distances reflect semantic relationships between the words. Foundational approaches include GloVe and Word2vec [45,46]. While these embeddings have generally been trained on large corpora such as Wikipedia articles or news articles, recent work has sought to train specialized word embeddings on legal corpora [47].

In the European context, several researchers have studied litigation prediction problems. Aletras et al. applied support vector machines (SVM) to n-grams (small groups of words) to predict the outcome of European Court of Human Rights (ECHR) decisions, obtaining a level of prediction accuracy of over 70%, although accuracy appeared to decline over time, suggesting drift in case law [13]. Medvedeva et al. more recently applied SVM

with TF-IDF (term frequency-inverse document frequency) and n-gram vectorization to ECHR decisions [48]. While they looked at patent text rather than litigation documents, word2vec embeddings were classified using with random forests (RF) ensemble decision tree methods to use textual features of patents along with other features (such as numbers of references and forward citations) to predict the litigation risk of patents [49].

Similar work has also been done with other features and traits of patents following Chien's groundbreaking work [17,50]. TF-IDF with an AdaBoost decision-tree based classifier has been used to predict motion outcome based on judicial complaints. In that work, the authors also vectorized document text using word2vec and doc2vec pretrained embeddings along with pre-selected features like attorney name, and coded case types were used to train classifiers that achieved an accuracy of over 60% on a 15 year data set of motions to strike complaints in Connecticut state civil tort and vehicular cases [51]. Convolutional neural networks (CNN) have been used with Bag of Words (BoW) vector representations for verdict prediction for cases in courts in India [52]. In this paper, we apply eXtreme Gradient Boosting (XGBoost) [53], which is a boosted decision tree-based ensemble machine learning method, to classify documents using TF-IDF vectors. Along those lines, researchers in China found that XGBoost outperformed methods including KNN, SVM, and Naive Bayes Classifiers for the prediction of theft crimes based on textual information [54]. Another recent study used XGBoost for prediction tasks on a database of consumer lawsuits in Brazil [55].

Shifting focus to PTAB proceedings in particular, researchers have used PTAB decisions as an empirical testbed for legal NLP applications; for example, developing document retrieval methods by extracting patents relevant to a PTAB determination [56]. Another work used pre-selected features from post-grant reviews to develop a model for patent quality [57]. The first work seeking to predict institution decision outcomes for PTAB proceedings employed Random Forests and SVM on TF-IDF vectorizations of patent claim text [58]. Although they reported greater than 50% accuracy, their study did not distinguish between cases where parties joined ongoing proceedings or settled proceedings by terminating them from petition grant and denial outcomes [58].

Subsequently, in 2018 another group also used SVM but on three different kinds of variables [59]. The researchers considered a "text-based approach", "entity network", and "domain knowledge". For the first, they computed paragraph-level attention using a CNN and hierarchical attention framework derived by [60] (further discussed below). The second, entity network, was an embedded graph using Node2Vec that included information on the patent owner, petitioner, and judges. The final domain knowledge category included characteristics of the patent, such as the number of references and forward citations, and the number of claims, and IPR proceedings, such as the number of words in the petition, and the institution rate of judges. They found an accuracy of 74% and 0.83 AUC for balanced data (they had a 67% institution rate in their dataset), noting that the entity network provided the greatest contribution to accurate prediction, while the weakest was domain knowledge. They also found that when testing temporally out-of-domain cases filed later in time than the training cases, the best results were for earlier cases in the test set. While the researchers carefully considered their selection of features, some of the results suggest potential overfitting. For example, one significant word the model identified was "DJC", which lacks general meaning. Furthermore, even temporally separated test cases and training cases may be related, which makes it a challenging area to establish generalization. Moreover, some features are duplicative. For example the identity of judges reflects the subject matter of patents, since PTAB judges are specialized to particular subject matter categories [20].

2.3. Deep Learning and Interpretability: Application to Litigation Analysis and Prediction

Recent advances in hardware and neural network architecture have enabled a movement away from TF-IDF and other document vectorization methods to deep learning: end-to-end machine learning to obtain document models and classifiers, in which the text

itself is the input to the model [47]. Deep learning can be both powerful and flexible, as models can learn contextual semantic relationships between words and phrases on their own. However, because deep learning approaches generally rely on neural network black box models, due to their multilayer nonlinear structure, they can predict an outcome but are difficult to interpret and understand [61].

By way of example of deep learning applications to litigation, in 2018, a group studying criminal cases in China developed a model that used long-short term memory (LSTM) recurrent neural network (RNN) model with an attention layer to do simultaneous criminal charge and attribute prediction based on analyzing text description of facts [12]. Further work by researchers working in the Chinese legal context found that extracting semantic features was superior to identifying features using SVM, and that increased model complexity was necessary to account for the nonstandard and out-of-vocabulary nature of legal terminology [62,63].

Deep learning methods also include transfer learning approaches in which word embeddings are pretrained on a broader corpus and then applied or fine-tuned in the classification. A popular neural network architecture for self-supervised training of word embeddings is BERT (bidirectional encoder representations from transformers), which Google developed to handle natural language search [64]. Recent work by Chalkidis et al. involved pre-training on a legal domain-specific LEGAL-BERT, which was found to outperform BERT in legal text classification tasks [65]. Another work used domain-specific pre-training with self-supervised learning to learn from the context of a case citation the case holding [66]. Another group recently combined Legal-BERT with TF-IDF to do legal statute retrieval [67].

One of the critical problems for applying AI to law is explainability or interpretability. Explainable AI/interpretable machine learning is a growing field of research, as it has become clear that for prediction models to be applied in practice, practitioners and consumers need to understand the basis for predictions. Interpretable models are also critical to gain theoretical insight from the classification of empirical data [19,68]. Interpretable machine learning is particularly important for legal practitioners since law is based on reasoning, and so it is essential to understand the basis for why decisions are made [69]. Interpretable models can also reveal ways in which the law is biased. More generally, for example, analysis of word embeddings in a large document corpus showed that embeddings learned a semantic correlation between vocabulary and gender that showed employment bias—for example, the term "computer program" was associated with men in [70].

Interpretability is not as much an issue for the classical machine learning and decision tree-based models described above. For example, TF-IDF can be viewed as a metric for relevance in addition to being a feature [71]. One group has proposed visualizing TF-IDF values of terms via word clouds to determine the basis for patent invalidity decisions at the Federal Circuit [72]. TF-IDF combined with SVM classification has also extracted the legal and factual basis of construction law-related verdicts [73]. XGBoost can be analyzed to identify feature importance through various scores that are based on characteristics of boosted tree models, such as the number of times a feature is used to split trees, or the gain in score towards the objective function obtained by splitting trees based on a feature [74,75]. The SHAP (SHapley Additive exPlanations) method can also be used to compute importance scores using a game theoretic approach [76].

Deep learning, by contrast, relies on neural networks which are not readily interpretable. Methods have been developed to examine neural network structure, through relevance propagation, sensitivity analysis, and decomposition methods for the network [77]. Another approach is propagating activation differences through a neural network [78]. Saliency maps have been proposed in CNN as a way to identify important regions in visual regions and then applied to regions in text [79]. Another method for interpretability is to do backtracking through a CNN to find important text [80].

While the aforementioned methods involve analyzing neural network structure, another approach is to add attention: a neural network architecture designed for sequence

prediction tasks, which can both learn and can be directly examined to identify key features in text. Classification may, for example, be done by a neural network that includes an attention layer after bidirectional-LSTM [81]. It has been shown that attention layer weights may be identified and used to highlight words using an attention encoder to summarize sentences [82]. A hierarchical architecture has also been developed employing attention with RNN to identify attention at the word, sentence, and paragraph level [60]. An important caveat is that while attention may identify features that were of particular importance when neural networks were learned in training, these may not be the most quantitatively significant features [83]. Empirical work suggests that attention can identify the actual important features in a classification task when attention is included as part of the development of the classifier, rather than serving merely as a window into the weights that are learned by the classifier [84].

In the legal domain, one work applied this method to U.S. Bureau of Veterans Appeals (BVA) decisions, but found that it was not scalable given longer text, and was very slow, reflecting the limitations of RNN architecture. It was therefore complemented by a semi-supervised approach focused on deriving text annotations that could help human interpretation of decisions [14]. Similar inaccuracy and scalability issues were found for BVA decisions when using CNN [16]. Another approach that was developed for medical coding is using attention on top of CNN to derive a per-label multilabel attention [85]. In one recent comprehensive study, interpretable machine learning using attention with RNNs, the hierarchical CNN with attention, multilabel attention, and BERT has also been demonstrated for judgment prediction in ECHR cases [86]. The authors found good classification performance. Notably, inferior performance was obtained using BERT when masking proper nouns, which suggested that BERT performed best when it overfit for the names of parties and other entities who had repeated patterns in cases.

In this paper, we employ a neural network that combines a CNN, with multiple attention layers for interpretability, followed by a dense classification layer. CNN was presented as a method to analyze text by Microsoft in an application for web search [87]. Our CNN method is based on a method for sentence classification presented in [88]. We employ CNN instead of RNN in part because CNN is much faster on modern hardware architectures, such as GPU and Google's specialized Tensor Processing Unit (TPU), as well as the method's tolerance for large text without over-consuming memory resources (a problem with transformer-based methods, which we evaluated but do not present in this work) [89]. It has also been suggested that CNNs may outperform RNNs in sequence modeling tasks even when memory is not at issue [90]. We obtain model interpretability by highlight key words using attention values, following [82]. We use hierarchical attention models inspired by [60], but rather then using a hierarchical model built on recurrent neural networks, we instead look at CNNs of different window sizes to look at long and short range effects. Our CNN and attention framework is also based in part on an interpretable attention framework for visual classification proposed by [91], as well as an approach of using CNN on each feature in a multifeature classifier developed for biomedical applications (predicting adverse drug reactions) [92].

Finally, another approach that has become very popular for explainable machine learning in text analysis is local interpretable model-agnostic explanations (LIME) [93]. LIME involves perturbing the features (e.g., words) of a sample around a point and developing a local approximation of the model sensitivity with respect to features, allowing for a ranking of features based on the extent to which they influence a classification result. The result is readily interpretable, as either a list of words with scores or text highlighted for key words. Efforts have been made to determine LIME globally over many samples, although that is difficult to do reliably and scalably [94]. LIME can also be unstable because of the randomization of LIME perturbations, although in practice that is more so for less stable models; a modified version of LIME has been proposed to address this issue [95]. Other drawbacks include that LIME is insensitive to the position of terms in the document. LIME is also not necessarily quantitatively intepretable in a meaningful way; rather, LIME

may be better as a visualization tool [96]. However, since LIME is model agnostic, it can be applied to a variety of approaches such as decision trees or neural networks.

3. Materials and Methods

3.1. Document Collection

The U.S. Patent & Trademark Office (PTO) makes the full files of post-grant review decisions available online at the PTAB Open Data website, https://developer.uspto.gov/ptab-web/#/search/decisions, accessed on 3 April 2022 (additional options are available to search for proceeding records and documents). Data are available for download via a PTAB API (https://developer.uspto.gov/api-catalog/ptab-api-v2, accessed on 3 April 2022), and there are additional bulk download options for patent text and subject matter codes. For this paper, information is accessed and documents downloaded through this API for all AIA post-grant review (CBM, IPR, and PGR) proceedings from the inception of the programs on 16 September 2012 through 31 December 2021. Manual inspection and the Docket Navigator database (http://www.docketnavigator.com, accessed on 3 April 2022) are used to identify proceedings in which an institution decision was reached, excluding proceedings in which a party joined a petition filed by another party and proceedings that terminated prior to an institution decision, which usually occurs either because the patent owner disclaimed the patent or the parties reached a settlement. We identified 11,047 proceedings that met these criteria.

The PTAB API can then be used to download petitions, preliminary response briefs, and written institution decisions using the identified proceeding numbers. While the API allows for the download of documents by type, we found that this was often unreliable, because types are recorded inconsistently in the system. For example, a preliminary response brief (POPR) should be identified as a preliminary response, but it is sometimes identified as a response, which then can create confusion with responses to procedural motions or the post-institution response brief. Similarly, institution decisions may be filed under a number of different document types. Accordingly, a variety of keywords in the "documentTypeName" and "documentTitleText" fields are used to download the correct document. The PDF files are then downloaded and converted to text using pdfplumber (https://github.com/jsvine/pdfplumber, accessed on 3 April 2022). The script used to obtain this paper's results is available at our Github, https://github.com/EESI/PTAB (accessed on 3 April 2022).

When preparing the data for this paper, not all documents could be downloaded. In many instances, this was because documents for certain proceedings were simply unavailable through the PTO's Open Data system due to technical issues and unreliability of the database. Additionally, some documents generated errors due to corrupted PDF files or errors in the PDF download and conversion to text processes that persisted even after repeated attempts. Ultimately, we downloaded documents for 10,462 cases; although, in some cases these included errors or a case may have only included one or two of the three types of documents. In sum, our dataset, which is made available on our Github, includes 9079 cases with written decisions, 10,060 cases with petitions, and 9283 cases with preliminary responses (although we have made the full data set available, petitions were not analyzed in this study).

3.2. Document Pre-Processing

This paper shows two classification tasks: outcome prediction based on the response brief, and classifying the written decision according to outcome. There are six possible outcomes: denial (no trial, i.e., the patent owner wins), denial on rehearing (after the patent owner loses they ask for the decision to be reconsidered and the judges reverse it), grant (trial instituted, i.e., petitioner wins), grant on rehearing, mixed (trial granted on some grounds or claims but not on others—this is no longer an option as of the Supreme Court's 2018 SAS decision), and indefinite. The latter "indefinite" category occurs for IPR reviews in which the judges determine that the claims are indefinite. Generally, this is grounds

for invalidating a patent, but by statute, a patent may only be invalidated on the basis of prior art in an IPR review; therefore, this counts as a "win" for the patent owner, even if it means that there is a likelihood that a court would nonetheless find their patent invalid. The analysis in this paper only considers pure denial and grant, because "mixed" outcomes are no longer possible post-2018, rehearing outcomes are too complex (the initial written decision is effectively found to have been erroneous), and indefiniteness findings are too rare to be able to have a sufficient number of training samples.

The model input is the document text. A limit of 4000 tokens is imposed. However, in part because of word count limits on briefing, and because many words are removed in pre-processing, no sample was found that had this many tokens. The text is tokenized using the Natural Language Toolkit (NLTK) Python package [97]. In our initial analysis, certain tokens were removed as follows: to avoid fitting to common words such as "there" and "only", stop words identified using the NTLK stop word list are removed. Certain abbreviations were found to be associated with specific cases or did not impart meaning. Therefore, any word with three or fewer letters was removed. Punctuation marks and words containing punctuation marks other than hyphens, as well as any words containing the "@" symbol (i.e., found in email addresses of the parties' attorneys) were removed. Finally, proper nouns were removed to take out entity names, following the suggestion of [86] as a way to avoid overfitting to entity names who are associated with particular outcomes, in particular, the names of parties who repetitively appear before PTAB, in the training set. As explained in the Results section, in the analysis of response briefs, only the CNN method provided classification results better than chance, even after extensive hyperparameter tuning. Thus, to better interpret the attention for CNN, the stop words and short words are *not* removed in the analysis of the response briefs. By leaving in stop words and short words, sentence structure can be preserved. Otherwise, text is preprocessed, and tokens are not removed as described in the foregoing.

All tokens are converted to lowercase, in order to avoid separating words that are capitalized because they begin sentences. At that point, for the TF-IDF + XGBoost model described below, the pre-processing script generates a TF-IDF vector for each document in which all terms (remaining tokens) are assigned their TF-IDF score (log of term frequency divided by document frequency) using the TFIDF_Vectorizer class in sklearn version 1.0.2 in Python 3.7 [98]. The terms and IDF scores are obtained for the training data set. For documents in the test set, terms that did not appear in the training set are ignored, and the IDF is not recalculated. For the CNN model, all tokens that appear in the training data are assigned an integer value, and each document is an ordered vector with the sequence of integers representing tokens. The sequences are padded with zeros to a length of 4000, which are subsequently masked. A value of one is used to represent any tokens in test set documents that were not found in the training data. (When reversed, the token appears as '<OOV>' to represent an out of vocabulary term).

3.3. Interpretable Machine Learning Using XGBoost Classification of Documents Based on TF-IDF Features

This paper shows two kinds of classification. One mode is classifying TF-IDF vectors of documents (prepared as described above) using XGBoost, a decision tree-based ensemble learning method [53]. Hyperparameter optimization was performed on a subset of the training data used in the classification task for decisions, in which we classify the decisions according to whether they are grants or denials of institution. After hyperparameter tuning, the following parameters were identified for subsequent studies: maximum tree depth of 15 (evaluating values from 5 to 25), the DART booster with a drop rate of 0.1 (0.1 to 0.5), skipping drop rate of 0.5 (0.1 to 0.5), L2 regularization of 2.0 (evaluating values from 0 to 2.5), L1 regularization of 0 (0 to 2.0), learning rate of 0.1 (0.0001 to 0.1), gamma of 0 (0 to 2.0), subsample rate of 0.8 (0.5 to 1.0), and a fixed level of 1200 estimators (300 to 2000; we also evaluated early stopping criteria based on training AUC). Individual samples are weighted in inverse proportion to relative class size to account for class imbalance. The same tuned

hyperparameters were used for other classification tasks. Alternative hyperparameter values were evaluated but no substantial differences in performance were observed. We use the standard predictor and GPU-optimized predictor depending on whether we are using an environment with a GPU. This paper's results were obtained using XGBoost with the Python 3.7 and xgboost 1.5.1 in the Google Colab CPU or GPU high-memory runtime environments. The feature significance was obtained from SHAP (Shapley Additive eXplanations) values of terms for the test data set using SHAP (https://shap.readthedocs.io/, accessed on 3 April 2022) in Python 3.7 [76].

3.4. Interpretable Deep Modeling with a CNN-Attention Model

The second classification method is a CNN with multiple kernel widths, pooling, and attention. Figure 2 shows a schematic of the neural network (a full schematic generated is shown in the Github, along with the layer definition using the Keras API for Tensorflow 2.0). Generally, tokens are embedded in vectors with weights trained during the classification task. The embedding vectors are then processed by a series of convolutional neural networks with different kernel widths, which can look at different ranges of context.

While, typically, the output dimensionality is reduced by maximum pooling, here the CNN outputs are fed to a self-attention layer following the structure proposed in [81], which was previously employed in a biological sequence context to do sequence attention visualization [99]. At this layer, there will be a L–W length vector, where L is the original sequence length and W is the kernel width of the respective CNN. That allows for sequence-level attention visualization of the CNN output at this point, along with providing trainable attention weights which mean that the attention will be part of model generation. As such, the attention may provide insight to what features in the text are important for classification at different ranges. The output of the attention is then concatenated and fed through two fully connected layers to obtain the ultimate classification result. We use a sigmoid activation function to generate the output as the tasks presented in this paper involve binary classification (so any output greater than or equal to 0.5 can be assigned to the positive class and otherwise the negative class); however, softmax output may be used for multiclass problems. To avoid overfitting, dropout is applied as shown in Figure 2. The attention values are visualized, as described in the Results section, to interpret the trained models, as well as, in some cases, LIME [93] using the LIME python implementation (https://github.com/marcotcr/lime, accessed on 3 April 2022).

The neural network model parameters shown in Figure 2 were selected based on evaluating hyperparameters for a subset of the training data in the decision classification task, as was done for XGBoost. The hyperparameter search evaluated values of 64, 128, 256, and 512 filters, and 6, 8, 10, 12, 14, or 16 CNNs, dropout layer values of 0, 0.1, 0.2, 0.3, and 0.4 for each of the two dropout layers, and an intermediate dense layer with a 64, 128, or 256-dimensional output. Individual samples are weighted in inverse proportion to relative class size to account for class imbalance. The model and training/testing are done using Tensorflow 2, with binary cross-entropy and mean squared error as the loss functions for classification and regression tasks, respectively. The learning rate hyperparameter was set to 1×10^{-4}, after evaluating rates of 5×10^{-5}, 1×10^{-5}, 1×10^{-3}, 5×10^{-4}, 1×10^{-2}, and 1×10^{-1}. Models were trained for 160 epochs, after evaluating epoch ranges from 50 to 200. Models were trained and evaluated in the Google Colab running Tensorflow 2.70 and Python 3.7.12 in the TPU run-time environment, which runs on Google Cloud Tensor Processor Unit (TPU) hardware.

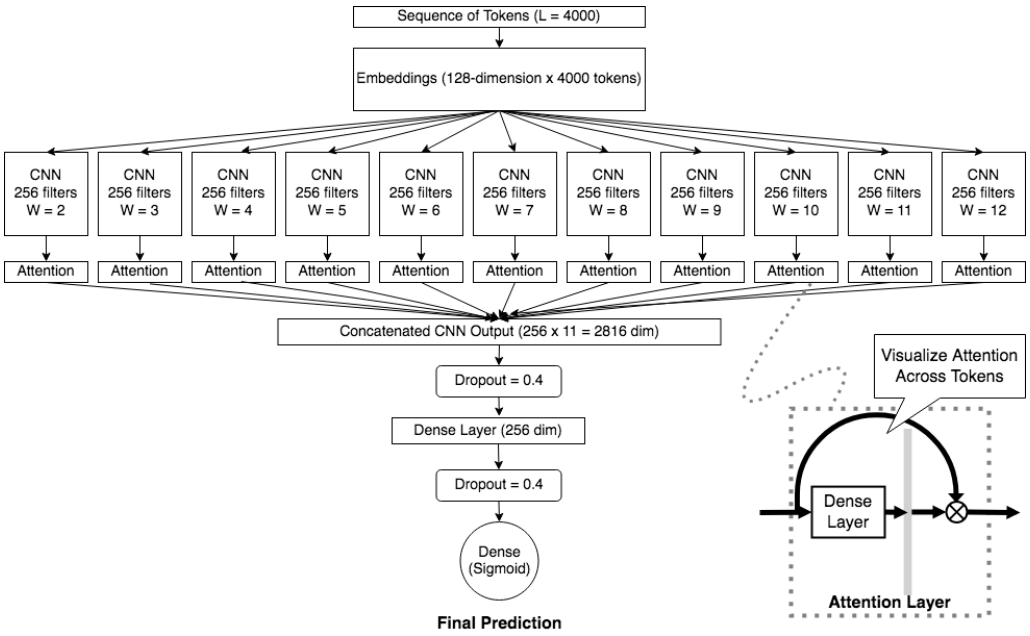


Figure 2. Schematic of CNN-Attention architecture used in this paper. The input is a sequence of L integer tokens, which is processed to a matrix of token embedding vectors with trainable embedding weights. The embeddings are fed into CNNs which vary by kernel width as shown in the diagram. The output is fed into a feed forward attention layer, which is shown in the inset at the bottom, based on the structure described in [81]. The attention at that layer will have a dimensionality equal to the sequence length minus the kernel width, allowing values across the sequence to be visualized. The outputs of the attention layer, which will have a dimensionality equal to the number of convolutional filters, are then concatenated to form an, in this case, 11 times 256 or 2816-dimensional output, which is in turn fed to a densely connected layer with a 256-dimensional output, and then a final densely connected layer with sigmoid activation function to make a binary classification.

4. Results

4.1. Determining Institution Outcome from Text of Written Decision

4.1.1. Comparing Performance of XGBoost and CNN Methods

To determine the viability of text classification for PTAB documents, and validate our approaches to interpretable machine learning, this paper starts with the classification of written institution decisions (institution decisions) issued by PTAB judges. As explained in the Background section, at the institution stage, PTAB judges determine whether a petition for post-grant review should proceed to trial. Their decision is announced in a written opinion that states the decision and explains the factual and legal bases for their decision as well as their reasoning. The decisions do not follow a fixed pattern or structure, and the judges may use different kinds of language to announce the decision; therefore, a form-based search cannot reliably derive the outcome from the text of the written decision. We pursue the task of using interpretable machine learning to classify written document text, and use that classifier to predict whether a decision is a grant or denial of institution.

One important consideration is that a patent challenger may file multiple petitions against the same patent, and, because patent owners may have filed a lawsuit against multiple defendants, different patent challengers may also challenge the same patent. Because the petitions may expose similar weaknesses in the patent, the result of the petitions may be correlated. Recently, PTAB judges have begun to discourage multiple petitions, for example by asking challengers to rank their petitions or rejecting duplicative petitions [100]. Notably, this is another example of how the problem changes over time in litigation. In any event, to avoid the complexities of redundant petitions, we eschew cross-validation and instead evaluate classification methods on a time-separated data set. The balance between grant and denial classes, however, also changes over time.

Figure 3 shows how the total number of institution decisions, and the balance between grant and denial, have changed over time. Notably, the low total number for 2012 is due to petitions not filed until the AIA post-grant review system went into effect in September. The low number for 2021 reflects the typical six month lag between petitions and institution decisions, since only petitions through June 30 will have been guaranteed to have an institution decision by data collection cutoff. Over this time, the grant/denial balance has fluctuated. For example, 60.5% of petitions in the data set that were filed in 2018 were granted, a fraction which fell to 58.7% in 2019, and rose back to 61.1% in 2020. This is unsurprising, as variations in the grant/denial ratio may be due to changes in how PTAB regards duplicative petitions, as well as changes in the quality of patents being asserted in litigation and parties' refinement of litigation strategies. Consequently, we use time-separated multi-year data sets for both training and testing. The training data are made up of written institution decisions of grant or denial (i.e., excluding mixed decisions) on petitions filed through 31 December 2017 (which generally implicates decisions through 30 June 2018, since the decision is due 6 months after the petition is filed). The test data are made up of institution decisions on petitions filed 1 January 2019 and published up to the data collection cutoff of 31 December 2021.

Moreover, as Figure 3 shows, up to 2018, a substantial fraction of institution decisions had mixed outcomes (some claims and grounds instituted, others now). As explained in the Background section, in April 2018 the U.S. Supreme Court ruled that mixed institution outcomes are unlawful. As the graph suggests, in the aftermath of the decision, the grant rate increased for 2018 petitions, as PTAB judges were obligated to grant a petition in full even if only a single ground or claim was deemed likely to succeed. Over time, this normalizes somewhat, corresponding to PTAB's increased rejection of petitions on the basis of being duplicative.

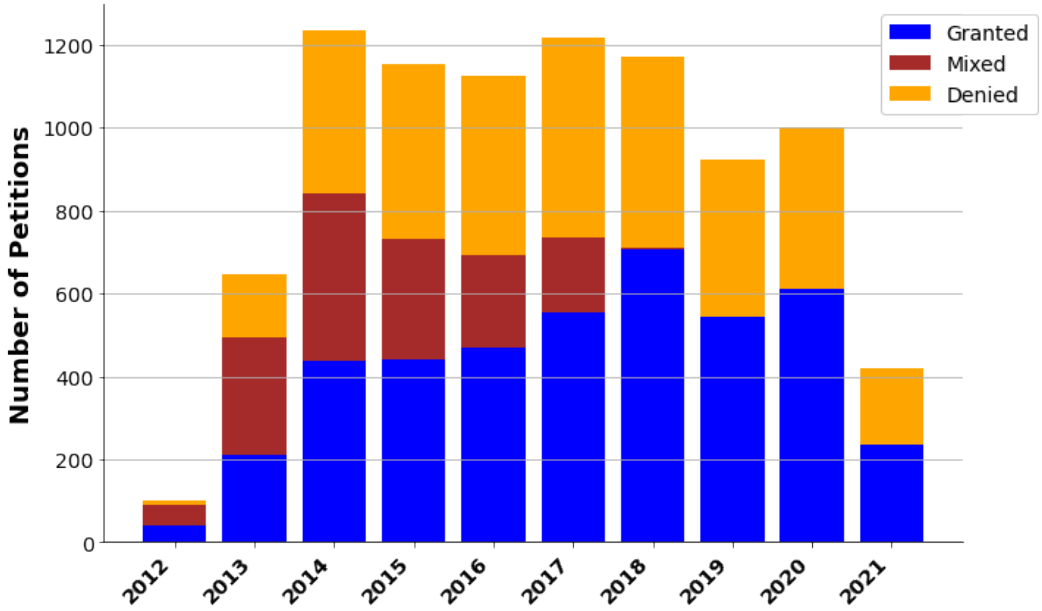


Figure 3. Institution decisions by petition filing year in the data set (combining IPR, PGR, and CBM as we do for the analysis in this paper) from the AIA's effective date in September 2012 through the data set cutoff of 31 December 2021, showing grant (blue, bottom), denial (orange, top), and mixed (brown, middle) outcomes. Notably, in 2018, the U.S. Supreme Court's *SAS* decision eliminating the possibility of a mixed outcome.

Figure 4 compares the performance of the two classification methods used for this paper, as detailed in Sections 3.3 and 3.4: (1) classifying documents represented as vectors of TF-IDF for all words in the vocabulary (after removing short words and stop words as described in Section 3.2), and (2) classifying documents represented as vectors of integer

tokens using the CNN model with attention shown in Figure 2. As a baseline, logistic regression was performed using sklearn in Python 3.7 [98] with a limited vocabulary. As Figure 4 shows, the most accurate and best method at avoiding type I and type II errors (i.e., with good specificity and sensitivity) is XGBoost with TF-IDF document vectorization. Logistic regression underperforms both XGBoost and the CNN-Attention model. Interestingly, the best performance was found for logistic regression when limiting the vocabulary size to the 2000 most frequent words. As the vocabulary size increases, logistic regression performance deteriorates, suggesting potential overfitting. The CNN-Attention method's underperfomance appears to be largely due to problems with identifying denials. This is not necessarily because denials were underrepresented in the training data. The training data set was 53% grants, whereas the test data set was 59% grants. As discussed above, this difference is because mixed outcomes were excluded from the training set, and the test set is entirely from after when mixed outcomes were disallowed. However, the time frames for the training and test sets are also not connected to the CNN-Attention model's underperformance on denials. The model run for training set petitions filed between 1 June 2018 and 31 December 2019, i.e., all after mixed outcomes were eliminated, and the training set petitions filed after 1 April 2020, provided similar results: the average recall for the granted class was 0.92 and the average recall for denial class was 0.62.

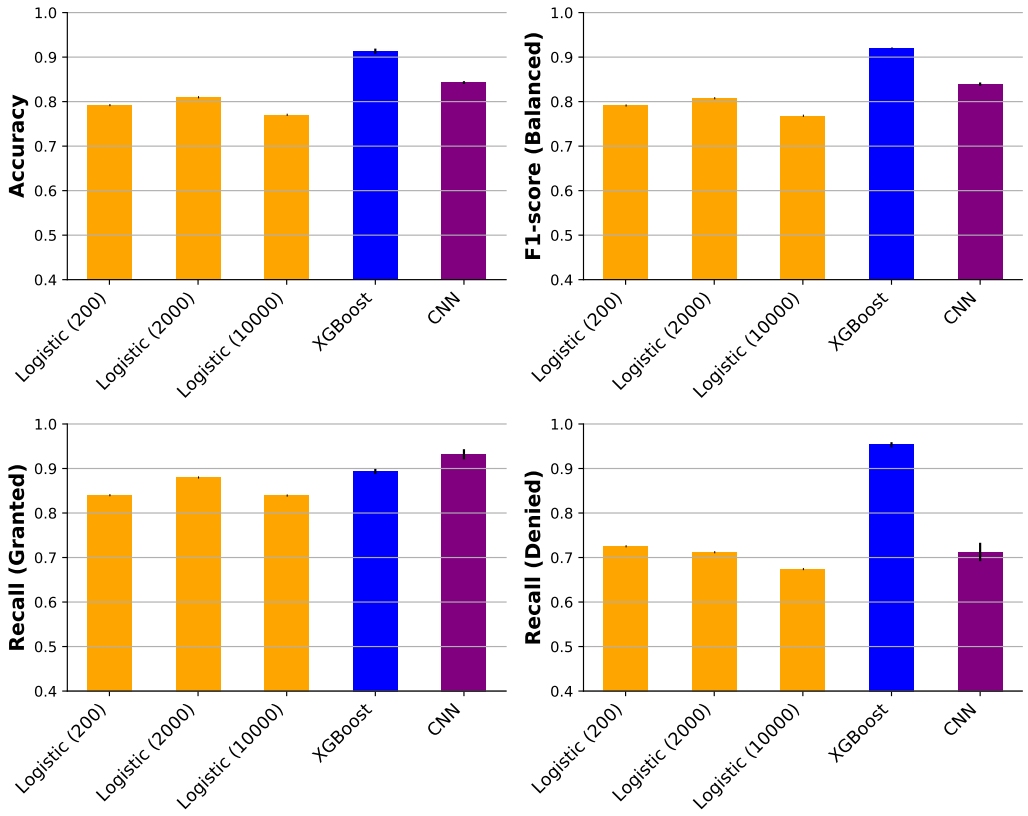


Figure 4. Results of predicting grant or denial in written decisions on petitions filed after 1 January 2019, for the model trained on written decisions for petitions filed up to 31 December 2017. The metrics are averaged over 5 runs for CNN-Attention and 3 runs for XGBoost (error bars indicate standard deviation). The graph at the upper left shows accuracy, calculated by dividing the number of correct predictions by the test set size. The graph at the upper right shows the sensitivity and specificity using the balanced *F1*-score, which is the harmonic mean of precision (true positives divided by all positive predictions) and recall (true positive rate, i.e., sensitivity), accounting for the imbalance between positive and negative classes. To illustrate the impact of the class imbalance rate, the graphs at the bottom left and right show the recall for the granted and denied classes separately. Which indicate that the CNN model is less sensitive to denials than XGBoost.

4.1.2. Interpreting XGBoost and CNN Predictions for Written Decisions

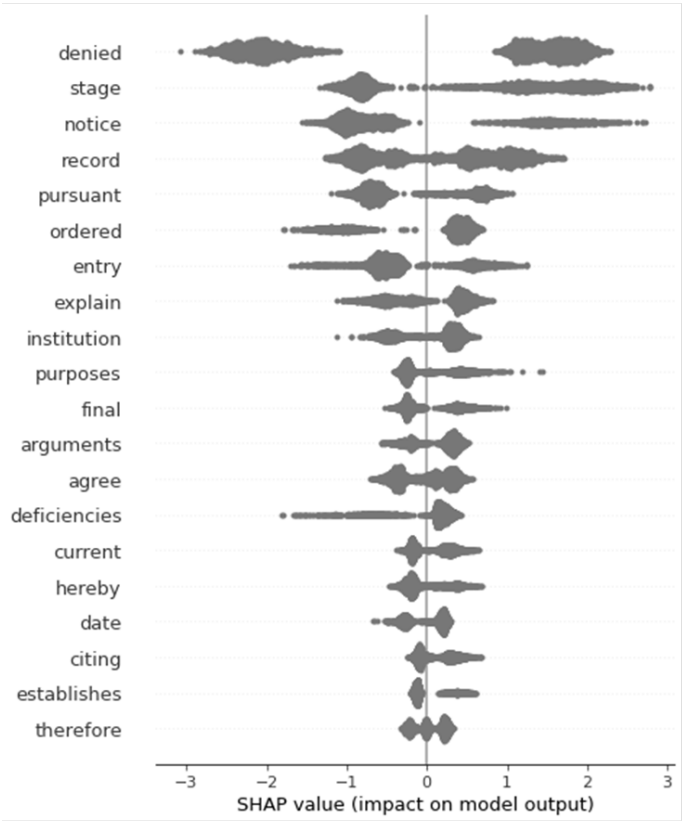
After classifying written decisions using both XGBoost/TF-IDF and CNN-Attention models, the trained models should be interpreted to obtain insight on what features are relevant to understanding the text of the decision. In particular, we can ask what features the trained model identified as important, or in the case of the deep model, attended to, in making its predictions on the test set.

As an initial matter, we evaluate pretrained word embeddings in addition to the TF-IDF and end-to-end deep learning methods shown in this paper. Two different sets of pretrained embeddings were used to transform written decisions in the PTAB dataset: GloVe [46] word embeddings pretrained on a corpus of 2014 Wikipedia entries and Gigaword 5 compendium of English-language news articles (obtained from https://nlp.stanford.edu/projects/glove/, accessed on 3 April 2022), as well as Law2Vec, which is pretrained like Word2Vec but on a corpus of international legislation, U.S. statutes, and U.S. Supreme Court decisions [101]. In both cases, pretrained embeddings miss words that are particularly relevant for our more specialized domain which includes both legal terms specific to U.S. patent law as well as the technical content of patents. For example, among frequently occurring words that have legal significance in patent law, Law2Vec missed: embodiments, inherency, unobvious, indefiniteness, and lexicographer. GloVe missed the latter as well as: unpatentable, unpatentability, nonobviousness, preambles, indefiniteness, nonobvious, reexaminations, and patentably. Among the frequently occurring technical terms, by way of example, Law2Vec missed: microfluidic, rotatable, subcarriers, playlist, oligonucleotide, applet, cached, and encrypted. GloVe missed relatively few common technical terms; however, it did miss certain technical terms that are often used in the patent context, such as "operably" and ''removably", while still missing certain technical terms such as "photosensor".

To be able to properly interpret documents, therefore, we utilize the document vocabulary (excluding proper nouns, stop words, and other short words) to generate TF-IDF vectors for XGBoost classification and learn embeddings through end-to-end training of the CNN-Attention model. Figure 5 shows the highly ranking SHAP values across the testing data set [76], as well as the highly ranked words based on the "gain", i.e., gain in score towards the objective function obtained by splitting trees based on that feature [74,75]. The highest rank words are similar in the two lists. Unsurprisingly, among the most significant words is "denied". That said, because common stop words and short words are excluded (as described in Section 3.2, words of negation such as "not" or "no" do not contribute to the classification. The classifier must instead make its decision based on longer words. Another significant word, for example, is "stage". This is consistent with our intuition that when the PTAB judges are deciding to make a decision to proceed to trial, they will likely note that their findings of a reasonable likelihood of success in the patent are only relevant to this initial stage of the proceeding. A similar intuition holds for the word "purposes", which is associated with judges making a preliminary factual or legal determination for the purposes of an initial determination, which they may revise after trial when they make their final decision in the case. Notably, for shorter lengths of text, or smaller vocabularies, it would be possible to look at combinations of words that appear together, such as bigrams, trigrams, n-grams. However, combinatorial explosion given the vocabulary size and document length makes n-grams untenable for analyzing full-length written decisions and briefs.

To analyze words in context, the attention vector of the CNN-attention model (see Figure 2) can be used to highlight features across the sequence of tokens [60]. As an example of a granted petition, Figure 6 illustrates case number IPR2020-00806, which was initiated by a petition filed by Comcast Cable Communications LLC on 17 April 2020 to invalidate Rovi Guides Inc.'s U.S. patent number 8,001,564, which covers an aspect of electronic television programming guides. Figure 6 shows an excerpt of the decision text, highlighting tokens according to the relative magnitude of attention at corresponding positions in the attention vector. Figure 6 shows that attention can highlight important arguments in the text, for ex-

ample, in this case, whether the petition was redundant of other cases and examination. The attention highlights references to "previously presented" asserted grounds, and the ability of the judges to use their discretion to deny the petitions on that basis.



Word	Gain Score
record	161
stage	145
denied	134
pursuant	119
explain	104
purposes	81
institution	80
notice	79
agree	77
hereby	74
discloses	68
arguments	65
deficiencies	63
entry	61
final	61
ordered	60
prior	59
therefore	57
used	57
date	52

Figure 5. Summary plot of the highest ranking words obtained using SHAP (**left**) and the highest ranking words identified by the gain score metric in XGBoost (**right**) in the test set for written decision classification. The SHAP plot shows clouds of points representing individual documents; a positive SHAP value indicates a contribution to a positive classification, i.e., denial, and a negative value a contribution to a classification as a grant.

To compare attention highlighting to other methods of identifying important features, we compare the ranking of the top 20 mean attention paid to particular words (across the positions at which they are found for a kernel width of 8, as shown in Figure 6) to the words with the top 20 words by SHAP value found by XGBoost classification, for the exemplary decision in IPR2020-00806. We further compare the results of using the LIME method, which identifies the features that affect classification the most by perturbing features around a particular example [93], on the IPR2020-00806 written decision. Table 1 shows the ranked lists side by side, along with the top 20 words by TF-IDF for the example. SHAP and attention values agree well with each other, which validates the attention values as a measure of the significance of word features. As the table shows, TF-IDF terms, by contrast, are unique to petitions on patents on distinct technological inventions. By contrast, both XGBoost and CNN classifiers tend to focus on terms more connected to legal and procedural aspects of the decision, such as "notice", "stage", and "ordered". The LIME classification, interestingly, includes both legal terms found by attention, and technical terms specific to the patent at issue in the case, such as "television", which was also a term with a very high TF-IDF in this document. This suggests that LIME's methodology of

finding local perturbations will tend to identify specific words to the document. Unlike LIME, attention values more consistently reflect terms that are more generally applicable across the category of documents.

Figure 6. Text of the written decision granting trial in IPR2020-00806, with words highlighted according to the value of the attention vector at their positions (zero-padding edges to account for the kernel width of the CNN). The intensity of green corresponds to attention above the median for the vector as a whole (excluding the attention at zero tokens where the text was padded), and the intensity of pink corresponds to the magnitude of attention below the median. At **top**, we show the attention for a kernel width of 2, and at **bottom**, the attention for a kernel width of 8, which captures longer range context.

Figure 7 shows an attention-highlighted excerpt for an exemplary written decision to deny institution. In the illustrated case, the highlighted terms in the written decision indicate that the opinion is based on using the fact that the original prosecution considered the same basis for invalidation as the petition in this case should be a basis for denial. The classifier pays less attention to the text at the beginning of the excerpt in Figure 7, which relates to the technical content of the decision. In the middle of the excerpt, there is another low attention stretch, which corresponds to the specific details of how the petition in this case duplicated original prosecution of the patent at issue. Attention visualization

therefore indicates that the deep learning is making its prediction based on the statements that the judge who authored the decision makes about legal conclusions.

each message contains parameters including mandatory receiver parameter identifies intended recipient message directed multimedia information control system code discloses information control systems automobile facilitate user retrieval dissemination information control vehicle particular discloses display interface user able obtain information control selectable functions automobile instrument panel navigation function directed system method providing menu data using communication code discloses communication system includes units network switching center service centers provide variety enhanced services code discloses user interface enables access global computing network institution inter partes review discretionary mandatory agency decision deny petition matter committed permitted never compelled institute argues exercise discretion deny institution relies upon prior overcome prosecution patent address directly however contends patent reference matter would issued combined references examiner render claims evaluating whether deny institution basis substantially prior arguments previously presented considers following factors similarities material differences asserted prior involved examination cumulative nature asserted prior evaluated examination extent asserted evaluated examination including whether prior basis rejection extent overlap arguments made examination manner relies prior distinguishes prior whether pointed sufficiently erred evaluation asserted prior extent additional evidence facts presented petition warrant reconsideration prior arguments designated precedential first paragraph begin brief overview relevant prosecution history analyze factors view parties arguments evidence relevant patent issued application application issued parent patent code during prosecution parent patent application rejected claims obviousness combinations various publications inventors after interview applicants amended independent claims issued advisory concerning newly discovered describing another communication language following another interview applicants agreed distinguish claims applicants amended independent claims allowed claims indicating cited references teach newly added limitations claims pending patent application also subject rejection amended include limitations similar added claims parent patent application after rejection amendment claims allowed factor similarities during reflected listing challenges asserted three challenges based combination discussed rejected claims pending parent patent application based part version effect patent application filed stated examiner consider information considered parent application examining list information need submitted thus determine constitutes prior involved examination patent application contends examiner reference however asserts contention incorrect involved examination particular points second page patent lists portion specifically cited first listing addressing seven parts specification argues made publicly available single explains filed seven exhibits size numbering accordingly present record agree representation find constitutes single document additionally find therefore involved examination patent also relies upon either combinations teach limitations several dependent claims neither directly addresses whether either references involved examination patent although record appear less emphasis additional references relied upon heavily relies upon references solely additional limitations several dependent claims largely focuses emphasizes alleged importance contend teachings additional references materially different teachings references argue citation limited solely accordingly find similarities lack material differences asserted prior involved examination weigh favor exercising discretion deny institution factor cumulative nature during reasons discussed respect first factor cumulative nature asserted prior evaluated examination also weighs favor exercising discretion deny institution factor extent during including discussed evaluated examination parent patent application relied upon rejecting claims accordance also would considered examination patent application additionally explained specifically cited listed second page patent although base rejection upon citation reference reflects evaluated reference further also explained appear present record evaluated examination patent application reasons explained context evaluation first factor less weight additional references accordingly find extent asserted evaluated examination including based rejection upon examination parent patent application weighs favor exercising discretion deny institution factor extent during relies this factor appears weigh slightly favor exercising discretion deny institution presents arguments sufficiently different made examination factor pointed evaluation discussed address directly explain examiner erred evaluating contrary takes position record discussed accordingly factor weighs strongly favor exercising discretion deny institution factor extent allegedly additional evidence directs attention reasons explained however disagree adequately explained constitutes additional evidence listed cited second page patent weighing factors considering factors whole find factors weigh favor exercising discretion deny institution based record find five factors weigh either favor strongly favor exercising discretion additionally particular significance overemphasis alleged lack examination record reflects reference fact accordingly light consideration factors record whole exercise discretion deny institution foregoing reasons exercise discretion deny institution substantially prior previously presented consideration foregoing hereby denied challenged claims patent inter partes review instituted

Figure 7. Attention visualization as described in Figure 6 for an example of a written decision denying institution in case number IPR2019-00839, filed 20 March 2019 by Microsoft Corporation against IPA Technologies Inc. Green-labeled words are locations where the attention (for the CNN module with kernel width 8) is greater the median for the sample (excluding the attention at zero tokens where the text was padded), with more intense green corresponding to higher levels of attention. Pink-labeled words are locations where attention is below the sample of the median, and more intense pink corresponds to lower attention at that location in the text.

Table 1. Top 20 words by mean attention (obtained from trained CNN classifier), LIME (for the CNN classifier), SHAP value (after XGBoost classification), and TF-IDF score (i.e., no classification) for the written decision in IPR2020-00806.

CNN-Attention	LIME (CNN)	SHAP (XGBoost)	TF-IDF
notice	record	notice	program
hereby	stage	stage	digitally
commence	proceeding	record	television
entry	patentability	entry	interactive
pursuant	citing	pursuant	programs
stage	interactive	institution	guide
final	teaches	ordered	directory
made	hereby	explain	associated
shall	television	hereby	claims
given	pursuant	final	presents
impact	recites	disagree	maintaining
different	notice	arguments	data
ground	using	citing	recordings
exists	demonstrates	therefore	constitutional
record	dispute	date	stored
institution	discussed	combination	previously
material	arguments	petition	teaches
patentability	entry	authorized	erred
disclosure	associated	must	user
identified	request	recites	storing

4.2. Predicting Institution Outcome Based on Patent Owner's Preliminary Response Brief

In post-grant review litigation before PTAB, the goal of the patent owner's preliminary response brief is to avoid an institution of trial. At the initial stage, factual inferences are considered in the petitioner's favor, and the petitioner need only demonstrate a reasonable likelihood of success. Therefore, it is the patent owner's goal to show that there is a gap in the petitioner's logic, or that institution should be denied at the discretion of the judges, for example because the petition is redundant with arguments that were already made during prosecution before the patent was issued, duplicative of other petitions, or redundant to other parallel proceedings [24,26]. To be sure, a patent owner may choose to waive their response, in which case, usually, although not always, trial is instituted; however, we do not consider those cases [22]. This paper focuses on developing a predictive and interpretable model that can predict a preliminary response brief's probability of winning based on the text and which, based on interpreting successful predictions, can provide insight as to what arguments and phrases may be connected to success.

As explained in Section 4.1.1, the training data set must be constrained to a time when there is stability in the law and PTAB jurisprudence. For example, the training data set should exclude proceedings decided before April 2018, when a mixed outcome was possible. As another example, in 2019 PTAB also established legal precedents based on the *NHK* and *Fintiv* cases that gave patent owners a clear roadmap to winning a denial of institution based on redundancy with parallel proceedings [26]. Accordingly, the training data set includes patent owners' preliminary response briefs in cases filed between 1 July 2018 and 30 November 2020. This provides a maximal number of training documents (2631 cases in total) to be sufficiently generalizable, while ensuring that it includes briefs which use contemporary legal arguments.

Figure 8 shows the results of predicting the case outcome based on the text of the preliminary response brief, using both the XGBoost/TF-IDF and CNN-Attention (Figure 2) methods. Two test sets given limited data in a stable time period (one that may have temporal correlation with the training since it starts the day after the training period and one with one month separation): (1) All cases filed between 1 December 2020 and 31 March 2021. (2) All cases filed between 1 January 2021 and 31 March 2021, to ensure complete temporal separation from the training data set. After 1 April 2021, prediction accuracy drops, thus

limited to this 4 month period for testing. This time-dependent drop in prediction accuracy has been observed in other legal prediction work. As Figure 8 shows, when stop words and pre-processing text are removed, as was done for the written decisions, neither the XGBoost nor CNN method could robustly predict both classes (grant and denial) better than chance. That is, the models perform better on the majority granted class than the minority denial class. The training data are biased towards grants of institution: 56.6% of training cases are grants, and the other 43.4% are denials. The CNN-Attention model produced better performance in a few training runs. Accordingly, subsequent studies described here focus on the CNN-Attention model.

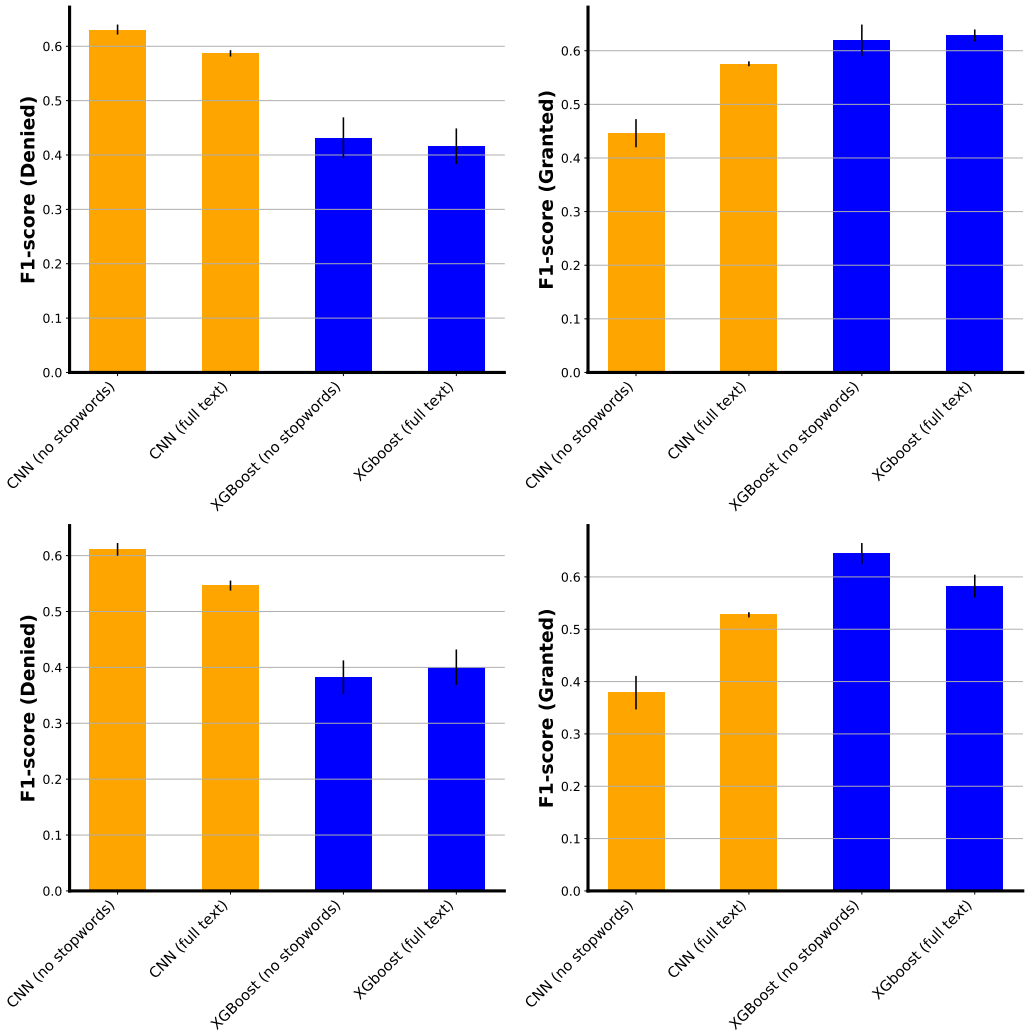


Figure 8. Metrics for XGBoost/TF-IDF and CNN-Attention models for predicting the institution outcome (grant/denial) using the full text of the patent owner's preliminary response brief. The plots show averages with standard deviations as error bars for five (CNN) and three (XGBoost) training runs. The training set consists of 2631 cases filed between 1 July 2018 and 30 November 2020. Results are shown for no maximum vocabulary size but rejecting stop words (labeled "no stop words" as described in Section 3.2, and limiting the vocabulary to 20,000 words but including stop words ("full text"). The *F1*-score (harmonic mean between precision and recall) is shown for the grant and denial classes for: (**top**) cases filed between 1 December 2020 and 31 March 2021, and (**bottom**) cases filed between 1 January 2021 and 31 March 2021 (see accompanying text for further details).

Because CNN-Attention takes into account word context, unlike the TF-IDF model, stop words (common words, e.g., articles such as "the" and "an") are included in the vocabulary. The vocabulary is capped at a maximum size of 20,000 words. Any other words are replaced with an out-of-vocabulary token ("<OOV>"). The maximum tokenized document

length is increased to 8000 to avoid cutoffs since inclusion of stop works will lengthen the integer-encoded document vectors used for the CNN model. Figure 8 shows that the CNN-Attention model can achieve adequate performance with the CNN-Attention model. However, the XGBoost/TF-IDF model still fails to provide adequate prediction performance for both classes. Further hyperparameter optimization for XGBoost was unavailing, therefore, it appears that Figure 8 essentially represents the best case scenario for XGBoost performance. By contrast, the CNN model is able to perform better than chance on both the granted and denial cases, with an everage of 57.8% accuracy on denial and 57.5% accuracy on grant outcomes.

The CNN predictions based on the response brief may be examined by visualizing the attention layers as discussed above. Figure 9 shows selected examples of cases in the test data set which were correctly predicted. Figure 9a shows attention most highlighted (colored most intensely green) where the patent owner argues that the petition "substantially overlaps" with another preliminary response (which suggests that they are not raising new arguments) and a technical argument about a "computer". The model is correctly classifying the example as a granted institution, i.e., defeat for the patent owner. The highlighting of this argument suggests that rather than making an argument about an *identical* redundant case, if an argument is about merely "substantially" identical claims, the argument will tend to lose. Figure 9b shows the highest attention level (most intense green color) for arguments in which the patent owner touts the worldwide attention that its technology received. Since the patent owner's brief in example (b) was correctly predicted as being granted, it suggests that arguments based on the adoption of the technology by "world leaders" may not be persuasive. Figure 9c shows excerpts from a brief that the model correctly predicted as a win for the patent owner (institution denial). In example (c), the highest-attention (most green) arguments concern the duplicative nature of the proceedings, which has recently favored denial as explained above. This contrasts with example (a), in which the patent owner argued that the parallel proceedings merely involved "substantially identical" claims. Notably, much less attention is paid to the patent owner's argument about specific delays on the part of the petitioner, perhaps because the specific dates at issue are idiosyncratic to this brief alone and thus ignored by the model. Figure 9d shows the same kind of pattern of high/low-attention as in (c) for a brief filed in a completely different case, which the model similarly predicted correctly as a patent owner win (institutional denial) in a brief filed in a completely different case.

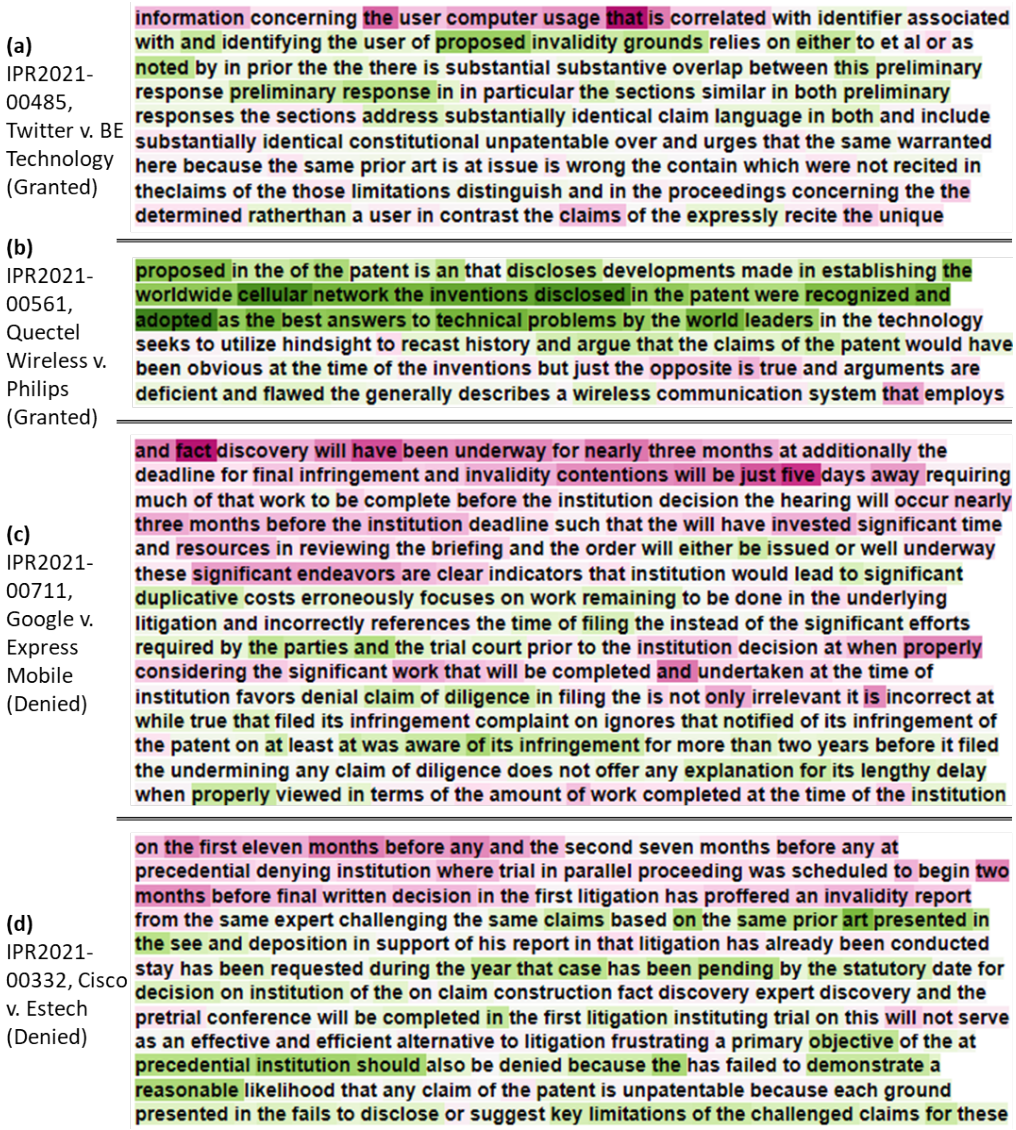


Figure 9. Excerpts of response brief predictions highlighting attention for a kernel size of 8. Green-labeled words are locations where the attention vectors have values greater the median for the sample (excluding the attention at zero tokens where the text was padded), with more intense green corresponding to higher levels of attention. Pink-labeled words are locations where attention is below the sample of the median, and more intense pink corresponds to lower attention at that location in the text. The excerpts are labeled with the case (proceeding) number and whether the result was grant or denial.

5. Discussion

This paper's goal is to demonstrate a proof-of-concept of machine learning techniques applied to a realistic and important problem in litigation analysis, post-grant patent review cases. The results demonstrate that applying XGBoost classification to a TF-IDF document representation can successfully predict whether a written decision represents a grant (success for the patent challenger) or denial (success for the patent owner) in the first phase of the patent review challenge. Then, XGBoost scores or SHAP scores calculated using the XGBoost model can rank terms based on their significance to the classification. The results further show that an interpretable deep neural network learning method based on multiple CNN and attention (using an architecture shown in Figure 2) can be successful as well. On the written decision classification problem, CNN performs less well than XGBoost/TF-IDF (e.g., 82% versus 90% accuracy, respectively), largely

due to underperformance on denial decisions. The predictions of the CNN model are also interpretable using the LIME method, as demonstrated by validating the key terms identified by LIME. The results also validate the use of the attention layer of the CNN model which can also be visualized for a document by highlighting key terms within their context in the document.

Predicting case outcome from the patent owner's preliminary response brief proves to be much more challenging. This is not unexpected. The written decision classification is not trivial—the documents have no formal structure and there is no consistent way in which judges state the decision—but the document itself discloses the outcome. The response brief will be much more variable because the patent owner must react to the issues the patent challenger brings up, and can make different strategic decisions about what arguments to pursue. XGBoost/TF-IDF was unable to predict outcomes based on the response brief. However, prediction accuracy of the CNN model proved better than chance (approximately 57%) for both the majority and minority classes in our problem domain (granted and denial, respectively). This paper thus establishes at least a preliminary proof-of-concept that complex deep network models can predict success based on the text of a litigation brief. Moreover, the results demonstrate that the attention layer visualization of the deep learning model is able to provide a highlighting of key terms and arguments that are important for determining whether the brief led to a grant or denial. The inherently interpretable deep learning architecture introduced in this paper has an important efficiency advantage over post-hoc analysis methods such as LIME or other sensitivity analysis methods. For LIME or sensitivity analysis, the model must be evaluated multiple times to generate perturbations [93], whereas attention is generated by evaluation the model just once.

There are important caveats to our analysis. Three major caveats are summarized in turn below. First, as shown in Figure 3, there are changes in grant/denial propensity over time. However, while those fluctuations are not necessarily large, there are significant differences in how PTAB adjudicates cases over time. Changes in governing law can be significant; for example, as shown in Figure 3, in 2018 the law changed such that a "mixed" outcome was no longer possible. As discussed in Section 4.2, for example, in 2019, PTAB began denying institution on new grounds of redundancy with trial court proceedings on the same patents. PTAB also began denying institutions on more parallel proceedings brought on the same patent than they did before [26]. This basis for denial proved critical in the success of response briefs after this change, as shown in Figure 9, in which key terms relating to arguments based on redundant proceedings are highlighted. Accordingly, predictive methods in litigation must be dynamic to account for changes in jurisprudence as well as governing law.

Second, training data sets are still small, which means that there is significant risk of overfitting and inability to generalize. The limited size of training data sets is inherent to the problem domain: only so many PTAB cases have been filed since the proceedings began in 2018. Future directions for addressing the limitation of training data set sizes may include looking to trial court cases in patent litigation where patent validity issues like those brought before PTAB are at issue. The shortcoming of this approach is that some issues, such as the aforementioned "redundancy argument", do not apply in district court.

Third, repeat litigants may skew results. The pre-processing method described in Section 3.2 endeavors to exclude entity-identifying information, such as proper nouns and the email addresses of attorneys. However, some entities are responsible for a large amount of patent litigation and correspondingly large numbers of PTAB proceedings. Among these are patent assertion entities (PAEs), perjoratively called "patent trolls" who sue many defendants, often over a long period of time and over different patents covering sometimes related or sometimes divergent subject matter [18]. Additionally, repeat litigants may be advancing different quality levels of patents. Some PAEs may be suing defendants on weak patents, seeking settlements—but others may be selective about asserting only strong patents to ensure durable returns [102]. In this study, we confirmed that the training

and test sets did not overlap in terms of parties; however, the presence of repeat litigants can bias training or skew test results. In particular, because training data sets are not large (the total data set only has over 11,000 examples at this point), repeat litigants risk systematically skewing results if they, for example, have similar patterns in their briefs. Future work is necessary to detect consistent patterns shown by repeat litigants, which may be helpful in removing bias from training sets or even learning patterns consistently associated with success or failure by such litigants. For example, deep learning may be able to identify a "signature" associated with PAE's litigation strategy or brief writing.

Another challenge that all real-world NLP must address is significant noise in the documents. The data set that we use is of PDF documents, which are not necessarily readily converted to clean text files. As a result, certain words can be distorted: in one example, the word "the" was replaced by "ttthhheee". Noisy words cause a problem for a classifier, because meaningful words are lost, and the training process can be skewed by overfitting to them. However, such noise is inherent to realistic litigation classifiers, at least in the United States context. In the United States, litigation documents are stored in PDF format and there is no requirement that they be machine readable noise-free. As of early 2022, court decisions and efforts in the U.S. Congress are at work are seeking to make all court documents freely accessible to the public (currently they are available only for a charge)—but these will all be in PDF format. Another challenge with these documents is that they include email addresses and other identifiable information that must be cleaned up to avoid improperly influencing classification.

One way to deal with the aforementioned document noise issue, and potentially improve prediction performance, is to use pretrained word embeddings which learn semantic relationships between words on a much larger document corpus through self-supervised training. As further discussed in Section 2.2, GloVe trained on general documents, as well as Word2Vec and BERT models trained on legal documents, have been developed and proven useful on other problems [46,65] or a version trained on legal corpus. However, as shown in Section 4.1.2, existing pretrained embeddings were missing important technical and legal terms necessary to analyze patent litigation. Future work in this area should focus on developing a broader pretrained embedding that can reflect both the technical and specialized legal content required for patent litigation, which is distinct from other forms of litigation as well as patent drafting or non-legal documents.

Finally, as machine learning and artificial intelligence methods are applied to the legal field, the academic and practitioner communities must attend to ethical issues. Because lawyers have a personal duty to their clients, offloading tasks to a computer risks breaching that duty [2,103]. Moreover, from the onset of computational methods for analyzing and predicting litigation outcomes, scholars have questioned whether computers can compete with a lawyer's professional skills and experience [104]. It is unclear how similar computational reasoning models may ever be to human legal reasoning for legal analysis [105]. Explainable and interpretable machine learning and deep models are crucial for practical applications to litigation. This is particularly so where the analysis of litigation outcomes is used to consult clients, given the need for openness and transparency [106].

6. Conclusions

This paper shows a proof-of-concept for predicting litigation outcomes from party briefs with interpretable machine learning methods that can identify what terms and parts of briefs were relevant to outcome prediction. The test case in this paper is a real-world, high-stakes litigation problem involving United States patent law: predicting the outcome of post-grant inter partes review (IPR) trial institutions for invalidating patents. Using outcome prediction based on the unstructured written institution decisions as validation, both XGBoost/TF-IDF (decision tree classification using term frequency-based document representation) and deep learning of text in context (with a novel multiscale CNN-Attention model architecture) are able to classify decisions according to outcome. Both XGBoost/TF-IDF and the multiscale CNN-Attention architecture are also able to

provide interpretable results that can accurately identify key terms relevant to the outcome prediction. At a proof-of-concept level, we show that end-to-end deep learning with the multiscale CNN-Attention deep learning architecture can also predict the outcome of PTAB institution decisions, with performance better than chance. Visualizing the attention layers of the trained deep learning models reveals key winning arguments, such as those based on the redundancy of PTAB proceedings with trial court cases on the same patents. Notably, these results come with important caveats that need to be addressed in future work: changes in jurisprudence and governing law over time limit the generalizability of predictions based on past cases to future cases, training data sets are limited in size, repeat patent litigants may skew results, and even after pre-processing, document noise continues to limit machine learning performance.

Subject to the foregoing caveats, however, the results in this paper reinforce the promise of using machine learning to analyze litigant briefs and potentially identify likely successful strategies and textual patterns in briefs. Moreover, while the proceedings analyzed in this paper occur in an administrative body, the U.S. Patent Trial & Appeal Board (PTAB), they are otherwise identical to adversarial proceedings in courts.

Author Contributions: Conceptualization, B.A.S., G.L.R.; methodology, B.A.S.; software, B.A.S.; validation, B.A.S.; formal analysis, B.A.S., G.L.R.; investigation, B.A.S.; resources, G.L.R.; data curation, B.A.S.; writing—original draft preparation, B.A.S.; writing—review and editing, G.L.R.; visualization, B.A.S.; supervision, G.L.R.; project administration, G.L.R.; funding acquisition, G.L.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by United States National Science Foundation (NSF) grants 1936791, 1919691 and 2107108.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data obtained from the United States Patent and Trademark Office for this study, in particular the briefs and written decisions, as well as other metadata for patent review proceedings and their encoded outcomes, are provided at the paper's GitHub, located at https://github.com/EESI/PTAB, accessed on 3 April 2022. Data were downloaded from the PTAB Open Data site, located at https://developer.uspto.gov/ptab-web/, accessed on 3 April 2022.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
AIA	America Invents Act
API	Application Program Interface
APJ	Administrative Patent Judge
BERT	Bidirectional Encoder Representations from Transformers
CBM	Covered Business Methods
CNN	Convolutional Neural Network
ECHR	European Court of Human Rights
IPR	Inter Partes Review
LIME	Local Interpretable Model-agnostic Explanations
ML	Machine Learning
NLP	Natural Language Processing
NLTK	Natural Language ToolKit
PGR	Post Grant Review
POPR	Patent Owner's Preliminary Response
PTAB	Patent Trial & Appeal Board
PTO	U.S. Patent & Trademark Office

RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machines
TF-IDF	Term Frequency – Inverse Document Frequency
XGBoost	eXtreme Gradient Boosting

References

1. Salmerón-Manzano, E. Legaltech and Lawtech: Global Perspectives, Challenges, and Opportunities. *Laws* **2021**, *10*, 24. [CrossRef]
2. Sherer, J.A.; Walters, E. Practical Magic: Law's Hands-on AI Revolution. *Law Prac.* **2018**, *44*, 32.
3. Jafari, P.; Al Hattab, M.; Mohamed, E.; AbouRizk, S. Automated Extraction and Time-Cost Prediction of Contractual Reporting Requirements in Construction Using Natural Language Processing and Simulation. *Appl. Sci.* **2021**, *11*, 6188. [CrossRef]
4. Brown, S. Peeking inside the Black Box: A Preliminary Survey of Technology Assisted Review (Tar) and Predictive Coding Algorithms for Ediscovery. *Suffolk J. Trial. App. Advoc.* **2015**, *21*, 221.
5. Yang, E.; Grossman, D.; Frieder, O.; Yurchak, R. Effectiveness Results for Popular E-Discovery Algorithms. In Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law, London, UK, 12–16 June 2017; Association for Computing Machinery: New York, NY, USA, 2017; pp. 261–264. [CrossRef]
6. Dale, R. Law and Word Order: NLP in Legal Tech. *Nat. Lang. Eng.* **2019**, *25*, 211–217. [CrossRef]
7. Rai, A.K. Machine Learning at the Patent Office: Lessons for Patents and Administrative Law. *Iowa Law Rev.* **2018**, *104*, 2617. [CrossRef]
8. Kang, D.M.; Lee, C.C.; Lee, S.; Lee, W. Patent Prior Art Search Using Deep Learning Language Model. In Proceedings of the 24th Symposium on International Database Engineering & Applications, Seoul, Korea, 12–14 August 2020; Association for Computing Machinery: New York, NY, USA, 2020.
9. Chen, L.; Xu, S.; Zhu, L.; Zhang, J.; Lei, X.P.; Yang, G.C. A Deep Learning Based Method for Extracting Semantic Information from Patent Documents. *Scientometrics* **2020**, *125*, 289–312. [CrossRef]
10. Krestel, R.; Chikkamath, R.; Hewel, C.; Risch, J. A Survey on Deep Learning for Patent Analysis. *World Pat. Inf.* **2021**, *65*, 102035. [CrossRef]
11. Callister, P.D. Law, Artificial Intelligence, and Natural Language Processing: A Funny Thing Happened on the Way to My Search Results. *Law Libr. J.* **2020**, *112*, 161–212.
12. Hu, Z.; Li, X.; Tu, C.; Liu, Z.; Sun, M. Few-Shot Charge Prediction with Discriminative Legal Attributes. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 12–16 August 2018; Association for Computational Linguistics: Stroudsburg, PA, USA, 2018; pp. 487–498.
13. Aletras, N.; Tsarapatsanis, D.; Preotiuc-Pietro, D.; Lampos, V. Predicting Judicial Decisions of the European Court of Human Rights: A Natural Language Processing Perspective. *PeerJ Comput. Sci.* **2016**, *2*, e93. [CrossRef]
14. Branting, L.K.; Pfeifer, C.; Brown, B.; Ferro, L.; Aberdeen, J.; Weiss, B.; Pfaff, M.; Liao, B. Scalable and Explainable Legal Prediction. *Artif. Intell. Law* **2021**, *29*, 213–238. [CrossRef]
15. Bansal, N.; Sharma, A.; Singh, R.K. A Review on the Application of Deep Learning in Legal Domain. In *IFIP Advances in Information and Communication Technology*; Springer: New York, NY, USA, 2019; Volume 559, pp. 374–381. [CrossRef]
16. Krass, M.S. Learning the Rulebook: Challenges Facing NLP in Legal Contexts. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montreal, QC, Canada, 3–8 December 2018.
17. Chien, C.V. Predicting Patent Litigation. *Tex. Law Rev.* **2011**, *90*, 283–329.
18. Allison, J.R.; Lemley, M.A.; Schwartz, D.L. Understanding the Realities of Modern Patent Litigation Symposium: Steps toward Evidence-Based IP. *Tex. Law Rev.* **2013**, *92*, 1769–1802.
19. Murdoch, W.J.; Singh, C.; Kumbier, K.; Abbasi-Asl, R.; Yu, B. Definitions, Methods, and Applications in Interpretable Machine Learning. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 22071–22080. [CrossRef]
20. Chien, C.; Helmers, C.; Spigarelli, A. Inter Partes Review and the Design of Post-Grant Patent Reviews. *Berkeley Technol. Law J.* **2018**, *33*, 817–854. [CrossRef]
21. Ragusa, P.A.; Zhang, Z. Opposing a Granted Patent in the USA: Post Grant and Inter Partes Review. *Pharm. Pat. Anal.* **2019**, *8*, 61–63. [CrossRef] [PubMed]
22. McClellan, F.; Wilson, D.; Armond, M. Filing Optional Reply Briefs Significantly Improves IPR Results. Law360. 1 May 2020. Available online: https://www.law360.com/articles/1260537/filing-optional-reply-briefs-significantly-improves-ipr-results (accessed on 3 April 2022).
23. United States Patent and Trademark Office (USPTO). Patent Trial and Appeal Board Consolidated Practice Guide. November 2019. Available online: https://www.uspto.gov/TrialPracticeGuideConsolidated (accessed on 3 April 2022).
24. Chen, F.C.; Lee, P.S. Inter Partes Review: Patent Killer No More? *Trends Biotechnol.* **2019**, *37*, 680–683. [CrossRef] [PubMed]
25. Jelsema, S.; Mason, A.; Vandenberg, J. Using a *Phillips* Construction in All PTAB Trials: The Impact on District Court Patent Actions and PTAB Proceedings. *Chi.-Kent J. Intell. Prop.* **2019**, *18*, 1–9.
26. Walsh, K.A. Institution Denied: The Evolution of Discretionary Denials of Inter Partes Review Under 35 U.S.C. § 314(A) Since Apple Inc. v Fintiv, Inc. *Am. Univ. Law Rev.* **2021**, *71*, 741–823.

27. Seeley, S.; Seeley, T. Establishment and Use of Non-Exclusive Factors to Deny Institution Under Secs. 314(a) and 325(d). *Chi.-Kent J. Intell. Prop.* **2021**, *20*, 169–179.
28. Unified Patents. PTAB Uses Discretion, Fintiv to Deny Petitions 38% in 2021 to Date. 22 September 2021. Available online: https://www.unifiedpatents.com/insights/2021/9/22/an-early-look-at-the-ptabs-use-of-fintiv-and-discretion-discretionary-denials-through-september-2021 (accessed on 3 April 2022).
29. Buchanan, B.G.; Headrick, T.E. Some Speculation about Artificial Intelligence and Legal Reasoning. *Stanf. Law Rev.* **1970**, *23*, 40–62. [CrossRef]
30. Frankenreiter, J.; Livermore, M.A. Computational Methods in Legal Analysis. *Annu. Rev. Law Soc. Sci.* **2020**, *16*, 39–57. [CrossRef]
31. Lawlor, R.C. What Computers Can Do: Analysis and Prediction of Judicial Decisions. *Am. Bar Assoc. J.* **1963**, *49*, 337–344.
32. Kort, F. Simultaneous Equations and Boolean Algebra in the Analysis of Judicial Decisions Jurimetrics. *Law Contemp. Probl.* **1963**, *28*, 143–163. [CrossRef]
33. Posner, R.A. The Theory and Practice of Citations Analysis, with Special Reference to Law and Economics. University of Chicago Law School, John M. Olin Law & Economics Working Paper No. 83. 1999. Available online: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=179655 (accessed on 3 April 2022).
34. Ruger, T.W.; Kim, P.T.; Martin, A.D.; Quinn, K.M. The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking. *Columbia Law Rev.* **2004**, *104*, 1150–1210. [CrossRef]
35. Jacobi, T.; Sag, M. Taking the Measure of Ideology: Empirically Measuring Supreme Court Cases. *Georget. Law J.* **2009**, *99*, 1.
36. Katz, D.M.; Bommarito, M.J.n.; Blackman, J. A General Approach for Predicting the Behavior of the Supreme Court of the United States. *PLoS ONE* **2017**, *12*, e0174698. [CrossRef]
37. Choi, J.H. An Empirical Study of Statutory Interpretation in Tax Law. *NYU Law Rev.* **2020**, *95*, 363–407. [CrossRef]
38. Alarie, B.; Niblett, A.; Yoon, A.H. Using Machine Learning to Predict Outcomes in Tax Law. *Can. Bus. Law J.* **2016**, *58*, 231–254. [CrossRef]
39. Ash, E.; Chen, D.L. *Vector Representations of Legal Belief*; Technical Report; National Bureau of Economic Research (NBER): Cambridge, MA, USA, 2018. Available online: http://users.nber.org/~dlchen/papers/Judge_Embeddings.pdf (accessed on 3 April 2022).
40. Varsava, N. Elements of Judicial Style: A Quantitative Guide to Neil Gorsuch's Opinion Writing. *N. Y. Univ. Law Rev.* **2018**, *93*, 75–126.
41. Salton, G.; Buckley, C. Term-Weighting Approaches in Automatic Text Retrieval. *Inf. Process. Manag.* **1988**, *24*, 513–523. [CrossRef]
42. Sebastiani, F. Machine Learning in Automated Text Categorization. *ACM Comput. Surv.* **2002**, *34*, 1–47. [CrossRef]
43. Kim, M.Y.; Xu, Y.; Goebel, R. Summarization of Legal Texts with High Cohesion and Automatic Compression Rate. In *New Frontiers in Artificial Intelligence*; Lecture Notes in Computer Science; Motomura, Y., Butler, A., Bekki, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2013; pp. 190–204. [CrossRef]
44. Kim, S.W.; Gil, J.M. Research Paper Classification Systems Based on TF-IDF and LDA Schemes. *Hum.-Centric Comput. Inf. Sci.* **2019**, *9*, 30. [CrossRef]
45. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. *arXiv* **2013**, arXiv:1301.3781.
46. Pennington, J.; Socher, R.; Manning, C. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 25–29 October 2014; Association for Computational Linguistics: Stroudsburg, PA, USA, 2014; pp. 1532–1543. [CrossRef]
47. Chalkidis, I.; Kampas, D. Deep Learning in Law: Early Adaptation and Legal Word Embeddings Trained on Large Corpora. *Artif. Intell. Law* **2019**, *27*, 171–198. [CrossRef]
48. Medvedeva, M.; Vols, M.; Wieling, M. Using Machine Learning to Predict Decisions of the European Court of Human Rights. *Artif. Intell. Law* **2020**, *28*, 237–266. [CrossRef]
49. Kim, Y.; Park, S.; Lee, J.; Jang, D.S.; Kang, J.H. Integrated Survival Model for Predicting Patent Litigation Hazard. *Sustainability* **2021**, *13*, 1763. [CrossRef]
50. Weires, R.; Rosefelt, J.; Meylor, K.; Shim, S.; Chong, L. Narrowing the Universe: A Machine Learning Approach to Patent Clearance. *Chi.-Kent J. Intell. Prop.* **2021**, *20*, 180.
51. McConnell, D.J.; Zhu, J.; Pandya, S.; Aguiar, D. Case-Level Prediction of Motion Outcomes in Civil Litigation. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, Sao Paulo, Brazil, 21–25 June 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 99–108.
52. Pillai, V.G.; Chandran, L.R. Verdict Prediction for Indian Courts Using Bag of Words and Convolutional Neural Network. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; pp. 676–683. [CrossRef]
53. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16), San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794. [CrossRef]
54. Qi, Z. The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model. In Proceedings of the 2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA), Dalian, China, 27–29 June 2020; pp. 1241–1246. [CrossRef]

55. dos Santos, P.T.C.; Henrique, F.; Garcia, V.; Ferreira, V.R.S.; dos Santos Neto, A.C.; Souza, J.C.; Manfredini, C.; França, J.V.F.; Boaro, J.M.C.; Junior, G.B.; et al. Multiclass Legal Judgment Outcome Prediction for Consumer Lawsuits Using XGBoost and TPE. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, ON, Canada, 11–14 October 2020; pp. 881–886.
56. Rajshekhar, K.; Zadrozny, W.; Garapati, S.S. Analytics of Patent Case Rulings: Empirical Evaluation of Models for Legal Relevance. In Proceedings of the 16th International Conference on Artificial Intelligence and Law (ICAIL 2017), London, UK, 12–16 June 2017.
57. Love, B.J.; Miller, S.P.; Ambwani, S. Determinants of Patent Quality: Evidence from Inter Partes Review Proceedings. *Univ. Colo. Law Rev.* **2019**, *90*, 67. [CrossRef]
58. Winer, D. *Predicting Bad Patents: Employing Machine Learning to Predict Post-Grant Review Outcomes for US Patents*; Technical Report UCB/EECS-2017-60; Electrical Engineering and Computer Science; University of California at Berkeley: Berkeley, CA, USA, 2017.
59. Yang, Y.H.; Cheng, P.J.; Chen, F.C. Predicting Institution Decisions in Inter Partes Review Proceedings. *J. Pat. Trademark Off. Soc.* **2018**, *100*, 697–717.
60. Yang, Z.; Yang, D.; Dyer, C.; He, X.; Smola, A.; Hovy, E. Hierarchical Attention Networks for Document Classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego, CA, USA, 12–17 June 2016; pp. 1480–1489.
61. Belinkov, Y.; Glass, J. Analysis Methods in Neural Language Processing: A Survey. *Trans. Assoc. Comput. Linguist.* **2019**, *7*, 49–72. [CrossRef]
62. Li, J.; Zhang, G.; Yu, L.; Meng, T. Research and Design on Cognitive Computing Framework for Predicting Judicial Decisions. *J. Signal Process. Syst.* **2019**, *91*, 1159–1167. [CrossRef]
63. Long, S.; Tu, C.; Liu, Z.; Sun, M. Automatic Judgment Prediction via Legal Reading Comprehension. In *Proceedings of the 18th China National Conference on Chinese Computing Linguistics, Kunming, China, 18–20 October 2019*; Lecture Notes in Computer Science Series; Springer: Cham, Switzerland, 2019; Volume 11856, pp. 558–572. [CrossRef]
64. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; Volume 1, pp. 4171–4186. [CrossRef]
65. Chalkidis, I.; Fergadiotis, M.; Malakasiotis, P.; Aletras, N.; Androutsopoulos, I. LEGAL-BERT: The Muppets Straight out of Law School. *arXiv* **2020**, arXiv:2010.02559.
66. Zheng, L.; Guha, N.; Anderson, B.R.; Henderson, P.; Ho, D.E. When Does Pretraining Help?: Assessing Self-Supervised Learning for Law and the CaseHOLD Dataset of 53,000+ Legal Holdings. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, Sao Paulo, Brazil, 21–25 June 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 159–168.
67. Wehnert, S.; Sudhi, V.; Dureja, S.; Kutty, L.; Shahania, S.; De Luca, E.W. Legal Norm Retrieval with Variations of the Bert Model Combined with TF-IDF Vectorization. In Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law, Sao Paulo, Brazil, 21–25 June 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 285–294. [CrossRef]
68. Linardatos, P.; Papastefanopoulos, V.; Kotsiantis, S.B. Explainable AI: A Review of Machine Learning Interpretability Methods. *Entropy* **2021**, *23*, 18. [CrossRef] [PubMed]
69. Yu, R.; Ali, G.S. What's Inside the Black Box? AI Challenges for Lawyers and Researchers. *Leg. Inf. Manag.* **2019**, *19*, 2–13. [CrossRef]
70. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man Is to Computer Programmer as Woman Is to Homemaker? Debiasing Word Embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *29*, 4349–4357.
71. Wu, H.C.; Luk, R.W.P.; Wong, K.F.; Kwok, K.L. Interpreting TF-IDF Term Weights as Making Relevance Decisions. *ACM Trans. Inf. Syst.* **2008**, *26*, 1–37. [CrossRef]
72. Raghupathi, V.; Zhou, Y.; Raghupathi, W. Legal Decision Support: Exploring Big Data Analytics Approach to Modeling Pharma Patent Validity Cases. *IEEE Access* **2018**, *6*, 41518–41528. [CrossRef]
73. Mahfouz, T.; Kandil, A. Construction Legal Decision Support Using Support Vector Machine (SVM). In Proceedings of the 2010 Construction Research Congress, Banff, AB, Canada, 8–10 May 2010; American Society of Civil Engineers: Reston, VA, USA, 2012; pp. 879–888. [CrossRef]
74. Ramraj, S.; Uzir, N.; Sunil, R.; Banerjee, S. Experimenting XGBoost Algorithm for Prediction and Classification of Different Datasets. *Int. J. Control Theory Appl.* **2016**, *9*, 651–662.
75. Elith, J.; Leathwick, J.R.; Hastie, T. A Working Guide to Boosted Regression Trees. *J. Anim. Ecol.* **2008**, *77*, 802–813. [CrossRef]
76. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Dutchess County, NY, USA, 2017; Volume 30.
77. Montavon, G.; Samek, W.; Müller, K.R. Methods for Interpreting and Understanding Deep Neural Networks. *Digit. Signal Process.* **2018**, *73*, 1–15. [CrossRef]
78. Shrikumar, A.; Greenside, P.; Kundaje, A. Learning Important Features through Propagating Activation Differences. In Proceedings of the 34th International Conference on Machine Learning (ICML 2017), Sydney, Australia, 6–11 August 2017; pp. 3145–3153.

79. Simonyan, K.; Vedaldi, A.; Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv* **2014**, arXiv:1312.6034.
80. Ce, P.; Tie, B. An Analysis Method for Interpretability of CNN Text Classification Model. *Future Internet* **2020**, *12*, 228. [CrossRef]
81. Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; Xu, B. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Germany, 7–12 August 2016; Volume 2, pp. 207–212.
82. Rush, A.M.; Chopra, S.; Weston, J. A Neural Attention Model for Abstractive Sentence Summarization. *arXiv* **2015**, arXiv:1509.00685.
83. Jain, S.; Wallace, B.C. Attention Is Not Explanation. *arXiv* **2019**, arXiv:1902.10186.
84. Vashishth, S.; Upadhyay, S.; Tomar, G.S.; Faruqui, M. Attention Interpretability Across NLP Tasks. *arXiv* **2019**, arXiv:1909.11218.
85. Mullenbach, J.; Wiegreffe, S.; Duke, J.; Sun, J.; Eisenstein, J. Explainable Prediction of Medical Codes from Clinical Text. *arXiv* **2018**, arXiv:1802.05695.
86. Chalkidis, I.; Androutsopoulos, I.; Aletras, N. Neural Legal Judgment Prediction in English. *arXiv* **2019**, arXiv:1906.02059.
87. Shen, Y.; He, X.; Gao, J.; Deng, L.; Mesnil, G. Learning Semantic Representations Using Convolutional Neural Networks for Web Search. In Proceedings of the 23rd International Conference on World Wide Web, Seoul, Korea, 7–11 April 2014; Association of Computing Machinery: New York, NY, USA, 2014; pp. 373–374.
88. Kim, Y. Convolutional Neural Networks for Sentence Classification. *arXiv* **2014**, arXiv:1408.5882.
89. Wang, Y.; Wei, G.Y.; Brooks, D. Benchmarking TPU, GPU, and CPU Platforms for Deep Learning. *arXiv* **2019**, arXiv:1907.10701.
90. Bai, S.; Kolter, J.Z.; Koltun, V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. *arXiv* **2018**, arXiv:1803.01271.
91. Zhang, Z.; Chen, Y.; Li, H.; Zhang, Q. IA-CNN: A Generalised Interpretable Convolutional Neural Network with Attention Mechanism. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
92. Dey, S.; Luo, H.; Fokoue, A.; Hu, J.; Zhang, P. Predicting Adverse Drug Reactions through Interpretable Deep Learning Framework. *BMC Bioinform.* **2018**, *19*, 476. [CrossRef] [PubMed]
93. Ribeiro, M.T.; Singh, S.; Guestrin, C. "Why Should i Trust You?" Explaining the Predictions of Any Classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 1135–1144.
94. van der Linden, I.; Haned, H.; Kanoulas, E. Global Aggregations of Local Explanations for Black Box Models. *arXiv* **2019**, arXiv:1907.03039.
95. Zhou, Z.; Hooker, G.; Wang, F. S-LIME: Stabilized-LIME for Model Explanation. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, Singapore, 14–18 August 2021.
96. Dieber, J.; Kirrane, S. Why Model Why? Assessing the Strengths and Limitations of LIME. *arXiv* **2020**, arXiv:2012.00093.
97. Bird, S.; Klein, E.; Loper, E. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*; O'Reilly Media, Inc.: Newton, MA, USA, 2009.
98. 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. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
99. Sokhansanj, B.A.; Zhao, Z.; Rosen, G.L. Interpretable and Predictive Deep Modeling of the SARS-CoV-2 Spike Protein Sequence. *medRxiv* **2021**. [CrossRef]
100. Grewal, M.; Petruzzi, H.; Gu, W. Ranking Parallel Petitions before the PTAB: A Survey. *Chi.-Kent J. Intell. Prop.* **2019**, *19*, 523–535.
101. Chalkidis, I. Law2Vec: Legal Word Embeddings. Available online: http://archive.org/details/Law2Vec (accessed on 3 April 2022).
102. Miller, S.P. What's the Connection between Repeat Litigation and Patent Quality: A (Partial) Defense of the Most Litigated Patents. *Stanf. Technol. Law Rev.* **2012**, *16*, 313.
103. Medianik, K. Artificially Intelligent Lawyers: Updating the Model Rules of Professional Conduct in Accordance with the New Technological Era. *Cardozo Law Rev.* **2017**, *39*, 1497–1530.
104. Wiener, F.B. Decision Prediction by Computers: Nonsense Cubed—And Worse. *Am. Bar Assoc. J.* **1962**, *48*, 1023–1028.
105. Davis, J.P. Artificial Wisdom? A Potential Limit on AI in Law (and Elsewhere). *Okla. Law Rev.* **2019**, *72*, 51–89. [CrossRef]
106. Pah, A.R.; Schwartz, D.L.; Sanga, S.; Clopton, Z.D.; DiCola, P.; Mersey, R.D.; Alexander, C.S.; Hammond, K.J.; Amaral, L.A.N. How to Build a More Open Justice System. *Science* **2020**, *369*, 134–136. [CrossRef] [PubMed]