May 2022

Prospects for Legal Analytics: Some Approaches to Extracting More Meaning from Legal Texts

Kevin D. Ashley

Follow this and additional works at: https://scholarship.law.uc.edu/uclr

Part of the Artificial Intelligence and Robotics Commons, Business Intelligence Commons, Computer Law Commons, Databases and Information Systems Commons, and the Legal Education Commons

Recommended Citation
Kevin D. Ashley, Prospects for Legal Analytics: Some Approaches to Extracting More Meaning from Legal Texts, 90 U. Cin. L. Rev. (2022)
Available at: https://scholarship.law.uc.edu/uclr/vol90/iss4/5

This Article is brought to you for free and open access by University of Cincinnati College of Law Scholarship and Publications. It has been accepted for inclusion in University of Cincinnati Law Review by an authorized editor of University of Cincinnati College of Law Scholarship and Publications. For more information, please contact ronald.jones@uc.edu.
PROSPECTS FOR LEGAL ANALYTICS: SOME APPROACHES TO EXTRACTING MORE MEANING FROM LEGAL TEXTS

Kevin D. Ashley*

TABLE OF CONTENTS

I. INTRODUCTION ................................................................. 1207
II. LEGAL TEXT ANALYTICS AND ITS LIMITATIONS .................... 1210
   A. Legal Text Analytic Methods ......................................... 1210
   B. Limitations of Legal Text Analytics -- Some Example .......... 1213
   C. Can HANs Explain Outcome Predictions? ........................... 1215
III. EXTRACTING MORE MEANING FROM LEGAL TEXTS .................. 1219
   A. Approach in SCALE -- Tag System ................................. 1220
   B. Why Identify Factors in Cases? ..................................... 1223
   C. How to Identify Factors in Case Texts ............................ 1225
   D. Applying Transformer Language Models to Extract
      Meaning from Case Texts ............................................. 1227
   E. Ranking Sentences by Explanatory Value .......................... 1229
   F. Identifying Legal Argument Structures ............................ 1230
   G. Teasing Meaning from Contract and Statutory Texts............ 1232
   H. One-shot Learning with GPT-3 ...................................... 1235
   I. Evaluating Extracted Legal Meanings in Use Cases ............. 1237
IV. CONCLUSIONS AND PROSPECTS .......................................... 1238

I. INTRODUCTION

Perhaps more than any other technology, text analytics is revolutionizing legal practice. Text analytics, also known as text mining, employs natural language processing (“NLP”), machine learning (“ML”), and other computational techniques automatically to extract meaning or semantics from text archives. In the legal domain, the focus is on analyzing case decisions, contracts, statutes, and other legal texts.

* Professor of Law and Intelligent Systems, University of Pittsburgh School of Law, and Senior Scientist, Learning Research and Development Center. The author can be reached at ashley@pitt.edu.

Editorial Note: At the invitation of Professor Felix Chang, Professor Ashley presented the substance of this Article in a talk entitled “New Developments in Legal Text Analytics”, part of a panel of talks on Modeling Legal Reasoning at the University of Cincinnati College of Law’s 2021 Corporate Symposium. The Law Review extends its gratitude to Professor Felix Chang, Professor Sean Mangan, and the entire Corporate Law Center for their continued support and growth of corporate law scholarship and innovation.

Argument mining, a subset of text analytics, focuses on text analytic discovery of argument-related information.\(^2\) This includes identifying elements of legal arguments in cases or memoranda such as premises and conclusions, argument and counter argument relationships, the roles sentences play in the legal arguments, and substantive strengths or weaknesses in a party’s claim.

Text analytic tools are widely applied in legal practice.\(^3\) They have changed the way law firms engage in discovery, perform due diligence, and even compose briefs and memoranda. Corporate legal departments are employing these tools to revamp their methods for monitoring their contractual obligations and compliance issues—not just in response to litigation or other emergencies but on a regular basis.

While the extent of the revolution in legal practice and its effect on legal employment may be inflated or disputed, text analytics have revolutionized research in Artificial Intelligence ("AI") and Law. “AI and Law” is a subarea of computer science that focuses on building computational models of tasks that require human intelligence, such as legal reasoning behaviors.\(^4\) Since the 1980s and before, the field has focused on representing legal knowledge to enable computers to model legal reasoning, argumentation, and explanation. These representations and models enabled computers to analyze cases and draw analogies to precedents, but they had one glaring weakness: they required manually representing the cases and precedents. Programmers had to read all of the cases which were to be modeled and fill in templates or schemes that computers could process. In other words, the need to manually represent the legal knowledge stymied the modeling efforts, a bottleneck that has long afflicted knowledge-based AI.\(^5\)

Today, the research paradigm in AI and Law has largely shifted. As we will see, text analytic techniques do not require engineering sets of legal concepts and features to represent legal knowledge. Rather, NLP techniques and ML can use deep learning neural networks to predict legal outcomes and classify legal texts statistically, directly from quantitative representations of the texts as vectors, that is, as a series of numbers.\(^6\) Unlike knowledge-based approaches, AI can make predictions without


\(^3\) Brian S. Haney, Applied Natural Language Processing for Law Practice, Intel. Prop. & Tech. F., B.C. L. SCH. 1, 22-31 (2020); Robert Dale, Law and Word Order: NLP in Legal Tech, 25 Nat. Language Eng’g 211 (2019); Ashley, supra note 1, at 1119.

\(^4\) Ashley, supra note 1, at 1117.


\(^6\) Haney, supra note 3; Ashley, supra note 1, at 1121-1122.
identifying elements of legal rules, issues, factors, values, or other kinds of legal knowledge. It can base its predictions on patterns and frequencies of words.

Text analytic tools have proliferated in the legal industry to the extent that, increasingly, law schools offer courses about the technology, its capabilities, and its limitations. These courses introduce students to various practical uses, or use cases, to which the technology has been applied in the legal domain. Such use cases include identifying e-documents that are relevant to a legal claim or subject to privilege, retrieving case decisions that involve a particular legal issue or answer a legal question, generating a draft memorandum supporting a particular type of motion, or predicting the statutory provisions that have been violated given a factual scenario. Some courses introduce law students to basic methods text analytic tools use, including ML and NLP, network analysis, and question answering. Students learn to apply libraries of off-the-shelf NLP and ML tools to databases of legal texts, to measure the results in terms of specialized metrics and to analyze errors by examining the examples an ML classifier gets wrong and why.

A key lesson involves helping students understand the limitations of text analytic techniques: the ML models can learn to extract useful information from legal texts, but they cannot read the texts as humans can. The models can predict outcomes but frequently cannot explain them in the terms that legal professionals normally employ. They also cannot learn to extract information implicit in the legal texts.

These limitations are important. Understanding them can save attorneys from misplacing their reliance on predictions or answers that

8. A "use case" denotes a context and set of interactions between a user and a system that would enable the user to achieve a kind of goal or to solve a type of problem. Michael Shrivathsan, *Use Cases—Definition (Requirements Management Basics)*, PROD. MGMT. INSIGHTS (Sept. 19, 2009), http://pmblog.accompa.com/2009/09/19/use-cases-definition-requirements-management-basics/.
10. See infra Section 2.1.
11. These include metrics such as accuracy, precision, recall, and F1. Accuracy is the ratio of correct predictions over the number of all predictions. Precision is the ratio of the number of positive predictions that are correct over the total number of positive predictions. Recall is the ratio of positive predictions that are correct over the number of instances that were positive. The F1 score combines precision and recall into one number that ranges from 0 to 1.0. A high F1 score like 0.8 means that both precision and recall are high. A low F1 score like 0.2 means both precision and recall are low. See ASHLEY, supra note 5 at 393, 396, 399-400.
12. Šavelka et al., supra note 7, at 171.
13. Ashley, supra note 1, at 1135-1144.
text analytic tools provide for certain problems. At the same time, these limitations are interesting in that they drive research efforts to address them.

This Article surveys recent research efforts aimed at increasing the extent to which text analytic methods can extract meaning from legal texts. These meanings may ultimately enable the tools to, if not read the texts as humans do, at least explain decisions in terms that legal professionals will understand and guide them in critically assessing the answers.

The Article begins by illustrating some examples of what current legal text analytics can accomplish using ML from legal text collections, such as predicting outcomes of claims for violations of human rights from descriptions of the facts. It uses these examples to introduce the limitation of legal text analytics mentioned above—the tools cannot explain their predictions or recommendations in terms lawyers would credit. The focus then shifts to research approaches for extracting more meaning from case texts. This includes identifying issues in the texts, factors for deciding domain name disputes, and why factors in case texts are key for explaining at least some kinds of legal decisions. New techniques are identifying sentences in cases that explain statutory terms and various kinds of argument structures in the decisions such as issues that a court addresses, its conclusions, and the reasons it employs to justify the conclusions. Similar techniques are being used to derive meaning from the texts of statutes and contracts to predict whether those texts entail, or logically imply, answers to given questions. The Article closes by laying out future prospects that turn critically on the extent to which the statistically intense approaches to modeling language can successfully address how attorneys reason, decide, and explain.

II. LEGAL TEXT ANALYTICS AND ITS LIMITATIONS

By combining ML from collections of legal texts, legal question answering, and legal network diagrams, legal text analytic applications have automated some key aspects of legal practice. These include assessing e-documents in litigation document review, analyzing contracts in due diligence searches, streamlining case information retrieval, legal question answering, and outcome prediction. Section A will provide a brief introduction to the techniques and some sample applications.

A. Legal Text Analytic Methods

ML programs employ statistical methods to “learn” models from data that can be used to classify a text or predict an outcome given the text of
a new case’s facts. In a training step, an ML algorithm takes sentences or other portions of text from cases as variables in a training set, which have been labeled according to classification types. The portions are represented as feature vectors of frequency-related information and a target label. The label is a binary decision whether a classification applies. Today, vector representations of texts include increasing statistical information about the contexts in which the terms of the texts are found, an aspect of their meaning. Using regression, decision trees, neural networks, and other techniques, an algorithm statistically models the correspondence between certain language features in the feature vectors and the target label. In the prediction step, given texts of new chunks from a test set, also represented as feature vectors, the ML model predicts the classification to assign to the text, if any. One can evaluate the ML model by comparing its classifications of the test set to manually assigned classifications.

Some ML applications in law involve classifying documents as instances of legal concepts. This includes classifying contractual provisions by topic and identifying key information, such as liquidated damage amounts or termination clauses. For example, a contract reviewing application, LawGeex, identifies thirty types of provisions in nondisclosure agreements. In a commercial study, the application identified issues, or topics, in five long nondisclosure agreements more accurately than twenty lawyers, and it did so in only twenty-six seconds as compared with the lawyers’ average of ninety-two minutes. LawGeex achieved an average accuracy of 94% versus an average accuracy rate of 85% for the lawyers. In e-discovery, ML models learn to classify e-documents as relevant to a litigation claim or subject to attorney-client privilege. In legal IR, it learns to classify court decisions as instances of a legal issue. As described below, ML can predict case outcomes:

---

14. See, e.g., Neil Araujo, iManage Acquires RAVN Systems, iMANAGE (May 25, 2017), https://imanage.com/blog/imnangeravnsystems/ (discussing the RAVN AI platform, which analyzes and summarizes large quantities of data); KIRA, https://kirasystems.com/ (last visited Aug. 17, 2021); Ashley, supra note 1, at 1136.


16. Id.


18. Paul Zhang and his team at Lexis Nexis employed ML and rule-based techniques to extract a network of legal issues from cases in the Lexis Nexis corpus. This would enable Lexis users to retrieve cases involving legal issues gleaned from the texts such as: “Thirteen-year-olds should not own a vehicle.” Paul Zhang et al., Knowledge Network Based on Legal Issues, in NETWORK ANALYSIS IN LAW 21 (Radboud Winkels et al. eds., 2014). U.S. Patent No. 9,336,305 (filed May 9, 2013).
Chalkidis, et al. applied deep learning to predict case outcomes directly from textual fact descriptions.¹⁹

Legal question answering uses NLP and ML to search large text collections to locate excerpts that match and answer a user’s question.²⁰ Programs measure similarity of the question and sentences by quantitatively comparing their vector representations, which, as noted, increasingly account for the contextual meanings of their terms.

ML plays a key role in legal question answering. An ML model learns from a training set of question and answer pairs and then assesses the likelihood that a short text answers a new query; it employs user feedback based on the users’ responses as to the utility of the system’s answers.²¹ The user feedback updates the system’s confidence in the responsiveness of its answer to the user’s version of a question. It also learns to recognize if a user’s query is one that it can answer.

Another technique of legal text analytics uses diagrams. These are graphs of the relations between objects such as cited legal cases or statutes. Law students and attorneys are most familiar with citation networks. Here, the “objects” are legal cases or statutory provisions, and the relations are the citations or cross-references. Legal IR tools make extensive use of citation networks.²² Communication connections among senders and receivers of corporate emails are an example of social networks. These can be an important tool in e-discovery. In statutory

¹⁹ Ilias Chalkidis et al., Neural Legal Judgment Prediction in English, PROC. 57TH ANN. MEETING ASSOC. FOR COMPUTATIONAL LINGUISTICS 4317 (2019).

²⁰ Ashley, supra note 1, at 1118; DANIEL JURAFSKY & JAMES H. MARTIN, SPEECH AND LANGUAGE PROCESSING: AN INTRODUCTION TO NATURAL LANGUAGE PROCESSING, COMPUTATIONAL LINGUISTICS, AND SPEECH RECOGNITION 494-520 (3rd ed. drft. 2021).

²¹ Until its closure in response to a copyright infringement lawsuit from Thomson Reuters, Ross was the best example of a legal question answering system. See ROSS, http://www.rossintelligence.com (last visited Feb. 15, 2022). Ross answered users’ queries based on matching sentences in cases, articles, and statutes. The model learned to recognize when a version of a question was one it could answer. From training with sets of question / answer pairs, it learned to assess the likelihood that a short text answers a new query. It also learned from user feedback. Upon answering a query, it invited users to “press thumbs up if the response is accurate” or “press thumbs down for another response.” (Ashley, supra note 6, Sec. 12.1.1) Other legal databases provide similar services. Lexis Nexis has developed its own question answering system, Lexis Answers. See You Ask a Question ... Lexis Answers™ Understands It. Thomson Reuters offers a legal question answering service connected with Westlaw Edge called WestSearch Plus. See WestSearch Plus. https://legal.thomsonreuters.com/en/products/westlaw-edge/westsearch-plus (last visited Apr. 30, 2022). Lastly, Casetext offers a legal question answering system called “Parallel Search” as well as free access to Casetext for law students and faculty. See CASETEXT, https://casetext.com/lawschool (last visited Feb. 15, 2022).

²² These include tools like Ravel, Casetext’s CARA A.I., and Google Scholar Cases. See RAVEL, http://ravel.com/ (last visited Aug. 17, 2021); CASETEXT, https://casetext.com/ (last visited, Aug. 17, 2021), and GOOGLE SCHOLAR CASES, https://scholar.google.com/ (last visited Aug. 17, 2021). For example, Google Scholar Cases’s “How Cited” Tool is able to determine the reason why a case is cited. It breaks down cases that cite an inputted decision into equivalence classes by the purpose for which the decision is cited. Presumably, Google Scholar achieves this by using ML to analyze the paragraph text surrounding the citation.
networks, relations among entities referred to by, and subject to, kinds of regulation across multiple statutes and jurisdictions can enable the visual comparison of regulatory frameworks.

**B. Limitations of Legal Text Analytics – Some Examples**

An impressive example of the power of current legal text analytic methods is its ability to predict case outcomes. Chalkidis, et al. applied deep learning to predict outcomes of cases before the European Court of Human Rights directly from textual descriptions of their facts. Their algorithms predict if the court found a violation of any provisions of the European Convention on Human Rights and of which provisions.

![Generic neural network architecture](image)

Deep Learning employs neural networks (“NNs”) for analyzing “big data,” that is, large quantities of data. As illustrated in Figure 1, NNs comprise input/output nodes connected to multiple layers of intermediary nodes via weighted edges. As it propagates inputs to an output node, the network combines the weights in a linear fashion. The goal is to learn a set of weights that minimizes the deviation of the computed output from the target output. The network learns the weights by training on a set of instances with known outcomes. Different architectures of networks, layers and depths are suitable for different analytic tasks. In a deep learning NN, hidden layers help learn features with predictive weight.

Chalkidis, et al. developed and trained their network with nearly 8,500

---

23. Chalkidis et al., supra note 19, at 4317.
cases of the European Court of Human Rights directly from textual descriptions of their facts. The researchers tested the trained network on 3,000 other cases that the system had not previously seen. Their algorithm based on BERT, a transformer language model discussed below, performed best with quite respectable results.

Such results are impressive, but Chalkidis’s work also illustrates a limitation of legal text analytics: it cannot explain its predictions in terms legal professionals can understand. Legal predictions require explanations. Otherwise, how is one to know if the prediction makes sense from a legal viewpoint?

Unfortunately, NNs are not ideally suited for explaining their results. To make predictions, they learn weights associated with the arcs in the network. Their “knowledge” is thus distributed across the network in a manner that makes it difficult to tease out explanations that legal professionals can understand. The nodes in the hidden layers of the NN do not correspond neatly to discrete legal features such as elements of a claim, statutory requirements, or other legal concepts that one might ordinarily employ to explain a legal conclusion. An NN cannot readily compare how different circumstances would affect its answers.

Another major limitation is that text analytic applications cannot read legal texts like lawyers can. Using statistical methods, these tools can only extract limited semantic information or meaning from legal text. That extracted information can be very useful; an application can use it to improve matching, retrieval, and ranking in legal IR. LawGeex’s identification of topics of contract provisions, for instance, could be instrumental in organizing contract provisions electronically by topic to streamline due diligence review. LawGeex, however, does not read the contracts as lawyers do. If one submitted a textual description of a particular scenario and asked if an NDA covers it, human attorneys would probably do much better. The program will know that a provision is, say, an NDA’s Definition of Protected Information but not what the text of the provision means or how it relates to the various legal issues (or due diligence risks) that may arise. It is not clear the extent to which current automated contract assessment systems can identify information implicit in contract texts or piece together information found at different places in a single contract or across multiple contracts in order

---

27. In terms of F1 Scores (see supra note 11) the model predicted whether there was a violation with an F1 score of 0.82. It predicted which provisions were violated with an F1 score of 0.60. As noted, a high F1 score like 0.82 means that both precision and recall are high.
to draw inferences about, for example, risk assessment. But think how important it is for a lawyer to read across documents and consider implicit information when reviewing contracts and engaging in legal reasoning, generally. Human lawyers accomplish it by reading texts and even by “reading between the lines.”

C. Can HANs Explain Outcome Predictions?

Researchers in legal text analytics such as Chalkidis, et al. have attempted to deal with this problem of NNs’ inability to explain predictions by using Hierarchical Attention Networks (“HANs”) to identify and highlight the portions of a text that have greater influence on the model’s outcome prediction. They hope that the highlighted portions will amount to an explanation of the prediction.

As mentioned, different architectures of networks, layers, and depths are suitable for different analytic tasks. HANs predict case outcomes and yield network attention weights, a metric reflecting the extent to which portions of the input text influenced the network’s outcome.


30. Given this limitation on A.I.’s ability to read and interpret text, when one hears claims about robots replacing attorneys, one might ask: “Who wants to hire an illiterate attorney?”
Figure 2: Hierarchical Attention Network: “Hierarchical neural model architecture for Board of Veterans Appeals cases. \( h_{\text{case}} \) is a learned function of \( h_{\text{issue}} \), built from the words in the issue section, and \( h_{\text{intro}} \), built from a hierarchical combination of the words-in-sentences and sentences in the case’s introduction section.”

A HAN is depicted in Figure 2 that computes attention weights for words (bottom right) and then for sentences (top right) in a document for which it makes a prediction. Karl Branting’s team adapted the hierarchical model to account for the structure of the types of legal case documents involved, WIPO\(^{32}\) and BVA\(^{33}\) cases. The WIPO cases comprised three sections, including history, background, and contentions. The BVA cases consisted of two sections: issue and introduction. In both, the texts of each section of a case were processed separately and the weights combined to represent the case.

---


Attention weights in HANs can then be employed to highlight the portions of a document that are most relevant in terms of weight for that prediction. For example, Chalkidis, et al. applied HANs to generate this highlighted fact description. Figure 3 shows the HAN’s attention over words (colored) and facts (vertical heat bars) in connection with predicting a violation of Article 3, prohibition of torture.\(^{34}\) The model attends more highly to the words colored in various shades of pink and red and fact sentences indicated with vertical heat bars.\(^{35}\)

The question, then, is whether HAN highlighting can be used to explain predictions and answers. Highlighting makes some intuitive sense: in dark red one sees highlighted terms including: “Suspicion,” “Police Station,” “Prosecutor,” “concussion,” and “bruises.” Terms in pink include “Overnight,” “police officers,” “prosecutor,” “forensic,” “examination,” “hospital,” and “damaged.” On the other hand, some terms meaningful to a human are not highlighted such as “beat.”

---

34. European Convention on Human Rights art. 3, Nov. 4, 1950, E.T.S. No. 155. (“No one shall be subjected to torture or to inhuman or degrading treatment or punishment.”).

35. Chalkidis et al., supra note 19, at 4319–4320. One notices that certain references to a specific city’s police station and prosecutor’s office are dark pink and thus highly weighted. Chalkidis ran predictions after removing potential bias-causing names and phrases, and found a fairly minimal effect from bias due to such demographic terminology.
Figure 4: Excerpt from a BVA case with attention weight highlighting. The blue highlighting marks the sentence with 74% of attention weight, the highest proportion. The yellow highlighting marks the next highest weight, 9%.36

Experiments by Karl Branting and his team suggest that such highlighted, high-scoring portions of texts fall short of meaningful explanation.37 They engaged sixty-one experts and non-experts on a task involving analyzing decisions of WIPO domain name disputes. The user interface of their program, which was used to predict case outcomes, employed attention-weight-based highlighting. Figure 4 shows a highlighted text from a Board of Veterans Appeals case (not a WIPO decision, but it conveys the same idea). In the experiment, “Each participant was asked to decide the issue of ‘No Rights or Legitimate Interests’ (NRLI) in two separate [WIPO] cases…and to provide a justification for each prediction.”38 In WIPO domain name disputes, having “No rights or legitimate interests” in a domain name or mark is an important issue.39

Participants in two of the four experimental conditions used a version of the program that highlighted portions of the case texts based on attention weights. The researchers observed that the highlighting had no effect on the correctness of the participants’ predictions.40 Some participants commented, however, “that they had difficulty understanding the connection between the highlighted text and the issue that they were supposed to decide.” Although one experiment is not determinative, this

36. Branting et al., Inducing Predictive Models for Decision Support in Administrative Adjudication, supra note 30, at Fig. 3.
37. Id. See also L. Karl Branting et al., Scalable and explainable legal prediction, 29 A.I. & L. 213 (2021).
39. See infra Section 3.1 for an explanation of NRLI.
40. Branting et al., Scalable and explainable legal prediction, supra note 36, at 220.
finding reduces hope that HAN attention weights can explain legal predictions. Branting recommends that “useful decision support should help the user understand the facts of cases in terms that correspond to the legal issues on which the ultimate decision depends.” Legal knowledge would be important in helping users understand this connection, but the question is: how does one integrate legal knowledge into this text analytic analysis?

As noted, deep learning NNs make predictions directly from vector representations of case texts. These text-based predictive methods have not generated any explanations or arguments supporting their predictions that legal professionals can follow. NNs do not have the legal concepts or argument structures to construct such arguments. They identify predictive features automatically. However, there are no guarantees that those features correspond to the elements of legal rules, issues, factors, values, or other legal knowledge that judges, attorneys, or law students employ to explain legal predictions. The predictive models do not necessarily represent reasons, causes, logical relationships, or argument structures to frame an explanation. As mentioned, NNs cannot read legal texts like lawyers can. They only extract some semantic information using statistical methods. This information can be very useful for automating contract reviews and other tasks, but it does not amount to reading.

III. EXTRACTING MORE MEANING FROM LEGAL TEXTS

Fortunately, text analytics research has made recent progress in automatically identifying more aspects of meaning in legal texts. These include extracting instances of legal concepts, such as factors, stereotypical patterns of fact that strengthen or weaken a side’s legal claim, and retrieving case sentences that explain statutory terms and statutory provisions relevant to a legal issue. They also incorporate rhetorical and argument structures including sentences that play particular roles in case decisions, such as stating a legal rule or a finding of fact or extracting a court’s issues, conclusions, and reasons.

Text analytics’ ability to extract more meaning from legal texts could help it explain its predictions and answers. Some of the earlier knowledge-based methods could provide ML with concepts and structures for constructing explanations. For example, ML models could be taught to identify factors in case texts; with those factors, knowledge-based methods could construct arguments and perhaps even test the ML conclusions. In addition, one who uses a legal question answering system

41. Id. at 221.
42. ASHLEY, supra note 5.
to pose a query about statutory terms could also retrieve sentences from cases that provide examples or counterexamples of a term’s application.

A. Approach in SCALE - Tag System

Recently, a promising research ML project called SCALE learned to annotate WIPO domain name dispute cases in terms of factual and legal findings.43 In order to create a training set, the researchers selected a small set of decisions and manually annotated sentences from their finding sections in terms of the labels in a tag (or type) system.

Text annotation involves marking-up texts of case decisions (or statutes) to identify instances of semantic tags or types of information. These are the concepts of interest in the texts (e.g., legal issues, case holdings, and roles of sentences in cases). The types of interest are organized into a hierarchy of concepts and relations so that an automated pipeline can automatically annotate the texts assigning them some semantic information. Supervised ML programs need sets of training instances that are manually annotated by humans to learn to classify texts by types.

The SCALE labels capture features linked to the types of findings, issues, factors, and attributes that arise in the WIPO domain name dispute cases. For instance, among other issues, Figure 5 shows the issue of “No Rights or Legitimate Interests” (“NRLI”), which was mentioned above, and the related factor of PriorBizUse. This issue comes from one of the UDRP rules governing domain name disputes. The rule states the importance of having a “Bona fide business use of the Domain Name …, prior to notice of the dispute.”44

43. Branting et al., Scalable and explainable legal prediction, supra note 36.
44. Id. at 234.
Figure 5: Tag System in SCALE

Figure 6: Four text spans annotated with factual and legal findings features.

Figure 6 shows samples of annotated text excerpts from two cases’ factual and legal findings sections. Human annotators assigned the labels that are shown in the last column. The second line from the bottom in the “Annotation” column of the chart shows an annotation of the text as a legal finding related to the issue of NRLI and the sub-issue of legitimate use. The attributes capture information in the text spans about citations of

---

46. Id. at 234 Fig. 5.
UDRP rules and polarity, that is, whether the sentence is positive or negative with respect to the issue.

The researchers leveraged their small set of manually annotated instances in order to label the sentences in the findings sections of all the cases in the WIPO corpus. For each tag, they represented with semantic vectors all the annotated spans. Semantic vectors employ word embeddings to represent each word’s “meaning” in terms of the most frequent contexts in which it occurred in the corpus. Thus, the vector representations of similar sentences tend to be closer to each other than to semantically dissimilar sentences. In this way, sentences that are located near enough to tagged sentences (actually, to the meaning of such sentences, that is, the tag centroid) are annotated in terms of that tag.

Next, the researchers employed these automatically projected tags to predict case outcomes, and they confirmed in an experiment that the tags achieved reasonable F1 values. Predicting outcomes involved learning two models, one for predicting a case’s applicable tags from the description of the facts and contentions in a case and the second for predicting decisions from these tags.

As a bonus, the tags represented substantive conceptual features, which could be used to explain the predictions. This, of course, is why Branting’s team pursued the SCALE approach. They wanted a system that could generate explanations that lawyers could understand, a system that could help the user understand the connection between relevant portions of the case record and the issues and reasoning of the case. As we have seen, it is not clear whether the hierarchical attention networks could do that. The networks highlight portions of text that weigh heavily in the prediction but do not know what the highlighted portions mean. Without the ability to identify findings, issues, and factors, HANs cannot help human users make the connection between the highlighting and the task they were meant to perform.

SCALE is an impressive step toward teasing more legal meaning from case texts, including factor-like factual patterns that have legal significance. It is subject, however, to several limitations. As Branting points out, “WIPO cases...have a high degree of stylistic consistency in the language used in Findings sections.” Beyond “stylistic” consistency, the WIPO UDRP cases are arbitration cases that not only follow a restricted format but involve domain name dispute scenarios involving a relatively small number of oft-repeated issues in oft-repeated factual contexts. Thus, one can expect to see the same kind of sentences repeated.

Despite the limitations, Branting is correct in suggesting that an

47. Id. at 222.
48. Id. at 232.
approach like SCALE can identify factors in cases beyond those involving WIPO domain names. However, let’s first consider another example of why that would be a good thing to do.

B. Why Identify Factors in Cases?

Law is a domain of rules, cases interpreting those rules, and arguments about whether to follow a rule or not. Various AI and Law approaches have modeled these phenomena computationally. One of these computational models, the Value Judgment-based Argumentative Prediction program (“VJAP”), works in the legal domain of trade secret misappropriation.49

![Image of VJAP's Domain Model of Trade Secret Misappropriation]

VJAP integrates reasoning with legal rules, cases, factors, underlying values, and effects of decisions on those values. It employs a model of the legal domain of trade secret law, which was shown in Figure 7. Legal rules defining a trade secret and how they are misappropriated (Figure 7, left) are represented with a set of logical rules (Figure 7, center top). The rules address a set of legal issues, elements of a trade secret misappropriation claim, including information valuable, maintain secrecy, info used, confidential relationship, or improper means. A set of 121 cases (Figure 7, center bottom) were represented with twenty-six factors, which are stereotypical patterns that strengthen or weaken a side’s trade secret misappropriation claim, and each were linked to the relevant

---

50. Id. at 91.
issues in the rules (Figure 7, center middle).

Crucially, Grabmair represented a set of four values protected by trade secret law, each of which was subject to four effects of decisions, pro or con, (Figure 7, top right), and these were keyed into relevant factors and issues. These values include plaintiff’s interests in property and in confidentiality and the general public’s interests in the usability of publicly available information and in fair competition. He developed a new way to connect the values to the factor-based case representations by describing in greater detail the different ways in which the factors affect protected values or interests. Certain factors may make the protected interest more or less legitimate. Others may waive the protected interest, interfere with it, or not interfere with it. This way of representing the relationship between values and factors provides more semantic information for a program to apply.

As suggested in Figure 8, when given a new case, VJAP generates an argument graph representing all the possible case-based arguments that can be made with all the relevant precedents. It propagates confidence values in those arguments based on weights it has learned through simulated annealing, a kind of iterative search for an optimal mapping of weights. It generates a quantitative prediction based on the highest-confidence arguments. In an evaluation of predictive accuracy, VJAP tied an ML algorithm, called naïve bayes, at 84%, but VJAP also generated

51. Id. at 92.
52. Id. at 94.
arguments explaining its results. Figure 8 shows an example of VJAP’s
textual arguments citing a case and drawing the analogy. As indicated by
the red-colored, underlined terms, it analogizes the case at bar as well as
two cited cases, the Dynamics and National Rejectors cases, in terms of
shared and unshared factors and also tradeoffs in underlying value
effects. Thus, a computational model of argument like VJAP does not simply
output outcomes or predictions. It also explains predictions. Its
explanations convey an understanding of the answer, some indication of
how it was obtained, and some information with which to assess whether
the answer made sense.

VJAP, of course, depends on manual representation of cases in terms of
factors. A human must read the cases to assign the relevant factors. In
short, VJAP is subject to the knowledge representation bottleneck that has
long stymied knowledge-based approaches in AI and law. If factors could
be identified automatically, however, that is all that would be necessary
to employ VJAP’s ability to generate arguments with the rules, cases,
values, and value effects. It is not obvious how any program, ML or
otherwise, could explain (or critique) its predictions of legal outcomes in
terms of arguments that lawyers will find intelligible without resources
such as VJAP’s domain model.

C. How to Identify Factors in Case Texts

Let’s now return to the question of whether the SCALE approach could
automatically identify factors such as those used in trade secret law.
Branting says that it may. The process would involve annotating text
spans of a representative set of decisions with the scalable and explainable
legal prediction factors, then matching these text spans to sentences in
unannotated cases whose vector representations are similar. Success,
according to Branting, would depend, among other things, on the degree
of congruence between the factors and the language of the decisions as
well as the stylistic consistency of the cases in the corpus.

    Rejectors, Inc. v. Trieman, 409 S.W.2d 1 (Mo. 1966).

54. Matthias Grabmair, Modeling Purposive Legal Argumentation and Case Outcome Prediction
    of Pittsburgh) (available at http://d-scholarship.pitt.edu/27608/).

55. Branting et al., Scalable and explainable legal prediction, supra note 36, at 232.
To understand the difficulty that automatically identifying trade secrets factors might entail, Figure 9 shows examples of sentences from which one can infer whether a particular trade secret misappropriation factor is present in a case. These sentences involve four factors and come from the Mason case, a trade secret dispute concerning the recipe for a cocktail, Lynchburg Lemonade.\(^{56}\) Intuitively, trade secret misappropriation cases involve more varied fact situations than WIPO domain name disputes. These cases involve many different types of information and methods for misappropriation. A recipe for a cocktail is just one example of the enormous range of ideas and information that trade secret law protects. In addition, unlike the WIPO UDRP arbitration decisions, the trade secret cases have since been litigated in a variety of courts and before different kinds of judges. As a result, the format and stylistics of their decisions vary.

Some progress has been made in extracting trade secret factors from case texts. Mohammad Falakmasir assembled a corpus of 1,600 trade secret cases from CourtListener and then employed word embeddings to represent the texts. He trained an ML algorithm, or Support Vector Machine ("SVM"), for each factor using a subset of manually classified cases. He employed 70% of the VJAP cases in the training set, holding out 30% of the cases as a test set. The model learned to predict the factors applicable in the cases in the VJAP corpus and achieved F1 values of 65%. This offered some indication that the texts contained enough factor-

---

related signals to enable ML to identify the factors.\(^\text{57}\)

Branting acknowledged another challenge for an ML algorithm to identify factors: the polarity of sentences regarding a factor. The issue with polarity is whether a system can tell the difference between a sentence that asserts a factor applies and a similar sentence that asserts it does not. In the SCALE work, they deal with this problem by assigning separate tags to positive and negative polarity forms of finding sentences.\(^\text{58}\) For some, perhaps many instances, however, the positive and negative polarity forms of such sentences may be quite similar. Compare the sentence example of pro-plaintiff factor F15: Unique-Product in Figure 9 with the following sentence: “It appears that one could order a Lynchburg Lemonade in many establishments other than that of the plaintiff.” The latter sentence would indicate that F15 does not apply, and yet it is very similar to the former.

Still, the work on SCALE is impressive and illustrates the promise of ultimately integrating ML-based predictions with conceptually meaningful explanations.

\(\text{D. Applying Transformer Language Models to Extract Meaning from Case Texts}\)

Language models are probability distributions for all the strings of characters in a language.\(^\text{59}\) Transformer language models, which are deep neural networks that are pre-trained on language modeling tasks, are achieving impressive results on a diverse set of NLP tasks. These language models include Bidirectional Encoder Representation from Transformers (“BERT”) and its variants as well as OpenAI’s GPT-3 and others. The original model is trained on a large corpus, such as Wikipedia, to perform weakly supervised tasks like predicting masked characters or the next sentence. One then adds a small extra layer to the core model to handle a targeted classification task such as classifying cases by legal area. For example, using a task-specific data set and cases manually classified by areas, the model’s parameters are further trained (i.e., fine-tuned) to perform more complex tasks. Since these models can leverage their previously learned general understanding of a language, they can be trained to perform a more complex task very well with relatively little available data.

BERT-based models have been applied successfully to legal texts. As

\(^{57}\) Mohammad Falakmasir & Kevin Ashley, Utilizing Vector Space Models for Identifying Legal Factors from Text, in LEGAL KNOWLEDGE AND INFORMATION SYSTEMS 183 (A. Wyner & G. Casini eds., 2017).

\(^{58}\) Branting et al., Scalable and explainable legal prediction, supra note 36, at 226.

\(^{59}\) Haney, supra note 3, at 7.
already mentioned, Chalkidis, et al. employed BERT to predict outcomes of cases before the European Court of Human Rights. Some of these applications are able to tease more legal meaning from case texts. BERT language models have been trained to classify Supreme Court judgments into legal areas, such as civil procedure, contract law, criminal law, etc., as well as decision texts, including judges’ arguments, as to whether the judge has accepted a claim. A BERT language model and a text similarity metric have also been applied to the task of case law entailment. Given a base case containing a text fragment and a target case, a program determined which of the target’s paragraph(s) entailed the text fragment. Savelka employed a BERT language model in learning-to-rank sentences from case law in terms of their utility in explaining statutory concepts. BERT models can also extract timeline events from case texts and classify decision texts involving Fourth Amendment searches and seizures by the type of test a court applied: a bright-line elements test or a totality-of-circumstances factors test.

In a recent project, researchers at Stanford University trained a BERT language model, Legal-BERT, on an enormous Harvard caselaw corpus. They sought to determine empirically the extent to which pretraining with legal domain texts affected the performance of ML predictors and classifiers that dealt with legal tasks and datasets. The tasks included identifying case sentences that overrule other cases as well as unfair terms of service clauses in contracts. Their own targeted task was particularly demanding: learning to select the correct case holding from among five choices, including the correct holding and four similar but incorrect

60. Chalkidis et al., supra note 19.
62. Charles Condevau et al., Weakly Supervised One-Shot Classification Using Recurrent Neural Networks with Attention: Application to Claim Acceptance Detection, 322 LEGAL KNOWLEDGE & INFO. SYS. 23 (2019) (showing that a one-shot learning approach with a small set of training examples worked well with another deep learning language model, ELMo).
65. See Erwin Filtz et al., Events Matter: Extraction of Events from Court Decisions, 322 LEGAL KNOWLEDGE & INFO. SYS. 33 (2020).
68. Id. at 162-63.
holdings. Identifying holdings is a key legal task. The researchers demonstrated that while Legal-BERT improved performance across the board, the increase was greatest for the more difficult and domain-specific legal tasks. 69

E. Ranking Sentences by Explanatory Value

Savelka’s approach to identifying and ranking sentences that explain statutory terms is an interesting example of how text analytics can extract more semantic information from legal texts. 70

In the study, law students manually classified sentences from cases with respect to their usefulness in explaining a corresponding statutory term. The students classified the sentences into four categories: high value, certain value, potential value, and no value. The sentences with high value elaborated the term’s meaning by providing definitions, a test for when the term applies, an explicit statement of what the term means, or an example or counterexample of the term. A sentence with certain value provided a basis for concluding what the term meant. A sentence with potential value provided extra information beyond that in the provision, but a sentence of no value offered no additional information.

After systematically analyzing the annotated sentences, Savelka demonstrated that while a sentence’s similarity to a provision indicates a sentence’s utility, other criteria are also important. These include novelty, topical similarity, and context. He concluded that a more sophisticated representation of the relationships among these criteria was required. One could either handcraft such a representation or let the learning algorithm do this automatically. Savelka did both.

He identified 129 features that model the retrieved sentences, their relationships to the terms of interest, and the statutory provisions from which they arose. He demonstrated that several ML algorithms learned to rank the sentences effectively based on these features. An ablation study revealed that the most useful features modeled the relationship between the source provision, the retrieved sentences, and the immediately surrounding text.

Savelka then applied BERT to the task. The BERT language models outperformed similarity-based methods (using BM25 or BM25-c) that compared the statutory provision and a retrieved sentence and its context. In effect, the BERT models were automatically discovering features that he had previously identified manually. Additionally, the BERT models

69. Id. at 164-65.

worked across statutes from multiple legal domains. They were learning something general about the usefulness of sentences for explaining statutory provisions. This is a paradigm example of how analytic methods are teasing more legal meaning from legal texts.

F. Identifying Legal Argument Structures

Researchers have made strides in teasing argument structures from case texts. This includes identifying the roles that sentences play in legal argumentation, such as stating a legal rule, a judge’s finding of fact, or an assertion of evidence.71

With this kind of argument role information, legal information retrieval (“IR”) systems could more effectively rank sentences depending on whether a user seeks legal rules, factual findings, or types of evidence.72 Using role information along with citations, an IR system could relate parts of decisions to the statutory elements a court addresses in the case.73 Role information can help a system select important sentences for automatically summarizing legal decisions.74 For example, Zhong, et al. (2019) developed a system to summarize BVA decisions. It selected important sentences to include from among the sentences it classified as predictive of a case’s outcome and as playing the roles of Reasoning or Evidential Support in the decision.75

At the University of Pittsburgh, we employ ML to identify appropriate sentences in case texts for summarizing legal decisions. The sentences correspond to legal argument triples comprised of issues, reasons, and conclusions (“IRCs”) with which to succinctly summarize some important features of a case. Issues are the legal questions which a court addressed in the case. Conclusions are a court’s decision for the corresponding issue. Reasons are sentences that elaborate on why the

71. See Kevin D. Ashley & Vern R. Walker, From Information Retrieval (IR) to Argument Retrieval (AR) for Legal Cases: Report on a Baseline Study, 259 LEGAL KNOWLEDGE & INFO. SYS. 29 (2013). Ashley and Walker advocated that legal argument retrieval (AR) systems were the next stage in the evolution of legal IR since lawyers are primarily interested in retrieving arguments and not documents.

72. Apoorva Bansal et al., Document Ranking with Citation Information and Oversampling Sentence Classification in the LUIMA Framework, 294 LEGAL KNOWLEDGE & INFO. SYS. 33 (2016).


74. See Paheli Bhattacharya et al., Identification of Rhetorical Roles of Sentences in Indian Legal Judgments, 322 LEGAL KNOWLEDGE & INFO. SYS. 3 (2019); Paheli Bhattacharya et al., A Comparative Study of Summarization Algorithms Applied to Legal Case Judgments, in EUROPEAN CONF. ON INFORMATION RETRIEVAL. 413 (2019); Paheli Bhattacharya et al., Incorporating Domain Knowledge for Extractive Summarization of Legal Case Documents, in PROC. 18TH INT’L CONF. ON A.I. & L. 22 (2021).

court reached the conclusion. The ML program was trained with a set of case summaries prepared by attorneys and annotated by law students who identified the IRCs. Xu, et al. (2021) have had some success in demonstrating that ML can learn to automatically identify argument triples, not only in case summaries, but also in the corresponding full texts of the cases.76

<table>
<thead>
<tr>
<th>17 U.S. Code § 117. Limitations on exclusive rights: Computer programs</th>
</tr>
</thead>
</table>
| (a) Making of Additional Copy or Adaptation by Owner of Copy.— Notwithstanding the provisions of section 106, it is not an infringement for the owner of a copy of a computer program to make or authorize the making of another copy or adaptation of that computer program provided: (1) that such a new copy or adaptation is created as an essential step in the utilization of the computer program in conjunction with a machine, and that it is used in no other manner, or (2) that such new copy or adaptation is for archival purposes only and that all archival copies are destroyed in the event that continued possession of the computer program should cease to be rightful.

Results
Showing top 5 results (4,682 total) for “essential step”

<table>
<thead>
<tr>
<th>Court Decision</th>
<th>Wall Data Inc. v. Los Angeles County Sheriff’s Dept.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Substantial evidence demonstrates that the Sheriff’s Department’s decision to copy the RUMBA software products was not an essential step, but a matter of convenience.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Court Decision</th>
<th>Madison River Man. v. Business Management Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>The right of an owner of a copy of a computer program to make a copy as an “essential step,” has been held to be no broader than the above-quoted rationale for the privilege, so that it is only a copy made by the very act of installing a program into a computer that is privileged.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Court Decision</th>
<th>William Krause Dba Special-T Software v. Titleserv, Inc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Utilization” of a computer program might refer exclusively to booting and running the program, in which case only limited modification, such as fixing bugs to prevent the program from crashing, might qualify as an “essential step” in booting or running the program.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Court Decision</th>
<th>Madison River Man. v. Business Management Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madison’s use of the protected ‘ICS data structure through access by its utilities and programs is not an essential step in the utilization of the computer program for the reasons set forth above.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Court Decision</th>
<th>Wall Data Inc. v. Los Angeles County Sheriff’s Dept.</th>
</tr>
</thead>
<tbody>
<tr>
<td>The “essential step” defense also ensures that a software user does not infringe when the user “copies” the software from the computer’s permanent storage (the hard drive, for example) onto its active memory (the random access memory, for example).</td>
<td></td>
</tr>
</tbody>
</table>

William Krause Dba Special-T Software v. Titleserv, Inc., 402 F.3d 119 (2d Cir. 2005)

**Figure 10: Sample target input/output for NSF project**

Our goal in a recent NSF project is to employ text analytics to expand access to legal sources, not only for legal professionals but also lay persons.77 Building on the above work, we will use deep learning legal language models to identify useful case sentences to explain statutory terms and summarize cases in terms of argument diagrams of court’s

---


issues, conclusions, and reasons. Figure 10 shows some of the sentences the program extracts for the Copyright Act’s concept of “essential step.” If the user selects a case sentence, say from the William Krause case in the middle, the goal is for the program to generate an argument diagram presenting an issue the court addressed in the case, the court’s conclusion regarding the issue, and its reasons. Ultimately, our goal is to deploy the tools through legal information institutes (“LIIs”) that provide free access to the public.

G. Teasing Meaning from Contract and Statutory Texts

Beyond legal case decisions, progress has also been made in teasing more meaning from other kinds of legal documents such as contracts and statutes. Beyond classifying contract provisions by topic, they can also be classified as consisting of obligations, rights, references, definitions, indemnities, prohibitions, etc. Additionally, they can be classified in terms of contract elements such as parties, important dates and durations, governing law, jurisdiction, etc.  

Figure 11: Using Long Short Term Memory (LSTM) networks, a kind of Recurrent NN, to mark up elements in textual contracts

---

Chalkidis, et al. (2017)\textsuperscript{80} employed deep learning to markup elements in textual contracts, such as parties, dates and durations, governing law, jurisdiction, etc., using a kind of sequence tagging with Long Short Term Memory ("LSTM") networks. LSTM models can learn long-term dependencies in texts; they decide which information to retain and which to discard from previous timesteps.\textsuperscript{81} As indicated in Figure 11, the BILSTM-LSTM-LR models outperformed the baselines (SW-LR-ALL), and they do not require manually constructing rules to extract the information.

These efforts include attempting to semi-automate aspects of due diligence review of contract provisions, such as red-flagging lease agreements for language presenting risks—extension periods, compulsory reconstruction, landlord repairs, sublease permissions, etc.\textsuperscript{82} Text analytic tools have also been applied to determine which of a half million contracts lacks forum selection or choice-of-law clauses and identify the likely template source of an agreement.\textsuperscript{83}

---

\textsuperscript{80} Id.
\textsuperscript{81} Noam Kolt, Predicting Consumer Contracts, 37 BERKELEY TECH. L.J. (forthcoming 2022).
Regarding statutory texts, recent efforts include automatically identifying semantic types of legal provisions, such as classifying statements in statutes as a prohibition, permission, definition, or reference. Savelka, et al. developed an interactive ML tool for this purpose, similar to that used for e-discovery. The tool is interactive in that it continually offers instances to a human expert to classify and uses the expert’s decisions to update its model incrementally. As indicated in Figure 12, the tool shows the user (a) an unprocessed statutory provision, (b) the features and terms deemed important in the current model and their weights, (c) summary statistics showing the distribution of relevant and non-relevant provisions up to that point, and (d) a list of the labeled provisions with confidence scores. It suggests the label for the unprocessed provision, indicates its confidence level, highlights the prominent features in the current provision at (a) and (b), solicits the user’s decision (top right), recomputes the model, and restarts the process.

Researchers are also making progress on using technology to make legal inferences from statutory provisions. For instance, they have developed a system for answering yes or no questions from the infamously difficult Japanese bar examination. Here is a sample question:

Is it true that a special provision that releases warranty can be made, but in that situation, when there are rights that the seller establishes on his/her own for a third party, the seller is not released of warranty?

The program identifies civil law statutory provisions that are textually similar to the question, identifies which of their sentences are most relevant to answering, extracts and compares semantic links between the question and relevant sentences, and then predicts if the statute entails an answer to the question. It employs unsupervised ML to obtain the entailment answer. Unsupervised ML does not learn from classified instances; it employs clustering algorithms to infer groupings of unlabeled instances based on their content. Despite rather limited data, it achieved accuracy considerably better than mere chance.

86. *Id.*
H. One-shot Learning with GPT-3

OpenAI’s GPT-3, a transformer language model, promises to make learning to respond to textual queries much more efficient and dramatically improve textual entailment. It uses an attention function to calculate the probability that a word will appear given surrounding words. Its 175 billion parameters apply greater or lesser weights to some aspect of the data reflecting what the network has learned.

The GPT-3 can learn from just a few examples. For example, we submitted the following inputs to GPT-3 by way of training it how to answer questions about the meaning of “essential step” in the Copyright Act, 17 U.S.C. § 117(a)(1), the subject of the example in Figure 10 above, an abstract of the Krause case, and three sample questions and answers concerning whether some activity with software constitutes and “essential step” within the meaning of § 117(a)(1). This training input is shown below.

GPT3: I am a highly intelligent question answering bot. If you ask me a question that is based on the passage below I will answer it correctly.

TRAINING INPUT:

Plaintiff Krause sued defendant Titleserv for copyright infringement. Krause alleged that Titleserv infringed plaintiff's copyrights by modifying the source code of eight computer programs Krause authored for Titleserv. Titleserv modified Krause’s programs by adding new features, such as check printing and providing for direct client access. These changes were designed to improve the programs' functionality in serving Titleserv’s business. The changes were not strictly necessary to keep the programs functioning. Krause argued that these modifications infringed his copyrights in his software.

Titleserv defended that its modifications were an essential step in the utilization of the program and did not infringe Krause’s copyrights. According to 17 U.S.C. § 117(a)(1) it is not copyright infringement for anyone who (i) owns a physical copy of a computer program, (ii) makes a copy or an adaptation "as an essential step in the utilization of the computer program in conjunction with a machine," and (iii) uses it "in no other manner." The CONTU Report of the National Commission on New Technological Uses of Copyrighted Works, which Congress followed, states that the rights granted by § 117(a) could "only be exercised so long as they did not harm the interests of the copyright proprietor."

The Court concluded that Titleserv's changes to its copy of Krause's programs were essential steps within the meaning of § 117(a)(1). The Court considered the concepts of "essential" and "utilization" to be ambiguous.

§ 117(a)(1) did not clearly restrict essential steps to only those strictly necessary to running the program, such as copying in the course of installing and running the program on a computer. Also, § 117(a)(1) did not clearly restrict essential steps to limited modification, such as fixing bugs to prevent the program from crashing. Finally, § 117(a)(1) did not clearly exclude from essential steps more extensive modification that involved adding new program features to enhance functionality. The Court construed § 117(a) in light of the CONTU Report. The Court found that the modest changes made by Titleserv were made only to its copies of Krause's programs. Krause enjoyed no less opportunity after Titleserv's changes, than before, to use, market, or otherwise reap the fruits of the copyrighted programs he created.

Q: If Titleserv modified Krause’s programs by adding new features, such as check printing and providing for direct client access, would that modification be an essential step within the meaning of § 117(a)(1)?
A: Yes, because even though Titleserv’s modification of Krause’s programs would enhance the programs’ functionality, the modification would be modest, the modification would only affect Titleserv’s copies of Krause’s program, and the modification would not harm the interests of Krause, the copyright owner.

Q: If Titleserv altered copies of Krause's software and marketed the altered copies to third parties, would that alteration be an essential step within the meaning of § 117(a)(1)?
A: No, because Titleserv’s alteration would affect more than Titleserv’s copies of Krause’s programs, and the alteration would harm the interests of Krause, the copyright owner.

Q: If Titleserv fixed bugs in Krause’s programs to prevent the programs from crashing, would that modification be an essential step within the meaning of § 117(a)(1)?
A: Yes, because Titleserv’s modification of Krause’s programs would be strictly necessary to run the program, the modification would affect only Titleserv’s copies of Krause’s programs, and the modification would not harm the interests of Krause, the copyright owner.

Figure 13: Test Input Questions and GPT-3’s Answers
Then, we submitted the two test input questions shown in Figure 13. The answers illustrate that with minimal training, GPT-3 can generate intelligible answers to such questions, answers that appear reasonable even if they may be wrong. GPT-3’s answer to the second test input question was incorrect. It appears that GPT-3 does not understand that copying a program from a computer’s hard drive to the computer’s random access memory is necessary to run the program on a computer or that actions necessary to run a program on a computer are likely to be “essential steps.” David Ferrucci, who led the IBM team that developed Watson, likes to call transformer models like GPT-3 “super-parrots,” noting:

They do well at narrow tasks, but only by constantly asking themselves one simple question: ‘Based on all the documents I’ve seen, what would a human likely say here?’ … What’s left unwritten … is the vast body of shared experience that led humans to write the words in the first place.”

In short, GPT-3 is capable of producing realistic appearing answers with little training, but one cannot yet rely on their correctness.

I. Evaluating Extracted Legal Meanings in Use Cases

The optimism that text analytics can extract more meaning from legal texts is qualified by the need to evaluate the extracted meanings in the context of particular use cases. As noted, “use case” means a context and set of interactions between a user and the system that would enable the user to achieve a kind of goal or solve a type of problem.

In the first place, a use case frames how one should interpret the results of an evaluation. As Walker stresses “whether performance is adequate is a function of the end use case.”

Second, in surveying research projects like those described above, one should distinguish between those that evaluate how well a program extracts legal meaning from text, an intrinsic evaluation, and those that also evaluate how well the extracted meaning serves in the intended use case, an extrinsic evaluation. Both are of interest, but the field needs more experiments that demonstrate how extracted meanings improve use-case performance. For example, an experiment could assess whether the extracted meanings improve the


reranking performance of a legal information system,92 or enable humans to solve a problem more effectively or efficiently, as in the evaluation of HAN highlighting of WIPO or BVA decisions.93

Work on case prediction and summarization illustrate this need. The highly touted prediction programs have been evaluated only intrinsically, not in the field. ML evaluations involve dividing the data into separate training and test sets, which ensures fairness, and then determining how well the trained ML model performs on the test set. As far as known, the resulting trained models have not been evaluated in the field as part of an application intended to guide human decision making about whether to bring a claim or settle it. Automatically generated legal case summarization has often been evaluated intrinsically in terms of rouge and bleu metrics. These metrics provide different ways to compare texts in terms of overlapping units of word sequences.94 Although useful, they do not measure whether a legal case summary will enable users to decide more efficiently whether to read the full case.

Evaluations of utility in use cases or with human users are expensive and difficult, but the potential intellectual and practical payouts are evident. Unfortunately, academic researchers have not conducted or published many extrinsic evaluations, probably due to the cost. If commercial researchers perform such extrinsic evaluations, they do not publish them often either, probably to protect confidential product information.

IV. CONCLUSIONS AND PROSPECTS

Even though new legal text analytic applications cannot yet read texts or explain their conclusions as lawyers can, they are extracting useful semantic information in ways that are changing legal practice. Meanwhile, as we have seen, researchers in text analytics are finding new ways to tease more meaning from legal texts. Ultimately, this will lead to new ways for ML to explain its predictions in terms that legal professionals can understand.

What are the prospects for teasing even more meaning from legal texts in the future? And what is the role of knowledge representation? The most promising methods we have seen include Branting’s SCALE, Savelka’s

92. Bansal, supra note 72.
93. Branting et al., Scalable and explainable legal prediction, supra note 36.
work on identifying useful explanatory sentences, and Walker’s research on identifying sentence argument roles, which all involved exercises in knowledge representation. These include SCALE’s type system of 122 tags in the WIPO domain name dispute corpus, Savelka’s descriptive features that model retrieved sentences explaining statutory terms, their relationships to the terms of interest, and the provisions from which they arose, and Walker’s sentence roles and use of rules to identify text patterns well enough for certain use cases. These exercises in knowledge representation provide systems a conceptual vocabulary for error analysis to understand why deep learning predictions worked or failed and conceivably for generating legal explanations.

Whether human or machine, those who would explain legal predictions need to know and manipulate the kind of legal concepts and argument structures that programs like the VJAP model employ. The prospects for ML to extract factors from case texts are good. Given a robust capability to do that, the capacity of such models to make argument-based predictions and explain them will enable users to assess ML predictions more critically and even test them by adding or removing factors.

Law schools can do a better job preparing students to understand the impact of ML and text analytics on legal practice. This includes making students aware of the problem of relying on text analytic tools, the need to learn how to evaluate the new technologies, and the opportunities the tools present for multidisciplinary interactions. At the University of Pittsburgh, my colleagues and I teach a law school course entitled Applied Legal Analytics and AI. Using basic Python Notebooks, law students engage in a progression of lessons in homework and lectures on ML and NLP. These lessons entail evaluating applications of ML and NLP to legal text data and aim to help students understand the implications and obligations of relying on such applications.

In legal education, law students learn to identify, apply, and manipulate patterns of conceptual structures through repetition and practice. In a sense, law students fine tune their more general models of reading and language to the demands of the legal domain. Can building even larger neural networks like OpenAI’s GPT-3 model achieve this too? In addition to lacking common sense knowledge about the world, deep learning approaches lack something else important. According to Marcus and Davis (2019) “deep learning … just has lots and lots of isolated bits of information known as features, without any structure.”

How can deep learning language models learn to explain and argue

95. Šavelka et al., supra note 7.
about legal conclusions without taking conceptual structures and legal knowledge into account? As the patterns and features comprising these structures become more abstract, from predicting the next character to the next word to the next sentence to the next conceptual structure or argument move, how will deep learning represent them and assign weights? Perhaps they can learn to take such structures into account, but we simply have not managed to adopt the right configurations of networks and layers that capture them and their frequency information. That is a question for a new generation of researchers in AI and Law and legal text analytics.