Informative Patents? Predicting Invalidity Decisions with the Text of Claims

JAMES HICKS*

Patent litigation presents a puzzle. On the one hand, lawyers and legal academics spend a huge amount of time and attention on the intricacies of patent drafting and the proper construction of claims. On the other, empirical studies of patent litigation consistently find that the intrinsic characteristics of a patent have little relationship with its success or failure in invalidity proceedings. The conventional wisdom is that by the time they reach a decision on the merits, validity decisions—like much of patent litigation—are a largely unpredictable affair.

In this paper, I use natural language processing to demonstrate that the conventional wisdom is incomplete. In fact, the content of patent claims is a surprisingly powerful predictor of invalidity decisions: the best performing model can correctly predict the outcome of nearly 73% of decisions in federal court, relying solely on the claim text of the disputed patent. This finding points to the potential of computational analysis to shed new light on patent validity and quality, and demonstrates the potential of predictive analytics in patent litigation.

---

* Academic Fellow, Berkeley Center for Law and Business; Ph.D. Candidate, Jurisprudence and Social Policy, UC Berkeley; james.hicks@berkeley.edu. I am grateful to John Allison, Mark Lemley, and David Schwartz for sharing their hand-coded litigation outcomes and providing thoughtful feedback on the project. For helpful comments and conversations, I thank Adam Badawi, Kristelia García, Ryan Hübert, Sonia Katyal, Aniket Kesari, Justin McCrary, Peter Menell, Tejas Narechania, Julian Nyarko, Kevin Quinn, Rachel Stern, Molly Van Houweling, and participants at the 2020 virtual IP Scholars Conference. I also thank commenters at a CPIP retreat for feedback on an unrecognizably early version of this idea. This work was supported by a grant from Berkeley's Law, Economics, and Politics Center.

I. Introduction....2
II. Prior Studies of Patent Litigation....6
III. Computational Analysis in Law....10
IV. Predicting Invalidity....15
A. Data....15
B. Text preparation....16
C. Methodology....19
D. Results....21
E. Interpretability....27
VI. Discussion....29
A. Litigation and patent quality....30
B. Caveats and limitations....31
C. Legal analytics: the coming change....32
Appendix....36
A. Receiver operating characteristic (ROC) curves....36
B. Separation graphs....37

I. INTRODUCTION

Does the text of a patent convey important information in litigation? This seems, on the face of it, to be a strange question. Deciphering and delineating the scope of the claimed invention is perhaps *the* central question of patent law. The preliminary stage of litigation, during which a patent's written claims are construed by the court—so-called *Markman* hearings—occupy a central place in patent commentary.[1] Landmark patent law cases involve the rules for the proper interpretation of claim language.[2] As Mark Lemley

---

1 *See, e.g.,* J. Jonas Anderson and Peter Menell, *Informal Deference: A Historical, Empirical, and Normative Analysis of Patent Claim Construction*, 108 NW. U. L. REV. 1 (2014) (summarizing *inter alia* the history of, and scholarly debate, on claim construction).

2 *See* Phillips v. AWH Corp., 415 F.3d 1303 (Fed. Cir. 2005) (en banc) (holding that words in a claim must be given their ordinary meaning in the context of the whole patent, as understood by a person skilled in the relevant art).

puts it: when drafting a patent claim, "every word matters."[3] And yet, when it comes to empirical studies of patent litigation, scholars find something surprising: with remarkable regularity, the observable, intrinsic features of the patent have no apparent relationship with the disposition of questions of validity.

How can we explain this apparent paradox? It could be that the process of selection into, and progression through, litigation results in a class of patents whose validity is highly uncertain to the litigants—coin flips, essentially.[4] Alternatively, it could be that patents have become highly technical scientific documents, largely impenetrable to juries (or even non-specialist district court judges) such that the actual text of the claims is, in practice, less important than other contextual factors of the litigation.[5] In this Article, I suggest and test a third explanation: that the content of patent claims *is* consequential, but our empirical tools have yet to identify that relationship.

Embedded in this third explanation is a notion of "patent quality," the identification of which has long been a question of interest to observers of the patent system.[6] Many economists and legal scholars have used litigation

---

3 Mark A. Lemley, *Without Preamble*, 100 B.U. L. REV. 357, 364 (2020).

4 This is an implication of the well-known Priest-Klein hypothesis. *See* George L. Priest and Benjamin Klein, *The Selection of Disputes for Litigation*, 13 J. LEGAL STUD. 1 (1984). I discuss questions of selection further in Part V, *infra.*

5 Patent litigation is an unusual area of U.S. law in which juries are regularly called upon to decide invalidity disputes, despite the fact that such decisions involve significant questions of law. *See* Paul R. Gugliuzza, *Law, Fact, and Patent Validity*, 106 IOWA L. REV. (forthcoming 2020).

6 A significant body of research, mostly in economics, focuses on observable characteristics such as forward citations (references to the patent in future applications) as measures of a patent's contribution to the general stock of knowledge. *See generally* ADAM B. JAFFE & MANUEL TRAJTENBERG, PATENTS, CITATIONS & INNOVATIONS: A WINDOW ON THE KNOWLEDGE ECONOMY (2002). Another line of research connects quality to value, searching for proxies and signals such as the payment of renewal fees or auction data. *See, e.g.,* Christina Odasso, Giuseppe Scellato, and Elisa Ughetto. *Selling Patents at Auction—An Empirical Analysis of Patent Value*, 24 INDUS. & CORP. CHANGE 417 (2015). However, in general, the extent to which a patent's quality aligns with its value (either to society or to the patentee) is unclear. See John R. Allison, *Patent Value*, in 2 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY LAW 47 (Peter Menell and David Schwartz eds., 2019) (contending that "quality in many cases is a necessary but insufficient condition for value").

to search for reliable proxies for quality,[7] but the concept has proven somewhat elusive, and the selection effects inherent in all studies of litigation complicate efforts to generalize the findings to non-litigated patents.[8] In his study of patent litigation involving non-practicing entities, Michael Risch concludes that "there were unobserved patent quality factors that affected whether to mount a challenge and whether that challenge was successful, but those quality factors are [not] on the face of the patent. . ."[9] It seems likely that patent quality is indeed an important determinant of eventual success or failure in litigation, but whether or not we can observe it remains an open question.

In this paper I demonstrate that some of those patent quality factors are, in fact, on the face of the patent, and that with modern computational text analysis tools we can identify them directly. To test this question, I combine an extant dataset of patent litigation outcomes with the claim text of litigated patents. Using a machine-learning technique,[10] I build a statistical model which relates each validity decision to the prevalence of specific words in the claims of the disputed patent.[11] The results are striking, showing that court decisions about patent validity are predictable to a much greater degree than has previously been recognized. With just the text of patent claims, a machine learning model can correctly predict 73% of invalidity outcomes in unseen "test" cases.[12] Combining the claim text with other metadata related to the patent and the litigation improves the model's accuracy further, although the text alone appears to encode much of the relevant information about the patent's likelihood of success.

In this paper, I make a series of contributions to the legal literature. First, I offer a new perspective on patent litigation, showing that it is more

---

7 *See infra,* Part II.

8 Allison, *supra* note __ at 73 ("…there are the many issues with selection effects and unmeasurable explanatory variables that plague studies of . . . the outcome of litigation").

9 Michael Risch, *A Generation of Patent Litigation,* 52 SAN DIEGO L. REV. 67, 122 (2015).

10 Specifically, I use an algorithm called a "random forest," which is a popular tool for classification but is relatively novel in legal analysis. *See* Part III, *infra.*

11 *See* Parts IV.A and IV.B, *infra.*

12 *See* Part IV.D., *infra.*

predictable than previously thought. This is highly salient in the patent system, which has long been decried by academics and commentators alike for its apparent lack of predictability.[13] As contingent-fee representation becomes a more significant part of the patent litigation landscape,[14] and the cost of enforcing (or defending against) a patent claim continues to rise,[15] accurate forecasting is becoming an ever-more important tool in legal practice. Second, I shed new light on a possible way to measure patent quality—a long-held goal for researchers. Although a full picture of quality is limited by selection concerns, the findings here demonstrate for the first time that claim text contains significant information about the likelihood that a patent will survive a validity challenge. Finally, I contribute to an emerging literature in empirical legal studies that demonstrates the potential of automated text analysis to reveal previously undiscovered patterns in legal text.

The balance of the paper proceeds as follows. In the next part, I discuss the current state of the empirical literature on patent litigation and patent quality. In part III, I describe the rise of computational methods in legal analysis, and situate this project within the emerging field. In part IV, I describe the data, my empirical strategy, and show how text can be used to

---

13 This (alleged) unpredictability takes many forms, including claim construction, jury decisions, arbitrary Federal Circuit reversals, and recently, Supreme Court interventions. *See, e.g.,* Gretchen Ann Bender, *Uncertainty and Unpredictability in Patent Litigation: The Time is Ripe for a Consistent Claim Construction Methodology*, 8 J. INTELL. PROP. L. 175 (2001); John F. Luman III and Christopher L. Dodson, *No Longer a Myth, the Emergence of the Patent Troll: Stifling Innovation, Increasing Litigation, and Extorting Billions*, 18 INTELL. PROP. & TECH. L.J. (2006) (asserting that the outcome of patent cases is inherently unpredictable); Gugliuzza, *supra* note __ ("jury decisions on technologically complex questions of patentability can be unpredictable"); Christopher M. Holman, *Unpredictability in Patent Law and Its Effect on Pharmaceutical Innovation*, 76 MO. L. REV. 645, 648 (2011) (noting a pervasive view in the pharmaceutical industry that patent enforcement has become "uncertain and unpredictable").

14 David L. Schwartz, *The Rise of Contingent Fee Representation in Patent Litigation*, 64 ALA. L. REV. 335 (2012) (detailing the growth of a contingent fee patent litigation practice that is surprisingly diverse across firm types).

15 *See, e.g.,* Anne S. Layne-Farrer, *The Cost of Doubling Up: An Economic Assessment of Duplication in PTAB Proceedings and Patent Infringement Litigation*, 10 LANDSLIDE 1 (2018) (noting that even for low-stakes patent litigation—with potential damages less than $10 million—the average cost to litigate to judgment was $2 million, with half of that amount incurred after discovery).

predict litigation outcomes. Part V concludes with a discussion of the implications of this work, along with some important caveats and limitations.

II. PRIOR STUDIES OF PATENT LITIGATION

Two decades ago, John Allison and Mark Lemley published one of the first rigorous and comprehensive empirical examinations of patent litigation outcomes. They showed that around 46% of challenged patents were found invalid,[16] and that non-obviousness and novelty provided the statutory bases for invalidity in the majority of decisions. Perhaps surprisingly, they also found no significant differences in invalidity rates between fields of invention. Since then, empirical studies have blossomed, exploring almost every aspect of litigation. Some of Allison and Lemley's initial findings have proven remarkably robust; others have been overturned with the times.

Scholars have since probed a wide variety of potential determinants of outcomes in patent litigation.[17] In an early study, two economists test the relationship of a variety of patent attributes (number of claims, forward and backward citations, and patentee portfolio size) with both decisions to litigate and litigation outcomes.[18] They find that "win rate outcomes are almost completely independent of observed characteristics of patents and their owners."[19] Legal scholars have paid closer attention to the characteristics of the process itself. For example, Cotropia, Lemley, and Sampat find that characteristics of the patent examination process are predictive: patents that were subject to re-examination are more likely to fail at litigation, all else equal.[20]

---

16 John R. Allison and Mark Lemley, *Empirical Evidence on the Validity of Litigated Patents*, 26 AIPLA Q.J. 185 (1998).

17 This is a large field, and the highly selective survey here serves to give a sense of the most important findings. For a recent summary of the broader literature, *see* Ronald Mann and Christopher Cotropia, *Empirical Studies in Patentability*, in 2 RESEARCH HANDBOOK ON THE ECONOMICS OF INTELLECTUAL PROPERTY LAW 281 (Peter Menell and David Schwartz eds., 2019).

18 Jean O. Lanjouw and Mark Schankerman, *Enforcement of Patent Rights in the United States,* in PATENTS IN THE KNOWLEDGE-BASED ECONOMY 145 (Wesley M. Cohen and Stephen A. Merrill eds., 2003).

19 *Id*. at 172.

20 Christopher A. Cotropia, Mark Lemley, and Bhaven Sampat, *Do Applicant Patent Citations Matter?,* 42 RES. POL'Y 844 (2013).

Similarly, Lemley, Li, and Urban consider the role of judicial experience, and note that district court judges who have more experience with patent cases are more likely to find non-infringement—though, interestingly, not invalidity.[21] Another notable change from the early studies is a growing discrepancy between industries and technology areas. In a 2012 paper, for example, Allison et al. find that "internet patents" (a subset of software patents) fare remarkably poorly, with overall patentee win rates of just 3% in their decade-long dataset.[22]

Allison and Lemley, this time with David Schwartz, returned to these questions in a trio of articles that explore patent litigation that commenced at the end of the 2000s.[23] Using a careful hand-coding of all merits decisions in cases initially filed in 2008 and 2009, they find evidence that jurisdiction makes a difference to case outcomes, as do some characteristics of the lawyers and judges involved.[24] They confirm prior wisdom that certain kinds of patents are less likely to be found invalid—particularly those relating to pharmaceuticals—and also observe that patentees lose nearly 75% of cases that are litigated to judgment.[25] In the third study, the authors analyze the role of non-practicing entities (NPEs) in litigation, finding that operating companies generally experience better outcomes in litigation than pure

---

21 Mark Lemley, Su Li, and Jennifer Urban, *Does Familiarity Breed Contempt Among Judges Deciding Patent Cases?*, 66 STAN. L. REV. 1121 (2014).

22 John R. Allison, Emerson H. Tiller, Samantha Zyontz, and Tristan Bligh, *Patent Litigation and the Internet*, 2012 STAN. TECH. L. REV. 3. *See also* Lemley et al., *supra* note __ at 1144–50.

23 John R. Allison, Mark Lemley, and David L. Schwartz, *Understanding the Realities of Modern Patent Litigation,* 92 TEX. L. REV. 1769 (2014); John R. Allison, Mark Lemley, and David L. Schwartz, *Our Divided Patent System*, 82 U. CHI. L. REV. 1073 (2015); John R. Allison, Mark Lemley, and David L. Schwartz, *How Often Do Non-Practicing Entities Win Patent Suits?,* 32 BERKELEY TECH. L.J. 235 (2017).

24 *Id., Understanding the Realities,* at 1799.

25 Allison et al., *Divided Patent System, supra* note ___ at 1112.

patent-assertion entities.[26] However, the authors caution that much of the effect is down to software patents, where NPEs are overrepresented as litigants, but no patentee fares well. Beyond this study, the role of NPEs has been the subject of significant recent scholarly attention,[27] with a mixed picture being drawn. Michael Risch, for example, finds that claims asserted by NPEs are settled at much higher rates than those by practicing entities, and that those that make it to a merits hearing see their patents held invalid at higher rates.[28]

Finally, in work most closely related to this Article, several papers have used litigation outcomes to explore the question of patent quality. Jonathan Ashtor develops a measure of the "informational content" of a patent, which combines various facially observable characteristics of the patent: number of claims, length of written description and abstract, number of inventors, and so on.[29] He first creates a set of weights by relating these variables to the number of forward citations (that is, citations to the patent by *future* applicants—a traditional measure of patent value for economists). He then tests that weighted index against validity decisions, finding that patents with more "informational content" are less likely to be invalidated.[30]

---

26 Allison et al., *Non-Practicing Entities, supra* note ___. Non-practicing entities are often referred to variously as "patent assertion entities" (PAEs) and "patent trolls." The umbrella terms represent a range of entities, including universities, individual inventors, and failed startups. Of particular note are patent-holding companies: PAEs who purchase patents and subsequently assert them in litigation, but who are not themselves involved in research and development or other innovative activities. There is some controversy over the social costs and benefits of such entities. *See, e.g.,* Christopher A. Cotropia, Jay P. Kesan and David L. Schwartz*, Unpacking Patent Assertion Entities (PAEs)*, 99 MINN. L. REV. 649 (2014). Regardless of one's position on this debate, from the perspective of litigation, there seems good reason to expect that they would behave in ways that are systematically different to other types of patent owner.

27 *See, e.g.,* Shawn Miller, et al., *Who's Suing Us? Decoding Patent Plaintiffs since 2000 with the Stanford NPE Litigation Dataset*, 21 STAN. TECH. L. REV. 234 (2018).

28 Risch, *supra* note __.

29 Jonathan H. Ashtor, *Does Patented Information Promote the Progress of Technology?*, 113 NW. U. L. REV. 943 (2019).

30 *Id.* at 980–85.

Meanwhile, Ronald Mann and Marian Underweiser explore validity decisions at the Federal Circuit.[31] Operationalizing patent quality as propensity to be held valid, the authors find that features of the patent's prosecution history are highly correlated with outcomes. (For example, a greater number of continuations and examiner-added prior art references are positively related to findings of invalidity.[32]) A particularly interesting feature of this paper is its text-based measure of the "alignment" between the patent's written description and its claims.[33] Although this metric has somewhat limited substantive application—alignment between the description and the claims are only relevant to a narrow subset of doctrinal bases of patent validity[34]—it is an intriguing early approach to incorporating the text of a patent directly into a model of litigation.

So, with all these studies, what do we know? While different scholars have emphasized quite different characteristics, we can draw some general conclusions. Certain factors seem reliably predictive of litigation outcomes. There are clear industry effects—for example, pharmaceutical patents are struck down at much lower rates than others, all else equal.[35] Jurisdiction effects, too, have been shown to be important, though the specifics are somewhat mixed.[36] Conversely, studies have generally found that forward citations provide little information about invalidity (or infringement) decisions.

---

31 Mann and Underweiser, *supra* note __.

32 *Id.* at 20–21.

33 *See infra,* Part III, for more general discussion of text-based "similarity" measures.

34 35 U.S.C. § 112 (2011).

35 Allison et al., *Divided Patent System*, *supra* note __ at 1114.

36 *Id.* For example, the Eastern District of Texas has long been thought to be patentee friendly. *See* Ofer Eldar and Neel Sukhatme, *Will Delaware Be Different? An Empirical Study of TC Heartland and the Shift to Defendant Choice of Venue*, 104 CORNELL L. REV. 101, 110-18 (2018). *But see* Lemley et al., *supra* note __ (finding that, for patent owners, D. Del. has generally been a *more* favorable venue than E.D. Tex., after conditioning on judicial experience).

As Eldar and Sukhatme describe, a recent Supreme Court decision limits plaintiff forum selection in patent suits and appears likely to change this effect in the future. Its full effect remains to be seen. *See* TC Heartland, LLC v. Kraft Food Brands Grp. LLC, 137 S. Ct. 1514 (2017) (requiring that patent infringement suits be brought in the defendant company's district of incorporation).

Most importantly for the project at hand, these studies generally find that observable attributes of the patent itself hold little to no explanatory power. As Risch summarizes: ". . .predicting which patents were invalidated had more to do with case-specific factors, such as the number of defendants, than with objectively measurable patent quality indicators."[37] The other feature that unites nearly all of these models, however, is that they are generally not very successful at explaining the data. Despite significant data collection and a host of theoretically justified covariates, tested across multiple different studies, the explanatory power of the regression models is consistently poor. As Allison, Lemley, and Schwartz conclude: "The pseudo R^2s in our regressions . . . are very low, revealing that most of the variation in patent litigation outcomes is not predictable, at least based upon the extensive variables we captured."[38]

All this prompts several questions. Are patent litigation outcomes truly unpredictable? Are we missing potentially valuable sources of information? In the next Part, I describe the computational methods that offer new ways to explore these questions.

III. COMPUTATIONAL ANALYSIS IN LAW

Legal analytics is a growing subfield both within law generally, and intellectual property specifically. The use of machine learning tools for text analysis has dramatically increased across the social sciences, but the tools hold particular promise in empirical legal studies—a field where we have large troves of text that, although highly salient, have until now been difficult to study empirically.[39] A set of machine learning approaches, referred to as natural language processing (NLP), enable rapid and flexible processing of

---

37 *See, e.g.,* Risch, *supra* note __ at 131.

38 Allison et al., *Modern Litigation*, *supra* note __ at 1799. Interesting, one feature of patent litigation—damages awards—does seem to be reasonably predictable, based on observable features of the patent and litigants. *See* Michael J. Mazzeo, Jonathan Hillel & Samantha Zyontz, *Explaining the "Unpredictable": An Empirical Analysis of U.S. Patent Infringement Awards*, 35 INT'L REV. L. & ECON. 58, 67 (2013).

39 For a broad introduction to this emerging field, *see* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS (Michael A. Livermore and Daniel N. Rockmore eds., 2019).

large quantities of text, allowing researchers to detect new and sometimes counterintuitive patterns in various kinds of legal documents.[40]

Within intellectual property, scholars have begun to employ creative uses of NLP in an effort to conceptualize and measure underlying legal doctrines. For example, a team of computer scientists and business school faculty have developed a measure of "novelty" that is based on the first entry of a word into the patent corpus.[41] Similarly, Jeffrey Kuhn and Neil Thompson propose a measure of patent scope that counts the number of words in the first claim of a patent.[42] The development of "similarity" measures is another active area of research. These metrics are based on the geometric distance between bodies of text, which are represented as (vast) vectors of words.[43] For example, Laura Pedraza-Fariña and Ryan Whalen compute scores of the mathematical similarity between the texts of thousands of patent pairs, and use those scores to derive a network-based measure of the non-obviousness doctrine.[44] Ashtor uses a similar pairwise approach in an effort to create a synthetic, algorithmic alternative to forward citations.[45]

My goal in this paper is somewhat different. Rather than recovering algorithmic measures of particular patent doctrines, my interest is to explore the variability of certain patent litigation outcomes. This is a different appli-

---

40 *See generally* Jens Frankenreiter and Michael Livermore, *Computational Methods in Legal Analysis*, 16 ANN. REV. L. & SOC. SCI. __ (2020).

41 Benjamin Balsmeier, et al., *Machine Learning and Natural-Language Processing on the Patent Corpus: Data, Tools, and New Measures*, 27 J. ECON. & MGMT. 535 (2018).

42 Intuitively, more words narrows the scope of a claim because—in general—a competing product must practice every element of the claim to infringe. Jeffrey M. Kuhn and Neil C. Thompson, *How to Measure and Draw Causal Inferences with Patent Scope,* 26 INT'L J. ECON. BUS. 5 (2019).

43 For a thorough explanation of this approach as applied to patents, *see* Kenneth A. Younge and Jeffrey M. Kuhn, *Patent-to-Patent Similarity: A Vector Space Model* (working paper, 2016), https://dx.doi.org/10.2139/ssrn.2709238.

44 Laura G. Pedraza-Fariña and Ryan Whalen, *A Network Theory of Patentability*, 87 U. CHI. L. REV. 63 (2020). Whalen in particular is at the leading edge of network-based computational approaches to innovation law. *See, e.g.,* Ryan Whalen, *Boundary Spanning Innovation and the Patent System: Interdisciplinary Challenges for a Specialized Examination System*, 47 RES. POL'Y 1334 (2018); Ryan Whalen, *Legal Networks: The Promises and Challenges of Legal Network Analysis*, 2016 MICH. ST. L. REV. 539.

45 Jonathan Ashtor, *Investigating Cohort Similarity as an Ex Ante Alternative to Patent Forward Citations*, 16 J. EMPIRICAL LEGAL STUD. 848 (2019).

cation of computational analysis to prior work in IP,[46] but the underlying tools are similar. I treat the question as one of prediction, and ask: given information about the text of a patent, how well can we predict its likely validity?

Typically, empirical studies of litigation use regression analysis of some form in order to describe the correlates of an outcome or draw causal inferences.[47] By contrast, the use of machine learning approaches to prediction

---

46 There is, however, some intriguing early work in computer science which incorporates text features in a model of selection into litigation. *See* Papis Wongchaisuwat, Diego Klabjan, and John McGinnis, *Predicting Litigation and Time to Litigate*, PROC. 16TH ANN. CONF. INT'L ASS'N ARTIFICIAL INTELLIGENCE & L. 257 (2017).

47 Traditionally, the researcher specifies a fixed (usually linear) model of the relationship between the outcome and a set of inputs—say, between plaintiff win rate and the jurisdiction hearing the case. (Usually this is provided by a theoretical model or some other domain-specific knowledge.) They then collect a sample of data, feed it to the model to produce estimates, and then assess how well it performs (how confident they are) using various goodness-of-fit tests. By contrast, the "algorithmic" culture of much modern statistics eschews theory, treating the structure of the relationship between inputs and outcomes as *a priori* unknown. Instead, researchers allow a statistical algorithm to "learn" the best model, and then validate performance by challenging the calibrated algorithm to predict outcomes from unseen data. This is a deep methodological divide. For a lively and classic discussion of the "two cultures" within (and beyond) statistics, *see generally* Leo Breiman, *Statistical Modeling: The Two Cultures*, 16 STATISTICAL SCI. 199 (2001).

and classification is quite novel in empirical legal scholarship.[48] One technique that has been deployed fruitfully is a classification tree (a type of decision tree).[49] Unlike a traditional logistic regression, a decision tree doesn't require any *ex ante* information about the structure of the relationship between the outcome and the predictors of interest. Instead, it partitions the outcome into an optimal set of "leaves," which are connected by "branches" that reflect the flow of a series of if-then logical decision rules.[50]

A landmark study uses decision trees to classify the outcome of Supreme Court cases, based on characteristics of each case and its procedural history.[51] The model was remarkably successful, correctly predicting 75% of case outcomes in the 2002 term—and significantly outperforming the panel of expert lawyers and academics who attempted the task in parallel. In the IP context, Cowart et al. demonstrate the potential of classification trees to im-

---

48 The critical literature on machine learning in *law*, on the other hand, is young but voluminous. Scholars at the intersection of law and technology, discrimination, and the criminal legal system have been particularly trenchant critics of the rise in algorithmic decision making and its troubling social consequences. *See, e.g.,* Solon Baracas and Andrew Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671 (2016) (noting the disparate impact of algorithmic approaches to employment discrimination which are based on historic patterns of prejudice); Rashida Richardson, Jason Schultz, and Kate Crawford, *Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing Systems, and Justice*, 94 NYU L. REV. ONLINE 192 (arguing that widely deployed predictive policing algorithms learn from, and help to perpetuate, past bias); Danielle Keats Citron and Frank Pasquale, *The Scored Society: Due Process for Automated Predictions*, 89 WASH. L. REV. 1 (2014) (arguing that individuals should be afforded meaningful opportunities to challenge the harmful impacts of algorithmic risk scores); FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015). *But see* Talia B. Gillis and Jann L. Spiess, *Big Data and Discrimination*, 86 U. CHI. L. REV. 459 (2019) (arguing that machine-derived decision rules can provide a framework for testing algorithmic discrimination); W. Nicholson Price and Arti Rai, *Clearing Opacity Through Machine Learning*, 106 IOWA L. REV. __ (forthcoming) (arguing that machine learning has the potential to shed light on non-intuitive, complex systems in the biomedical sciences).

49 *See* Jonathan P. Kastellec, *The Statistical Analysis of Judicial Decisions and Legal Rules with Classification Trees*, 7 J. EMPIRICAL LEGAL STUD. 202 (2010).

50 *Id.* at 209–13.

51 Theodore W. Ruger, Pauline T. Kim, Andrew D. Martin, and Kevin M. Quinn, *The Supreme Court Forecasting Project: Legal and Political Science Approaches to Predicting Supreme Court Decisionmaking,* 104 COLUM. L. REV. 1150 (2004).

prove the clarity of models of patent litigation.[52] Despite their flexibility, however, decision trees have a tendency to "overfit" to the data, often failing to predict future cases. In more recent work, a legal analytics research team use a methodological extension of decision trees—random forests—to successfully forecast the resolution of employment law litigation based on observable characteristics of the cases.[53]

This latter project is also interesting for its incorporation of case features that are extracted directly from the text of docket sheets. This is an increasingly common approach. Beyond the IP examples mentioned above, applications of automated text analytics are finding increased use in empirical approaches to law. For example, Julian Nyarko, David Pozen, and Eric Talley use the text of the Congressional Record to demonstrate the changing partisanship of "constitutional" political language over two centuries.[54] Elliott Ash and various collaborators use the text of written opinions to explore and classify judicial ideology.[55] In the corporate context, Adam Badawi shows that the text of complaints can be used to predict the outcome of securities litigation.[56]

---

52 Tammy Cowart, Roger Lirely, and Sherry Avery, *Two Methodologies for Predicting Patent Litigation Outcomes: Logistic Regression Versus Classification Trees*, 51 AM. BUS. L.J. 843 (2014) (highlighting the interpretative and flexibility advantages of classification tree analysis).

53 Charlotte S. Alexander, Khalifeh al Jadda, Mohammad Javad Feizollahi, and Anne M. Tucker, *Using Text Analytics to Predict Litigation Outcomes, in* LAW AS DATA: COMPUTATION, TEXT, & THE FUTURE OF LEGAL ANALYSIS 271-308 (Michael A. Livermore and Daniel N. Rockmore eds., 2019).

54 David Pozen, Eric Talley, and Julian Nyarko, *A Computational Analysis of Constitutional Polarization*, 105 CORNELL L. REV. 1 (2020).

55 *See, e.g.*, Elliott Ash and Daniel L. Chen, *What Kind of Judge is Brett Kavanaugh?*, 2018 CARDOZO L. REV. DE NOVO 70; Carina I. Hausladen, Marcel H. Schubert, and Elliott Ash, *Text Classification of Political Ideology Labels in Judicial Opinions*, 62 INT. REV. L. ECON. 1 (2020).

56 Adam Badawi, How Informative is the Text of Securities Complaints? (unpublished manuscript) (October 1, 2019) (on file with author). *See also* Adam Badawi and Elisabeth de Fontenay, *Contractual Complexity in Debt Agreements: The Case of EBITDA* (Duke Law School Public Law & Legal Theory Series No. 2019-67, 2019), https://dx.doi.org/10.2139/ssrn.3455497 (using NLP to analyze EBITDA definitions in credit agreements).

The approach I use here mirrors these recent legal NLP applications. I treat the claim text of a patent as a proxy for a (latent) measure of quality, and explore its relationship with invalidity decisions. In brief, my strategy is to break down the patent claims into their constituent words, and use the prevalence of those words in a given patent (with some adjustments) to predict the likelihood of success or failure in a validity adjudication. In Part IV, *infra*, I describe these steps in more detail, and introduce the dataset of litigation that I use.

IV. PREDICTING INVALIDITY

A. Data

The litigation data used in this analysis come from a set of recent studies of patent litigation conducted by John Allison, Mark Lemley, and David Schwartz. I refer the reader to those authors—in particular the first paper in their trilogy, *Understanding the Realities of Modern Patent Litigation*—for a full description of their careful hand-coding procedure.[57]

The dataset includes every outcome from a merits decision in patent cases filed in U.S. district court in 2008 and 2009.[58] The unit of analysis in the data is the patent-case combination. Patents are often litigated more than once, and if so they would show up appear multiple times in the data.[59] Simi-

---

57 Allison et al., *Understanding the Realities*, *supra* note __. For a fuller discussion of some of the complexities of coding patent litigation data, *see* Jason Rantanen, *Empirical Analyses of Judicial Opinions: Methodology, Metrics, and the Federal Circuit,* 49 CONN. L. REV. 227 (2016).

58 The cases under study commenced over a decade ago. Given the long pendency of litigation, some cases did not reach a definitive conclusion until as late as 2015, but it is nevertheless reasonable to suspect that at some of the specific patterns detected in this period would not hold true today. Patent law has experienced significant doctrinal and statutory upheaval in the past ten years. I discuss the generalizability of the core result in Part V, *infra.*

59 A patent (or individual patent claim) that is found invalid by a court is permanently lost by the patentee. A patent that has been subject to an affirmative finding of "no invalidity" or, more commonly, simply not ruled invalid remains at risk in future litigation. *See* Mark A. Lemley, *The Fractioning of Patent Law*, in INTELLECTUAL PROPERTY AND THE COMMON LAW 504 (Shyamkrishna Balganesh ed., 2013) (noting the asymmetry in which a patentee has to win every time a patent is challenged).

Table 1: Summary of outcomes (validity decisions in case-patent pairs)

	Ruled not invalid	*Ruled invalid*	*Total*
Observations	205	159	367

larly, if a case includes judgments on multiple discrete patents, the disposition of each patent is coded individually.

The key outcome variable is the result of an adjudication of invalidity ("valid" or "invalid").[60] This includes all doctrinal bases (eligible subject matter, novelty, obviousness, and so on), and decisions that occur at any stage in the litigation. If a district court judgment was subsequently overturned on appeal, then the invalidity outcome reflects that. Importantly, the outcome includes judgments of partial (in)validity. In other words, if only a subset of the claims was at issue, but all of those were found invalid, the outcome is coded as a finding of "invalidity" on the patent.

In one important respect, I treat the data differently to the previous authors. Fifteen cases (out of 216 unique cases in total) involved a split decision on a particular patent, in which some claims were found invalid while others were upheld. Although this is not a modeling problem in principle, it does make it difficult to compare the text of these patents to the other examples, all of which are coded at the case-patent level (even if only a subset of claims was at issue). To ensure all the outcomes are comparable, I drop these observations from the analysis.

B. Text preparation

I supplement the litigation data with the full claim text of each patent, which I obtain from the USPTO.[61] Because the outcomes are coded at the level of the patent rather than the individual claims, I concatenate the text of

60 Strictly speaking, courts find that a patent is not invalid.

61 USPTO Patents View, *Data Download* (Dec. 31, 2019), https://www.patentsview.org/download. The post-processed claim text will be available in an online appendix.

all claims into a single "document" for each patent.[62] Then, to make the text amenable to quantitative analysis, I put it through a series of typical "preprocessing" steps.[63]

First, I break down the patent into its constituent words, convert the words to lowercase, and remove all punctuation, numbers, and other symbols. Note that in the process, I discard all information about word order and syntactical structure. This representation of a text—known as a "bag of words"—may appear quite aggressive, but prior research in social science has shown that these bags of words convey the essential meaning of a document with remarkable power.[64] (New tools in NLP offer increasingly elaborate ways to represent the text, some of which explicitly maintain information about word co-occurrence and semantic relationships.[65] However, the traditional bag of words approach has the advantage of being very transparent, and performs well here.)

---

62 In other words, I use the combined text of every claim in a given patent to predict a decision on that patent, even when a court is asked to consider only a subset of the claims. This obviously risks introducing a certain amount of noise—in the sense of irrelevant data—into the model. On the other hand, this is a conservative choice: if the words in non-adjudicated claims are not relevant to the validity decision, they should not contribute useful information to the prediction model.

63 The method I describe is a common approach to handling text of this kind. *See, e.g.,* Eric Talley and Drew O'Kane, *The Measure of a MAC: A Machine-Learning Protocol for Analyzing* Force Majeure *Clauses in M&A Agreements*, 168 J. INSTITUTIONAL & THEORETICAL ECON. 181 (2012); Gabriel Rauterberg and Eric Talley, *Contracting Out Of The Fiduciary Duty Of Loyalty: An Empirical Analysis Of Corporate Opportunity Waivers*, 117 COLUM. L. REV. 1075 (2017).

64 Justin Grimmer and Brandon M. Stewart, *Text as Data: The Promise and Pitfalls of Automatic Content Analysis for Political Texts*, 21 POL. ANALYSIS 1, 6–7 (2013).

65 In particular, modern NLP applications often use word or document "embeddings," which represent text as low-dimensional numerical vectors that incorporate information about the co-occurence rates of words within the corpus. *See, e.g.,* Elliott Ash, Daniel L. Chen, and Arianna Ornaghi, *Gender Attitudes in the Judiciary: Evidence from U.S. Circuit Courts*, (working paper, 2020), http://elliottash.com/wp-content/uploads/2019/11/200205_Ash-Chen-Ornaghi.pdf.

Next, as is common (although not universal), I remove "stopwords" from the corpus.[66] These are phrases such as "the," "or," and "to," that are very common, but do not usually convey any semantic meaning. It is worth noting that the patent context is somewhat unusual: many traditional stopwords are themselves terms of art that have been subject to significant litigation about their interpretation.[67] The contested nature of these words may add complexity to more elaborate semantic models of patent text, but for our purposes, the choice about whether to include them is essentially an empirical question—which is more predictive? I found that predictive accuracy was degraded with stopwords included, so I remove them in the analyses that I present here.

Finally, to avoid overfitting to idiosyncratic words, I filter the dataset to exclude any term that appears in fewer than ten unique patents, and then select the 300 words that appear most commonly across the entire corpus (that is, the set of all words that appear at least once in any document). Using this filtered set of words, I count the number of times each term appears in each patent. The output of this process is set of vectors, one per patent, which contain a count of the number of times that each of the 300 terms appears in that patent. Finally, to ensure that the results are not driven primarily by the length of a patent, I normalize each word by dividing its frequency by the sum of all word counts in that patent.[68]

Taken together, all these steps result in a grid of normalized term frequencies, with a row for each patent, and a column for every word in the corpus. This representation of the text is known as a document-term matrix (or "DTM"). Table 2 shows the first few rows and columns of the trans-

---

66 I use the Snowball stopword list. *See* Stopwords 2.0, https://stopwords.quanteda.io. The full list is 175 words long, and includes pronouns, prepositions, and conjunctive words, as well as common contractions ("shan't," "won't," and so on). Additionally, I remove the word "claim," which appears very frequently in every patent but primarily provides signposting information (for example: "the device of claim 1, wherein…").

67 *See, e.g.,* Dan L. Burk and Mark A. Lemley, *Fence Posts or Sign Posts? Rethinking Patent Claim Construction*, 157 U. PA. L. REV. 1743, 1751–1753 (2009).

68 Of course, we can also control for the length of the patent in the analysis. This step simply ensures that the effect of the words themselves and the overall length are considered separately.

formed DTM for this patent corpus.[69] To create the final analysis dataset, I merge the DTM with the litigation outcomes and metadata based on the involved patent in each adjudication.[70]

Table 2: Transformed document-term matrix

	absorbent	*acceptable*	*access*	*accordance*	*according*	*acid*	...
patent1	0	0	0	0	0	0	…
patent2	0	0	0	0	0	0	…
patent3	0	0	0	0	0	0	…
patent4	0	0.031	0	0	0.062	0	…
patent5	0	0	0	0	0	0	…
patent6	0	0	0	0	0.048	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	

C. Methodology

To accommodate the textual data, I adopt a somewhat different empirical strategy to prior work in this area. The typical approach is to use a logistic regression, which is a generalization of linear regression designed for binary (rather than continuous) outcomes. A traditional linear model is problematic here for two reasons. First, textual data is inherently very "high-dimensional"—in most cases I have nearly as many variables as observations, which renders linear and logistic regression unreliable at best, and impossible at worst. Second, all these variables (text and otherwise) interact with each other in highly complex, non-linear ways, which are not generally easy

69 The grid is largely populated with zeroes; in statistical language, it is extremely sparse. For this reason, researchers often take steps to further reduce the dimensionality of the data, such as computing the singular value decomposition of the matrix. *See, e.g.,* Talley and O'Kane, *supra* note ___. For purposes of this project, I am interested in preserving information about the distinct words, so I leave the DTM as is.

70 Note that each row of the DTM will appear as many times as its patent was litigated.

to anticipate or specify *ex ante*. Machine learning offers promising solutions to both these problems.

To model the data, I use a statistical tool known as a "random forest."[71] Random forests are a generalization of the basic decision tree introduced above.[72] Decision trees are highly interpretable and flexible, but they tend to overfit and are quite sensitive to small perturbations in the data, such that the output can be quite variable from tree to tree. To circumvent this, a random forest grows thousands of trees in parallel, using randomly chosen subsets of data and covariates each time. The algorithm then combines their predictions by aggregating the "votes" of each individual tree, resulting in a stronger prediction from the whole ensemble together.

For each model, I follow a common procedure to train the algorithm and produce predictions.[73] First, I split the observations at random into a "training set" (90% of the data) and a "test set" (10%). Then, I estimate a statistical model that best fits the training data.[74] In each case, the outcome of interest is whether or not a patent is held invalid. Then, using this trained model, I predict the outcomes in the previously unused test set.[75] I repeat this procedure nine more times, each time using a distinct set of test data, such that I end up with an "out of sample" prediction of invalidity or validity

---

71 *See* Leo Breiman, *Random Forests* 45 MACHINE LEARNING 5 (2001). For the canonical textbook treatment, *see* TREVOR HASTIE, ROBERT TIBSHIRANI & JEROME FRIEDMAN, THE ELEMENTS OF STATISTICAL LEARNING 587–601 (2d ed., 2009). In the Appendix, I report results for several alternative algorithms.

72 *See* footnote __ and accompanying text.

73 *See, e.g.,* Pozen et al., *supra* note __ at 31.

74 Note that, in principle, this could be *any* statistical model—from a simple linear regression to a complex neural network.

75 Traditional measures of statistical performance assess how well a model performs "within sample." Perhaps the most well-known example is R-squared, which is used in classical linear regression to measure the proportion of variation in the data that is explained by a given model. (There are analogous metrics for logistic and other generalized models.) However, optimizing for in-sample performance risks overfitting the data—that is, producing a model that works well in one context or in a particular sample, but is not readily generalizable. Applying an algorithm to unseen data provides a more challenging and realistic assessment of its predictive accuracy. *See generally* GARETH JAMES, DANIELA WITTEN, TREVOR HASTIE & ROBERT TIBSHIRANI, AN INTRODUCTION TO STATISTICAL LEARNING 29–37 (2013).

for every observation in the data.[76] We can then compare these predictions to the hand-coded "true" outcomes in order to assess how well our algorithm performs at predicting court decisions that it has not previously encountered.

To evaluate the models, I use two common measures of performance. The first is "accuracy"—or equivalently, correct classification rate (CCR)—which is simply the number of "successes" divided by the total number of predicted outcomes. In other words, if a model correctly classified 75 out of 100 outcomes and got the other 25 wrong, its CCR would be 75%. The CCR is useful and intuitive, but comes with an important caveat. In practice, the model produces a probabilistic, rather than definitive, prediction of whether a patent will be held invalid. The CCR implicitly assumes that 50% is the correct cutoff—that is, any probability above 50% indicates (in our case) invalidity, and anything below 50% indicates validity. In some cases, and especially when the data are evenly balanced between the two classes, this can be a reasonable assumption. However, there are sometimes reasons to prefer a different cutoff, and so it is common to report a more flexible score known as the "area under the curve" (AUC).[77]

D. Results

I begin with a high-level test: how predictive are the words of the claims on their own? It is important to have a baseline against which to compare the performance of our classification algorithm. An obvious possibility is simply to predict outcomes at random—that is, to assign a 50% chance of invalidity to every case. (Intuitively, a predictive model that correctly classifies fewer than 50% of outcomes is worse than guessing.) Alternatively, a more conservative choice is to assign the *most common* outcome to

---

[76] This procedure is known as k-fold cross-validation, where "k" is equal to the number of splits of the data. Because I have relatively few observations, I use ten folds to ensure a reasonable number of cases in each training set.

[77] Rather than assuming a 50% threshold, AUC varies the cutoff from 0% to 100%, and computes the true positives (correct prediction of invalidity) and false positives (incorrect prediction of invalidity) at each point. We plot each of these pairs of values on a graph, and calculate the area under the resulting curve. An AUC of 0.5 indicates a completely uninformative classifier (that is, true and false positives are nearly equal at every possible cutoff), while an AUC of 1 reflects a model that perfectly discriminates. An example of these curves is given in the Appendix.

every case. In the dataset at hand, around 56% of patents were actually held not invalid, so a naive classifier that simply predicted that every single case would end in a finding of validity would be correct 56% of the time. Clearly, neither of these strategies produce insightful predictions—like a stopped clock, they are both right some of the time—but they provide a floor for the quality of any model.

Second, it is important to compare the performance of the new approach to a more traditional, theory-driven model. As I discuss above, empirical scholars of patent litigation have tested a huge range of potential correlates of validity, and there is no universally agreed-upon set of controls. I construct an arbitrary "canonical" logistic regression model, which includes controls for a range of attributes that are generally thought be related to validity outcomes or which have proven informative in earlier studies. I include metadata related to the patent, litigation, technology class, jurisdiction, and type of assertion entity.[78]

The main results, presented in Table 3, are striking. The random forest model correctly predicts 72.5% of validity decisions, using *only* the text of the patent claims. Meanwhile, the theory-driven model achieves an accuracy of just 64% on the same data. In other words, the text-only model achieves an

---

78 The full logistic specification uses the following covariates:

(1) Patent metadata: foreign origin; adjusted number of citations; total prior art references; and age at filing of the litigation.

(2) Litigation metadata: number of patents asserted in the suit and the number of defendants.

(3) Dummies for five primary technology areas (one excluded), and an indicator for patents that are implicated in abbreviated new drug applications. I adopt the technology areas used in the original data: mechanical, electronics, optics, biotechnology, chemical, and software.

(4) Mutually exclusive dummies for individual, failed startup, university, and patent-assertion entity (operating company is the excluded category).

(5) Dummies for three of the most important jurisdictions—the Eastern District of Texas, the Northern District of California, and the District of Delaware (all others excluded for statistical power).

Table 3: Overall performance

	Guess the most common outcome	*"Canonical" logistic regression*	*Text-only random forest*
Correctly classified	56.3%	64.4%	**72.5%**
AUC	-	0.69	**0.81**

eight percentage-point improvement on the more complex theoretical model.[79]

Table 4 shows the breakdown of correct and incorrect predictions for the text model. It correctly identified 142 of the true decisions in favor of the patentee (69%), but incorrectly predicted the remaining 63. The model performs slightly better for findings of invalidity, correctly predicting 122 out of 159 decisions (77%).[80]

Another way to break down the predictive accuracy is to look at its success (and failure) across different groups of the data. Recall that—unlike the theory-driven model—the algorithm was given no explicit information about a patent, its owner, or the litigation. To the extent that there are informative differences contained in the text of for example, electrical and software patents, the model deduced those patterns itself. Figure 1 shows a breakdown of the model's success rate across (a) different primary technology areas and (b) the type of entity asserting the patent. The graphs compare the accuracy of the text and theory-driven models, as well as a baseline that

---

79 Note also that the text-only model performs well despite being "handicapped." Several patents appear multiple times in the analysis, but the model can only produce one prediction for the text of a given patent. For example, if a patent is upheld twice before ultimately being found invalid in a third case, the model will be wrong at least once (if it predicts validity) or twice (if it predicts invalidity). In the results reported in Table 5, *infra*, I relax this constraint by including additional case-level variables.

80 In settings where one of the outcomes has more observations, random forests are known to favor that class in its predictions. *See* HASTIE ET AL., *supra* note __, at 317. To address this, I slightly upweight the "invalid" decisions at the training stage, such that the model sees more of them in each iteration. The overall accuracy varies slightly as this parameter is changed and the model favors validity decisions more or less, but this does not substantively affect the conclusions (the AUC remains steady and high).

Table 4: True outcomes versus predictions (text-only model)

		Actual decision	
		Valid	*Invalid*
Model prediction	*Valid*	142	37
	Invalid	63	122
		205	159

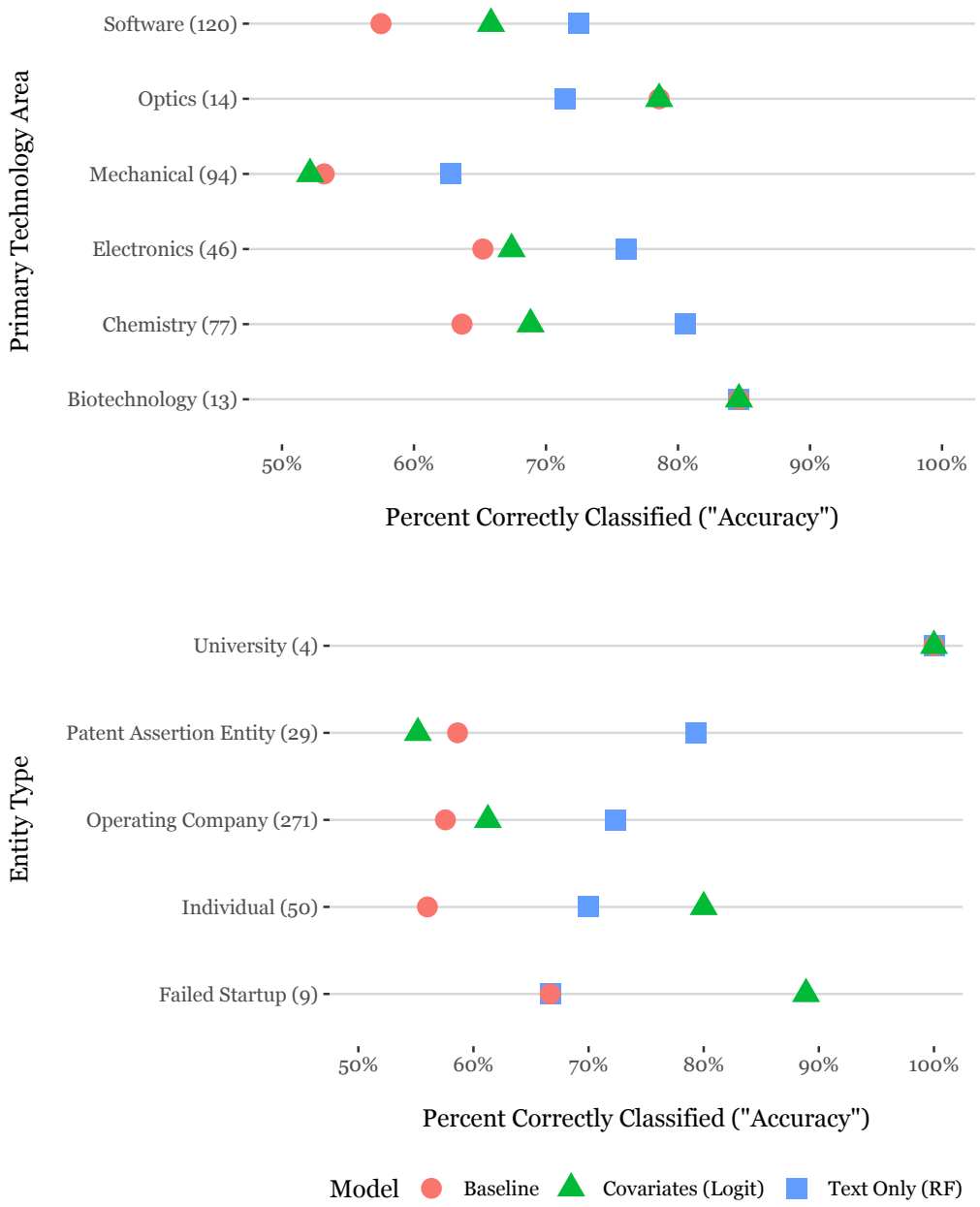
simply guesses the most common outcome for each primary technology area or asserting entity, respectively.

In general, the text model predicts well across the whole range of industries. It is notably more accurate than the traditional approach for software, chemistry, and mechanical patents. Meanwhile in two sectors—biotechnology and optics patents—the text approach performed worse or equal to the baseline. Both those technology areas, however, had relatively few challenged patents (13 and 14, respectively). If there is something unique about the claim text within each area, the algorithm likely did not have enough data to learn it for these two.

Between different types of patent owner, the text-only model had more mixed success. It is a clear improvement on past models (and on guessing) for operating companies, who make up by far the majority of patentees in this set of litigation. The accuracy amongst patents asserted by PAEs (80%) is a particularly striking improvement on the theoretical model. On the other hand, it predicted less accurately than the theory-driven model for individuals (50 observations). This is an interesting exception to the generally strong performance of the text model, and what drives it is unclear. It may be that individuals assert more idiosyncratic patents, such that any patterns are hard to discern.

Finally, I ask: how much information is contained in text *alone*? To test this, I reestimate the random forest on the same data, but this time include both the words and metadata together. Table 5 reports the accuracy and AUC for the "combined" model (the first column repeats the text-only re-

Figure 1: Accuracy of validity/invalidity predictions (higher is better)



NOTE: Number of observations in parentheses. The "baseline" predicts the most common outcome within each technology area or entity. The most common outcome for biotechnology and mechanical patents was a finding of invalidity; for chemistry, electrical, optics, and software, it was the opposite. Amongst asserting entities, the most common outcome for individuals was a finding of invalidity; for all others it was validity.

sults from Table 3 for comparison purposes). The inclusion of the litigation and patent metadata brings a very small performance improvement, taking the overall accuracy to 73.4%. (Table 6 translates this into concrete outcomes: the combined model correctly predicts two additional validity findings, and one additional invalidity finding.) It seems that a substantial amount of information about the patent is encoded in the text of the claims.

Table 5: Comparison of Random Forest model performance

	Text-only random forest	*Combined random forest*
Correctly classified	72.5%	73.4%
AUC	0.81	0.83

Table 6: True outcomes versus predictions (combined model)

		Truth	
		Valid	*Invalid*
Model prediction	*Valid*	144	36
	Invalid	61	123
		205	159

E. Interpretability

While random forests are superior to the traditional approach in terms of predictive power, they are significantly less interpretable. This is the infamous "black box" character of many machine learning models, and is often an unfortunate tradeoff in this area.[81] Unlike traditional regression analysis, the random forest model does not produce specific estimates for the effect of any single variable. (Indeed, the variables might relate to the outcome in highly non-linear ways; this flexibility is one of the model's strengths.) However, there are ways to explore what factors might be particularly important to the overall accuracy of the model.

The random forest produces "variable importance" scores, which indicate how far the accuracy of the model falls when a given variable is removed.[82] (For example, a variable importance of 0.01 would be equivalent to a 1 percentage-point decline in accuracy.) Two notes of caution are in order. First, in situations where variables are highly correlated, the variable importance scores can be artificially low. For example, if two words provide very similar (but important) information about the outcome, such as "pharmaceutical" and "pharmaceutically," then overall accuracy may not decrease when only one of them is removed. Second, the precise variable importances can

---

81 As Arti Rai has described, the "black box" nature of an algorithm can be a result of various different factors. Arti Rai, *Machine Learning at the Patent Office: Lessons for Patents and Administrative Law*, 104 IOWA L. REV. 2617 (2019). In this case of this paper, it derives from complexity. The inputs, outputs, and algorithmic steps are all quite clear, but it is difficult for humans to comprehend the workings of 2,000 trees simultaneously. *But see* Cynthia Rudin, *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*, 1 NATURE MACHINE INTELLIGENCE 206 (2019) (arguing that in areas where explanation is a particularly salient social value—for example, prediction of recidivism—we should always prefer interpretable models).

82 In practice, the algorithm randomly rearranges (permutes) the values of variable in question, so that any real relationship between that variable and the outcome is destroyed. It then re-estimates the model, and notes how much predictive accuracy falls as a consequence.

vary slightly as (apparently innocuous) model parameters are changed.[83] For both these reasons, we should interpret these scores lightly: a guide to what the model is thinking, rather than precise estimate of any particular word effect.

With those cautions in mind, figure 2 shows the fifteen most predictive words for the text-only model, estimated on the entire dataset. A few points stand out. "Amount" and "pharmaceutically" are both amongst the most predictive words. On closer inspection, they correlate very strongly with ANDA and the chemical technology area, indicating that they are serving as a proxy for drug patents. Ditto "circuit" and electrical patents. It is striking that the model picks up these relationships so clearly, and their predictive importance reinforces the salience of technology areas that has been highlighted in prior work.[84] The word "wherein" is also reliably predictive, suggesting that the relative presence of dependent claims helps to discriminate between a valid and invalid patent. In general, though, there are relatively few terms of art. Finally, all of these effects are strikingly small—no single word dominates. This indicates, unsurprisingly, that the words are related to the outcome and to each other in complex ways.

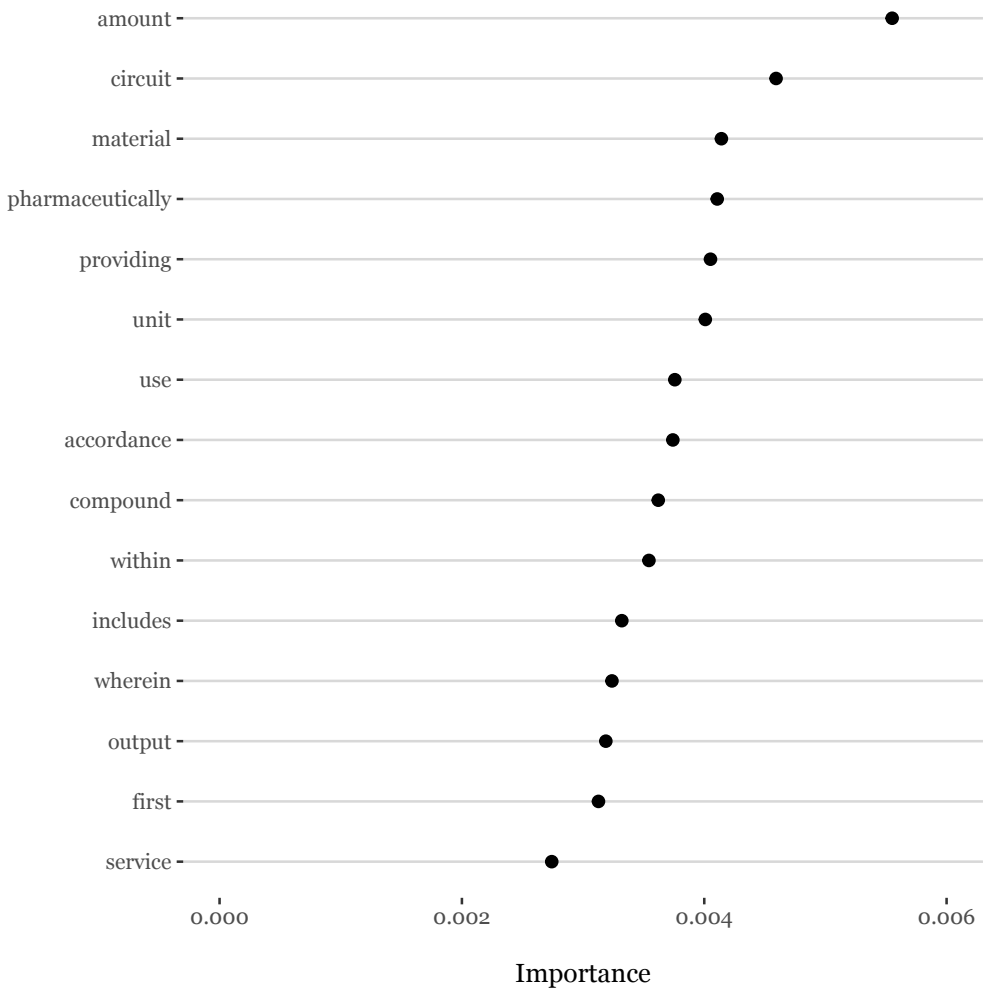
One further note of caution is warranted: nothing about these findings implies that claim text has a causal effect on outcomes. The words proxy for some underlying measure of quality, but they don't *cause* invalidity decisions in and of themselves. Consider a word that is used almost entirely in the context of pharmaceutical patents. We know from past research that those patents are held invalid at lower rates than other types of utility patent. A good algorithm would therefore be likely to find, *inter alia*, that words that were strongly associated with drug patents were contributors to predictive accuracy (as indeed this model does). But it would make no sense to try to deploy this language outside of the pharmaceutical context. The model finds patterns in the world, not causal relationships.

---

83 One example of such a parameter is the rate at which we sample from "invalid" and "valid" decisions when building the model. The potential for a multiplicity of models and algorithms to perform at similar levels of accuracy is a broader problem in machine learning, which Breiman refers to as the "Rashomon effect." Breiman, *supra* note __.

84 Allison et al., *Divided Patent System*, *supra* note __.

Figure 2: Top fifteen most predictive words for text-only model



VI. DISCUSSION

My central findings are twofold. First, that patent litigation is considerably more predictable than previous research has realized. And second, that the words on the face of the patent contain significant information about the patent's propensity to be found invalid in court. Even with a relatively small dataset—just 367 adjudications of invalidity—the algorithm is able to correctly predict a substantial fraction of outcomes using no more than the text of

the claims. These findings have important but distinct implications for research and the practice of patent litigation.

A. Litigation and patent quality

On the academic side, there are multiple avenues for future research. While the text of claims appears to encode much of the information that is also contained in the traditional correlates of litigation outcomes, it appears that the claims contain additional information about the likelihood of a validity decision—above and beyond the usual metadata. Although the goals of prediction and explanation are in some tension here, there are potential ways to incorporate textual measures of quality into more traditional, interpretable models of litigation, particularly given sufficient data.[85]

What can we learn about patent quality more generally? The possibility of directly extrapolating this result to a broader set of non-litigated patents is hampered by serious selection concerns. It is widely understood the patents that are subject to a decision in litigation are not representative of all patents, and it would be a mistake to assume a textual model of the quality of a disputed patent could be applied directly to the broader population.[86] However, the methods presented here can be applied to help better understand the patents at different stages in their lifecycle. An obvious possibility is to use analysis of the text of litigated and non-litigated patents to improve models of selection into litigation. This is an area of significant interest to academics and industry alike. As Colleen Chien has argued, our limited understanding of propensity for litigation renders patent litigation mostly uninsurable, and drives companies to amass large patent portfolios as defensive arsenals to

---

85 For example, with more litigated patents to draw on, it might be possible to compute a low-dimensional characterization of the description and claim text that could be incorporated into linear regression, as a text-based control for patent quality. *See, e.g.,* Badawi, *supra* note __.

86 *See, e.g.,* Michael D. Frakes and Melissa F. Wasserman, *Do Patent Law Suits Target Invalid Patents?, in* SELECTION AND DECISION IN JUDICIAL PROCESS AROUND THE WORLD: EMPIRICAL INQUIRIES 6 (Yun-chien Chang ed., 2019); Allison, *supra* note __ at 56–7.

ward off the threat of it.[87] But despite various academic efforts,[88] there remains a great deal of ambiguity about exactly which characteristics of a patent reliably predict entry into litigation.[89] The results in this Article suggest that the text of patents can provide important insight on this question.[90]

B. Caveats and limitations

Beyond the general concern about the limits of explainability discussed in Part IV, *infra*, there are some specific limitations to the findings here. A major caveat of the study is its relatively short time window. The data I use here are extremely high quality, and permit us to compare the performance of computational text analysis to existing studies, but they only account for cases filed in 2008 and 2009.[91] In the past ten years, patent law has been in significant flux—the changes have been both doctrinal[92] and statutory[93]—and

---

87 *See* Colleen Chien, *Predicting Patent Litigation*, 90 TEX. L. REV. 283 (2011) (finding that after-acquired characteristics are somewhat predictive of selection into litigation).

88 *Id. See also* John Allison, Mark A. Lemley, Kimberly A. Moore, and Derek Trunkey, *Valuable Patents*, 92 GEO. L.J. 435 (2004) (litigated patents are younger, contain more claims, and are more highly cited than a non-litigated matched group); Jean O. Lanjouw and Mark Schankerman, *Characteristics of Patent Litigation: A Window on Competition*, 32 RAND J. Econ. 129 (2001) (finding that litigated patents are, *inter alia*, more highly cited and have more claims).

89 *See* Lee Petherbridge, *On Predicting Patent Litigation,* 90 TEX. L. REV. SEE ALSO 75, 79 (2011) (noting the need for a more "specific and sensitive test" given that litigation is a rare event relative to the population of issued patents).

90 I intend to explore this question in future work.

91 The average duration of patent litigation filed in this period was just over three years, such that most cases were resolved by the end of 2012, although certain cases which were appealed were pending for considerably longer.

92 *See, e.g.,* Mayo Collaborative Servs. v. Prometheus Labs., Inc., 566 U.S. 66 (2012) (mere applications of laws of nature are not eligible subject matter in the absence of some additional inventive step); Ass'n for Molecular Pathology v. Myriad Genetics, Inc., 569 U.S. 576 (2013) (isolated DNA sequences that occur naturally are not eligible subject matter); Alice Corp. v. CLS Bank Int'l, 573 U.S. 208 (2014) (abstract ideas are not rendered patentable merely through a generic computational implementation); Nautilus, Inc. v. Biosig Instruments, Inc., 572 U.S. 898 (2014) (patent claims must inform those skilled in the art of the scope of the invention, with reasonable certainty).

93 Leahy-Smith America Invents Act, Pub. L. No. 112-29, 125 Stat. 284 (2011). *See also* Mark A. Lemley, *The Surprising Resilience of the Patent System*, 95 TEX. L. REV. 1, 1–6 (2016) (describing some of the major changes to and in the patent system in the last three decades).

it would be a mistake to extrapolate this *particular* model to a case filed today and expect it to predict well. Although I intend in future work to extend my analysis to recent litigation, my contribution in this Article is primarily conceptual. At that level, there appears to be little reason to expect that claim text would have become less informative as a result of the intervening legal changes even if, as seems likely, the precise *structure* of the patterns has changed.[94]

At the same time, patent litigation *itself* may have changed. There has been a significant increase in the volume of litigation brought by NPEs.[95] In the study window, relatively few cases fall into this category. But the following years played host to a stark change: Cotropia et al. report that nearly 50% of patent litigation was brought by NPEs of some kind in 2012, up nearly 30 percentage points since 2010.[96] This change, primarily driven by patent holding companies, is consequential to the extent that we think patents litigated by these new PAEs are different *in kind* to other patents. For example, it may alter the selection mechanism if PAEs are more prone to advance low-probability cases to merits decisions. Again though, there's no *a priori* reason to assume that the text model would perform poorly in this context—indeed, it was remarkably successful at discriminating between the valid and invalid patents asserted by PAEs in 2008 and 2009.

C. Legal analytics: the coming change

Although there are valuable insights in this work for scholars, perhaps the most fertile ground for the predictive analytics technology demonstrated here is in the profession. A central component of litigation practice lies in answering what are, essentially, questions of prediction. How likely are we to succeed on this eligibility question? How strong is this patent? What are

---

94 In one important regard, this claim might be controversial. The recent line of Supreme Court cases addressing patentable subject matter (described in note __, *supra*) has been widely criticized for creating an insolubly ambiguous standard for patent eligibility. But recent survey evidence indicates that patent prosecutors (though not litigators!) are still able to predict outcomes with some degree of confidence. *See* Jason D. Reinecke, *Is the Supreme Court's Patentable Subject Matter Test Overly Ambiguous? An Empirical Test*, 2019 UTAH L. REV. 581. This is an obvious area for future research.

95 Cotropia et al., *supra* note __.

96 *Id*. at 674.

the chances of a successful invalidity counterclaim? The results in this paper suggest that validity decisions can be forecast with greater confidence than previously realized. To be sure, uncertainty remains; even for this model, many predictions fall into a "grey area."[97] But the higher probability predictions allow patentees to make more informed, confident choices about their litigation strategy, and ultimately to reduce unnecessary litigation costs.[98] From the perspective of alleged infringers, better *ex ante* prediction tools can also serve as a prophylactic against strategies which assert many patents in a single suit—some of which may be of questionable validity—in order to overwhelm defendants.[99] Accurate prediction is also highly salient given the growth of contingent fee representation in patent litigation.[100]

These changes sit in the context of broader developments in law and legal analytics.[101] A burgeoning literature on the use of predictive analytics in legal practice evinces both enthusiasm from proponents of "legal tech,"[102] and reservations from those concerned about its impact on the profession and the fair administration of justice.[103] On the other hand, David Engstrom and Jonah Gelbach argue that in most areas of law, NLP will remain too su-

---

97 The graphs in Appendix Part B give a visual sense of how well the various models succeed in separating true valid and true invalid decisions. In cases where the text model gives a predicted probability of validity above 70% (or below 30%), there are very few false positives.

98 *See supra* note __.

99 *See* Schwartz, *supra* note __ at 375-6.

100 *Id.*

101 Engstrom and Gelbach catalog some of the many companies now operating in the "legal tech" space. Most provide sophisticated data aggregation and presentation (e.g., Lex Machina; Gavelytics), but a number focus explicitly on questions of prediction, including Blue J Legal (tax) and Colossus (insurance). *See* David Freeman Engstrom and Jonah B. Gelbach, *Legal Tech, Civil Procedure, and the Future of Adversarialism*, 169 U. PA. L. REV. ___ (2020).

102 *See, e.g.,* Daniel Martin Katz, *Quantitative Legal Prediction—Or—How I Learned to Stop Worrying and Start Preparing for the Data-Driven Future of the Legal Services Industry*, 62 EMORY L.J. 910 (2013); Benjamin Alarie, *The Path of Law: Towards Legal Singularity*, 66 U. TORONTO L.J. 443 (2016); John O. McGinnis and Russell G. Pearce, *The Great Disruption: How Machine Intelligence Will Transform the Role of Lawyers in the Delivery of Legal Services,* 82 FORDHAM L. REV. 3041 (2014).

103 *See, e.g.,* Frank Pasquale and Glyn Cashwell, *Prediction, Persuasion, and the Jurisprudence of Behaviourism*, 68 U. TORONTO L.J. 63 (2018); Citron and Pasquale, *supra* note __.

perficial to account for the range of nuance that is implicit in legal reasoning, which incorporates both "synoptic" and "subtle" judgments.[104] Put differently, legal decisions simultaneously implicate regular, observable patterns, as well as subtle shades of grey, and current NLP technology excels primarily at the former. The evidence in this paper broadly supports this supposition. In the context of a small sample and with little instruction, NLP proved remarkably good at identifying consistent patterns in the data. But while the models in this paper are a significant improvement on the state of the art, there remains a clear grey area: even with both text and metadata combined, the model still misclassified 25% of decisions. More training data will no doubt improve the accuracy of machine predictions,[105] but forecasting in this area will likely remain a human-guided process—albeit with machine assistance.[106]

Finally, a more troubling implication of natural language approaches to patent litigation lies in the future of patent prosecution. To the extent that future applicants are able to identify specific linguistic techniques which tend to find favor with courts—but which have little to do with the disclosure of information about the underlying innovation—we might be concerned about further divorcing extant patent law from its underlying social purpose. At this stage, such a concern is probably premature. Patent doctrine is a dynamic area of law, responding to constant changes in technology and social norms, such that particular patterns in patent language seem unlikely to be stable for long. However, to the extent that computational approaches are able to identify durable patterns in decision-making, it is probably more appropriate to view this as evolution rather than revolution. Like all legal pro-

---

104 Engstrom and Gelbach, *supra* note __ (distinguishing between the "synoptic" and "subtle" aspects of legal judgment, and arguing that machine learning is well suited to the former, but not to the latter).

105 Because patent law is a fairly dynamic area, the more successful applied uses of the model developed here would need to be regularly updated with contemporaneous decisions. Fortunately, this is an area in which comprehensive data collection is feasible. Patent litigation is relatively low volume, and case data are comprehensively collected by popular third-party analytics services such as Lex Machina.

106 In this regard I concur with Eric Talley, who argues that law is "irreducibly complex," and will continue to need "significant human input." *See* Eric Talley, *Is the Future of Law a Driverless Car? Assessing How the Data-Analytics Revolution Will Transform Legal Practice*, 174 J. INST. & THEORETICAL ECON. 183, 185 (2018).

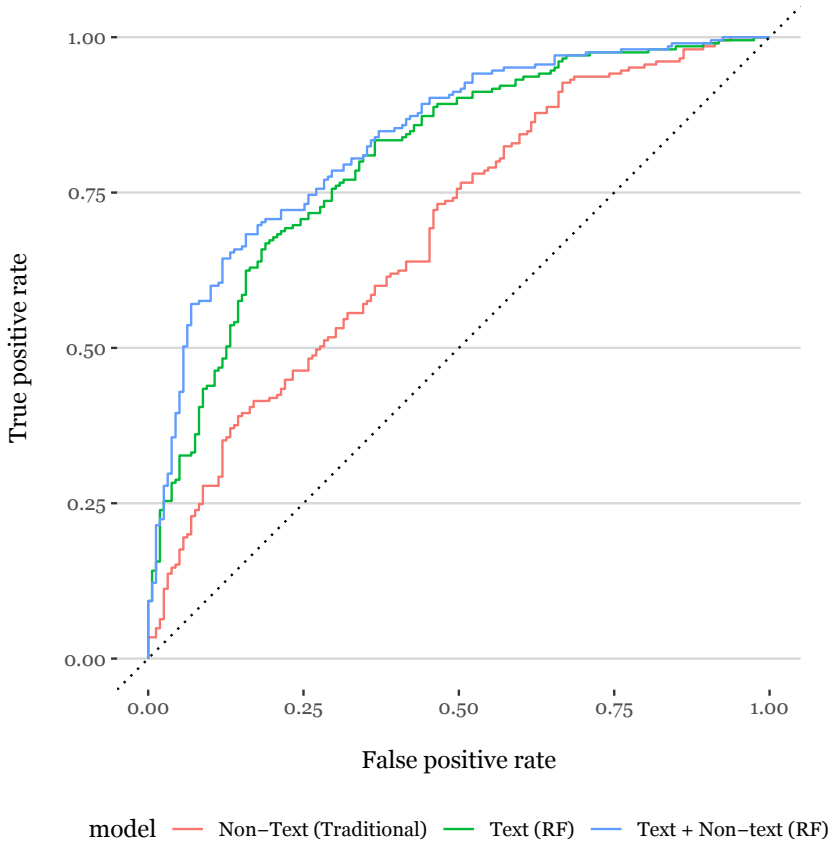
fessionals, patent prosecutors are, to some degree, pattern recognition experts. Claims drafting has already evolved into a highly specialized art, as patent attorneys respond to (and try to anticipate) Federal Circuit decisions on a range of esoteric but consequential questions of linguistic interpretation. Machine learning tools may accelerate this process, but they are ultimately different in capacity, rather than different in kind.

APPENDIX

A. Receiver operating characteristic (ROC) curves

Logistic regression and random forests, like many other classification models, generate a probability for every outcome, and then use this probability to assign a predicted class (in this case, "valid" or "invalid"). By default, most classifiers use 50% as the assignment cutoff, but depending on the balance of the underlying data—or cost of false positives/negative—it is often desirable to set a different threshold. An ROC curve shows the effect of varying this cutoff.

The ROC varies the threshold from 0 to 100, and at each level calculates the rate of "true positives" (in our case, correct prediction of invalidity) and "false positives" (incorrect predictions of invalidity). The dotted diagonal line is the baseline: every point on this line is equivalent to guessing the outcome with 50% probability. Curves that are closer to the top left corner indicate an algorithm that is more successful at discriminating between outcomes. (The AUC measure, described in Part IV.C., refers to the area under this curve.)



B. Separation graphs

The graph below shows the density of predictions for each model, split into valid and invalid patents. The blue curve shows the "true" invalid patents, while the red curve plots the "true" valid patents. The bottom axis shows the model's predicted probability of validity. The traditional model is notably worse at separating the valid and invalid decisions—a substantial fraction of the true invalid patents have a validity prediction greater than 50%.