Deep Learning, Text, and Patent Valuation

Po-Hsuan Hsu
National Tsing Hua University

Dokyun Lee
Carnegie Mellon University

Prasanna Tambe
Wharton School, U. Pennsylvania

David H. Hsu
Wharton School, U. Pennsylvania

November 16, 2020

Abstract

This paper uses deep learning and natural language processing (NLP) methods on the US patent corpus to evaluate their predictive power in estimating two measures of patent value: (i) investor reaction to patent announcements as measured in Kogan et al., 2017 and (ii) forward citations. While forward citations have traditionally been used as measures of economic value in the literature, their utility is mainly retrospective. Contemporaneous predictions of patent value, as embodied in investor reactions to patent grants, can be important for managers and policy-makers for prospective decision making. We compare the prediction performance of models using the structured features of the patent (number of citations, technology class, etc.) to deep learning and NLP methods. Relative to linear regression models using the same features, deep learning models reduce mean absolute error (MAE) by approximately 32%. Incorporating patent text further lowers the MAE by 13%.

Keywords: patent value; deep learning; natural language processing.

Acknowledgements. We are grateful for valuable research assistance from David Fan, Yan Gao, and Tong Zhou. We thank the Mack Institute for Innovation Management at the University of Pennsylvania for research funding.

Introduction

Using patents to measure corporate innovation has been important for advancing scholarship in economics and business.[1] Patent documents are valuable indicators of corporate innovation because they are thoroughly documented, standardized, and how they are constructed and recorded has legal implications.[2] Over the past 50 years, academic analysis of patents has principally relied on structured information from the patent such as technology domain (e.g. biotech, semiconductors) (Lerner, 1994), assignees (Marco et al., 2015), citations (Harhoff et al., 1999), claims (Lerner, 1994), inventors (Li et al., 2014), location (Jaffe et al., 1993), and other standardized fields. Recent work has only begun to derive new information from the structured data (e.g. inventor ethnicity and diversity) to further advance measurement capabilities (Singh and Fleming, 2010).

Much of this interest has been focused on predicting the impact or value of patents. The dominant method is application of traditional regression-based techniques on the structured features of patent data (scope, originality, classification, etc.) to estimate value (e.g. Lerner, 1994; Hirshleifer, Hsu, and Li, 2018). Within the last decade, however, deep learning prediction methods have been shown to out-perform these traditional methods for many prediction tasks, and especially those in which unstructured data, such as text, can play a role. Deep learning allows algorithms to mimic what human examination of the patent might reveal; it yields non-linear insights from the structured features, about how useful or valuable a patent might be, and it can perform the task at scale.

Deep learning may be especially promising for patent analysis because the massive quantities of unstructured patent text have been under-utilized in the social science literature. Patent filings are rich textual documents that describe an innovation, its scope, important mechanisms, what it protects, and other important details. According to Trajtenberg (1990, p. 173): "It has long been thought that the detailed information contained in the patent documents may have a bearing on the importance of the innovations

[1] The economics literature includes Kamien and Schwartz (1975), Griliches (1981, 1998), and Hall, Jaffe, and Trajtenberg (2001, 2005). For a more contemporary literature review, see Cohen (2010).
[2] As suggested by Griliches (1990), "[n]othing else even comes close in the quantity of available data, accessibility, and the potential industrial, organizational, and technological detail."

disclosed in them and that it may therefore be possible to construct patent indicators that could serve as proxies for the *value* of innovations." Although patent text has been available to researchers for decades, the scope and scale of the patent text have made it difficult to use for large-scale inferential analysis (there has been some limited work in the computer sciences literature, e.g. Hasan et al., 2009 and text analysis for social science insights is a rapidly expanding area, e.g. Gentzkow et al., 2019). There are millions of patents with thousands of words each and the language processing required to draw meaningful insights from patent-based text is demanding, not only because of its scale but also because of the sophistication of the language used in patent filings. Recently, textual patent analyses aim to predict patent similarity (e.g., Arts, et al. 2018; Whalen, et al. 2020).

Our objective in this paper is to evaluate the effectiveness of deep learning and natural language processing (NLP) methods in predicting patent value. Our deep learning models complement existing approaches, so to provide a baseline to evaluate their performance, we (i) start with a linear regression model based on structured features, (ii) test the performance of popular supervised machine learning models ("shallow" techniques) such as Ridge regression, Random Forest, and XGBoost that use the structured features of the patents, and (iii) use deep learning architectures as well as XGBoost (for comparison) that incorporate text. For the deep learning task, we use three convolutional neural net (CNN) layers, with 100 filters each, and with kernel sizes of 2, 3 and 4 applied in parallel. These filters extract local-level features in text. A bidirectional long short-term memory (LSTM) network with a hidden-layer size of 256 then processes the sequential local-level textual features further, which is then passed to a multi-layer perceptron (dense) layer of size 256 with a ReLU (Rectified Linear Unit) activation function to produce the final patent representation. Finally, to predict the objective measure, we use an additional linear layer which takes the structured features and text representation as inputs and outputs a single scalar value.

We apply this deep learning architecture to predict economic patent value. The number of forward patent citations received by focal patents is commonly used as a proxy for economic value in this literature (Office of Technology Assessment and Forecast, 1976; Trajtenberg, 1990; Harhoff, Narin, Scherer, and

Vopel, 1999).[3] However, recent research has found that stock market reaction to firm patent grants may be a superior correlate of firm growth (Kogan et al., 2017). Moreover, such investor reactions are forward-looking, whereas most traditional measures of patent value (such as forward patent citations, and patent renewal and litigation decisions) are backward-looking. Because managers and policy-makers likely value both perspectives depending on use-case, we employ both measures of patent value.[4] We do emphasize predicting the investor reaction-based Kogan et al. (2017) values, as a key advantage of this valuation method is that it is contemporaneous, though we show significant improvements in prediction using deep learning and NLP methods on both measures of patent value.

This paper makes two contributions to the literature. First, it evaluates the utility of deep learning and NLP methods for estimating patent value. These methods are evaluated against statistical methods that have been widely used in the literature. It compares several different approaches -- regression, traditional supervised learning with and without text, and deep learning with text -- to evaluate the incremental contribution of each for our ability to estimate patent value. In addition, using modern deep learning techniques in the innovation literature is itself relatively new, so the paper makes an interdisciplinary contribution. Nevertheless, these methods can be applied to other objectives or value estimates (e.g., licensing fees and royalties, traded patent prices, etc.).

Second, we demonstrate that these methods can be used to effectively estimate values for patents that cannot be valued in other ways. Using these methods raises the accuracy of predictions by more than 46% in comparison with linear regression methods based on a rich set of patent and corporate features. We also find that the full text of patent documents, after being processed by deep learning methods, can be converted into useful information that helps us to make more accurate patent value predictions. This is

[3] The Office of Technology Assessment and Forecast (1976) states, "if a single document is cited in numerous patents, the technology revealed in that document is apparently involved in many developmental efforts. Thus, the number of times a patent document is cited is a measure of its technological significance." Also, Harhoff, Narin, Scherer, and Vopel (1999) estimate that each (one) forward citation is worth one million dollars. Nevertheless, we also acknowledge issues related to the use of forward citations (Jaffe and de Rassenfosse, 2017).

[4] These measures are also publicly-available and widely used in the literature (e.g., Almeida et al., forthcoming). By contrast, patent value in contexts such as patent licensing and patent litigation are not widely observed and result from a highly selected process. In our empirical work, we drop patents with values in the top 1% to ensure that our estimation is not driven by outliers.

supported by the finding that our predictions based on the brief summary of patent documents outperform those based on claims (which define patent rights in litigation).

Literature Review

A group of prior studies examines patent valuation from the perspective of corporations, essentially imputing such value with the economic returns associated with the right to exclude in a particular domain. Taking advantage of corporate patenting and renewal data, Schankerman (1998) documents that the technology field, nationality of patent inventor, and patent application year are significant correlates of patent value. Ziedonis (2004) and Galasso and Schankerman (2010) find that the value of a patent is subject to its patent thicket, which is the fragmentation of patent rights measured by backward citations. As patent renewal and especially realized patent value is rarely observed, several papers use stock prices as instruments of market-perceived patent value. They show that R&D expenditure (Pakes, 1985; Cockburn and Griliches, 1988), backward citations of scientific studies (Deng et al., 1999), ratio of patents over R&D (Hall et al., 2005; Hirshleifer et al., 2013), diversity of backward citations (Hirshleifer et al., 2018), and innovation exploration or exploitation strategy (Fitzgerald et al., 2019) are factors that can affect patent value. Using the valuation of startup firms by venture capitalists as an instrument of patent value, Lerner (1994) shows the effect of patent scope, measured by the number of international patent classes (IPCs) in which a patent is assigned.[5]

Some papers have also confirmed that *ex post* measures, such as future infringement and renewal (Lanjouw, 1998; Lanjouw, Pakes and Putnam, 1998; Harhoff, Scherer and Vopel, 2003), number of forward citations (Harhoff et al,1999; Sampat and Ziedonis, 2005), and knowledge complementarity, measured by the ratio of forward non-self citations of a patent over forward non-self citations of all patents in a

[5] Other studies also use the university patenting and licensing data and analyze patent value from the perspective of universities. For example, Thursby and Kemp (2002), Thursby and Thursby (2002), and Lach and Schankerman (2008) show that inputs from within universities, such as faculty size and quality, size and age of technology transfer office, and royalty sharing incentives; and support from outside of universities, such as ownership, federal and industry sponsorship, improve patent licensing revenue (Siegel and Wright, 2015).

technology class (Galasso and Schankerman, 2010), are all associated with potential patent value. Finally, there are other ex post approaches to inferring patent value by examining patent rights reassignment in a market for intellectual property (Galasso, Schankerman and Serrano, 2013) and based on inventor surveys (e.g., Harhoff et al, 1999). A common denominator to all of the approaches listed in this section, however, is that there is a high degree of sample selection and voluntary disclosure, likely drawing from the more valuable part of the patent value distribution.

A wholly separate literature from the patent and innovation work, but related to our efforts, is the rapidly-developing field of text-analytics and machine learning with economic data. More broadly, the revolution in "big data" has led many to argue that the availability of high-frequency, granular data at scale has the potential to revolutionize some fields of economic inquiry (Einav and Levin, 2014). This may be particularly true of unstructured data, such as images, sounds, and text that are not conveniently encoded in a format that they can be included in statistical models, but nevertheless have enormous amounts of information content. The increased use of deep learning models, in particular, has made it possible to derive valuable information from these volumes of unstructured data.

A number of economic studies have begun utilizing text and text analytics to develop new insights in economics and social science. Hansen et al (2018) use natural language processing to study the effects of transparency on communications on deliberations issued by the Federal Reserve. Gentzkow et al (2019) apply computational methods to congressional speech and find that partisanship has been growing over time. NLP techniques have also been used in the patent domain, and in the paper that is perhaps closest to ours, Kelly, Papanikolaou, Seru, and Taddy (forthcoming) create a measure of patent importance from patent text similarity with existing patents and they identify "important" patents and show that the indices that they derive from them are able to capture waves of technological change over time. Raymond and Stern (2019) are also interested in predicting which patents will fall into the most selective tier of forward patent citations using textual data.

Our paper contributes to this growing field that uses volumes of text to understand economic phenomena, in this case, understanding drivers of innovation measured as patent value.

Key data sources and measures

Patent filings data and features

This analysis relies on several data sets. The first is the widely-used NBER patent database, which contains information on patents granted by the US Patent and Trademark Office ("USPTO"). A patent is a property right granted to an inventor, and is obtained by an inventor after filing a document with the USPTO. For use during the patent review process, this document includes all of the critical information about the patent, including the inventors, date of filing and date of grant, a description of the invention, what aspects of the invention should be protected by the patent, and other key fields. As such, these documents contain an enormous amount of information about R&D output and innovative activity that can be connected to specific organizations and individuals. Patent data and fields in the patent data can be viewed at https://www.patentsview.org/download/ at the USPTO Patentsview database.[6] These data are available for patents filed beginning in 1926.

Much of the research that uses patent data focuses on the question of whether it is possible to assess the impact of a patent (e.g. the number of times the patent is cited by later patents or alternatively, the economic value of the patent), given the information contained in the patent document, such as which technological subsection it is assigned to or the identity of the inventor or organization that filed the patent. Since we are mainly interested in predicting the impact of a patent, we only consider the information content of the patent that is known at its grant date (i.e., the date when the patent is officially assigned to its assignee).

The premise of this study is that most of the research in this area has used the *structured* patent fields to assess patent impact, i.e. categorical or numerical fields such as references (backward citations), assignee name and type, inventor name, patent grant date, and technology classification. However, the *unstructured* textual data that comprise most of the patent are also available (for US patents since 1976),

[6] The NBER patent database is a version of this data where the data have been cleaned and that has been used for many studies on strategy, innovation, finance, and economics (Hall, Jaffe, and Trajtenberg, 2001).

drawn from different sections such as the patent summary, the claims, and the description of the patent and this text is a potentially useful and unexplored asset for prediction tasks. This analysis uses both the structured and unstructured data from the patent documents disclosed to the public upon grant dates.

The structured fields in the patent document that we use in the analysis are meant to reflect those used in the existing literature. We discuss all patent-level variables in the following groups (detailed variable definitions are provided in Appendix A, and their summary statistics are reported in Appendix B):

a. Claims: Claims denote a series of statements that explicitly define the legal rights covered by a patent granted to the patenting organization. When courts adjudicate patent infringement cases, they only rely on claims. Thus, the descriptions of claims are very important to the patent owner in terms of both economic value and enforcement. A simple measure that has been used in prior research is the number of claims in the patent document (Lerner, 1994): more claims suggest greater coverage. Some research also analyzes the length of the first claim (Kuhn and Thompson, 2019): the shorter the first claim, the broader a patent covers, as shorter length suggests fewer qualifiers.

b. Technology classification: While there are a variety of different technology classification systems used across and within patent jurisdiction offices, we focus on the CPC (Cooperative Patent Classification system jointly developed by the European Patent Organization and the US Patent & Trademark Office) subsection code, which consists of one digit of letters and two digits of numbers.[7] One patent can be assigned to several subsection codes. In addition to positioning a patent in terms of technological property, we also consider a variable, scope, which is defined as the number of different subsection codes a patent is assigned with. This variable reflects how broad a patent covers in technology space (e.g., Lerner, 1994).

c. Backward citations: A patent document includes a list of references inserted by applicants and patent examiners. This list includes prior patents, reports, or any documents that are closely related to the patent. Thus, the list can be regarded as the "paper trail" of knowledge sources used to develop the patent

[7] https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html

(Jaffe, Trajtenberg, and Fogarty, 2000) and provides rich information about the technological inter-connections of different patents.[8] A simple measure based on backward citations is the number of backward citations made. In addition, we include the number and ratio of backward citations to basic science (such as journal articles or technical reports) because patents based to a greater extent on basic science are more important (Trajtenberg et al., 1997; Fleming and Sorenson, 2004).

We also consider the duration of citations between the focal patent's grant year and cited patents' grant years, which reflects the life cycle length of a technology (Trajtenberg et al., 1997). Moreover, we consider if the backward citations made by a patent are new or old knowledge to the patent owner and construct measures for exploration and exploitation (Benner and Tushman, 2002) and depth and scope (Katila and Ahuja, 2002). Using the information about backward citations and technology classifications, we also construct a patent originality variable, which measures the breadth of different technology classifications covered by backward citations made by a focal patent (Trajtenberg et al., 1997; Hirshleifer et al., 2018). We also combine the information about backward citations and the ownership of prior patents covered by these backward citations to examine the number and ratio of self-citations that reflect the specificity and redeployability of a patent (Lanjouw and Shankerman, 2004; Hoetker and Agarwal, 2007; Marx, Strumsky, and Fleming, 2009) as well as patent thickets that reflect the diversified ownership of prior patents cited by the focal patent (e.g., Ziedonis, 2004).

d. Family: When a patent is filed to foreign patent offices, it will have a patent family identifier that indicates how many other offices have registered the focal patent. Prior work has shown that a patent that has been filed overseas is more valuable (e.g., Hsu et al., forthcoming). In addition, the size of a patent family may also suggest the coverage of a bundle of patents.

e. Year of grant: The year in which the patent application was granted by the USPTO reflects when a patent owner receives legal protection for a patented technology.

[8] When patent X is listed in the references of patent Y, the literature has interpreted patent X as a knowledge source of patent Y with associated knowledge "flows" (Trajtenberg et al., 1997; Jaffe, Trajtenberg, and Fogarty, 2000).

f. Assignee: The USPTO also provides the information of the assignee of a patent. This field not only allows us to link patents to public firms, but also enables us to measure the patent thicket that reflects the fragmentation of patent ownership in commercialization of patents (Ziedonis, 2004).

Firm features

In addition to structured data in the patent document, we also consider an extensive list of firm characteristics that have been shown to influence patent values in the existing literature (detailed variable definitions are provided in Appendix A, and their summary statistics are provided in Appendix B). The first set is related firms' financial and accounting variables and includes R&D expenditure, advertising expenditure, capital expenditure, market capitalization in logarithm, the market to book ratio that reflects a firm's market opportunities, the ratio of property, plant, and equipment (PPE) to total assets that reflects asset tangibility, firm age, the industry classification, ROA, ROE, financial leverage, the ratio of cash flows to total debts that reflects a firm's liquidity, the ratio of cash holdings to total assets, industry concentration (i.e., the Herfindahl-Hirschman Index based on sales of all firms in one industry) that is an inverse indicator of industry competition, and the number of employees. These characteristic variables are collected from the Compustat/CRSP database.

The second set is related to firms' patent portfolio characteristics and includes the number of patents owned by a firm (i.e., patent portfolio size, see Galasso, Schankerman, and Serrano, 2013),[9] the originality score based on all patents granted to the firm, the number of inventors, the duration of backward citations of the firm's patent portfolio, the number of different technology classifications covered by patents of the firm that reflects the breadth of the firm's patent portfolio, the number and ratio of backward citations to basic science of the firm's patent portfolio, the originality, exploration, exploitation, self-citation, scope, and depth of a firm's patent portfolio, and patent thicket a firm faces.

Patent value

[9] Kamien and Schwartz (1975) survey the literature and note that "[n]evertheless, systematic study of patenting behavior has led Schmookler, Scherer, and others to conclude that the number of patents granted a firm is a usable proxy for inventive outputs."

The second data source used in the analysis is a measure of the economic value of a patent. Assessing patent value is a challenging task, and scholars have used different approaches to estimate the value of patents as was briefly surveyed above. We use two measures of patent value: i) forward citations and ii) market reactions to patent announcements.

A patent's number of forward citations denotes the number of citations it has received by subsequent patents that cite it, and has been commonly used as a proxy for the economic value of the patent (e.g., Trajtenberg, 1990; Harhoff et al., 1999; Hall et al., 2005). It is worth noting that patent value and forward citations are two correlated yet distinct measures for the economic value of a patent.[10] Patent renewal is another indicator of patent value because more valuable patents are more likely to be renewed (Schankerman, 1998). Moreover, patent litigation indicates patent value because more valuable patents are more likely litigated (Galasso and Schankerman, 2010). A critical limitation of these indicators is that they are *ex post* measures and cannot be observed by researchers immediately after a patent is granted (Kogan et al., 2017).

One recent approach that has been used by many follow-on studies and that does not suffer from this limitation is that used by Kogan et al (2017), who construct a measure of patent value by analyzing the market reaction around a patent announcement.[11] The key idea behind the construction of their value measure is that holding all other factors constant, the change in market value around a public patent announcement should reveal the net present value (NPV) of the patent rights granted to the firm. The authors show that the measures of patent value that they generate using this method contain information that have explanatory power beyond other measures of patent impact such as forward citations, patent renewal, and patent litigation, and that these measures are predictors of economic outcomes such as future productivity.

Using these methods, the authors compute estimates of the economic value of thousands of patents, and these patent value estimates have been made available for research. For our analysis, we accessed this

[10] Some studies, however, argue that it is inappropriate to use forward citations to measure a firm's patent value (Lerner and Seru, 2017; Jaffe and de Rassenfosse, 2017).
[11] The procedure has adjusted for stock return volatility, day-of-week fixed effect, and the firm-year joint fixed effects.

data from https://iu.app.box.com/v/patents on January 20, 2020. The data set used in the initial published paper included patents from 1976 through 2010. This data set gives us a measure of the value of a patent drawn from one class of decision makers (i.e. investor reactions).

Overview of approach

Our goal is to evaluate the extent to which deep learning and text processing can aid in the patent value prediction task. We first collect the information from all 1,335,177 patents granted to public firms in our sample period which spans the years 1976 to 2010. We drop variables with missing data and remove the top 1% by patent value (due to heavy skew) and we are left with 1,200,333 rows.[12] We then randomly train-test split the data to obtain 1,110,333 patents to be used as the training sample and the remaining 90,000 to be used as the test sample. As shown in Appendix B, the mean patent value in our training sample is USD 9.46 million, and the mean patent value in the test sample is USD 9.53 million.

To provide consistent baselines throughout the analysis, we focus on three categories of models for performance comparison. For each, we convert patent values to logs and fit the logarithmic patent value to features in the training sample. Then, we use the trained model to predict the logarithmic patent value.

The three classes of models we evaluate are described below: First, we replicate models used in the existing literature on patent value. These principally rely on OLS regressions using the structured numerical and categorical features in patent documents. This literature analyzes the economic value and technological merits of patents and can be traced back to the log-linearized production function of innovation used by Griliches (1981, 1988) and Kortum and Lerner (1998). Second, we expand the class of models to supervised machine learning models, including both shallow and deep learning models, but continue to constrain our feature set to those derived from the structured patent data. These models include some that are non-parametric and that allow for interaction effects among the structured features, so although they use only structured data, they impose different tradeoffs than the linear regression-based models. Finally, we expand the feature set by using text and NLP methods in the deep learning models. When incorporating

[12] In the training sample, the patent value in the 99th percentile is USD 146 million, which is more than 10 times the sample mean of USD 12.14 million before the truncation at the 99th percentile.

unstructured patent text, we take advantage of recent advances in multi-view (i.e. able to process multiple modalities such as text and structured data) deep learning models augmented with pre-trained embedding models (e.g., GloVe, FastText), which are aware of semantic similarities and linguistic statistical structures and enable these models to better capture textual signals.[13] This step is "supervised" (given data x and label y, find a function f such that f(x)=y) because the objective function is specified (market reaction or forward citations). Additionally, we also ran another supervised learning model that incorporates feature-engineered patent text to compare against deep learning models that incorporates textual data natively.

To assess the performance of each of these models when predicting patent value or impact (i.e. market value or forward citations), we use the mean absolute error (MAE) loss metric. Mean absolute error is defined as the average absolute distance between each predicted patent value and its "true" value from the labels assigned by Kogan et al. (2017). Relative to other error metrics, such as mean squared error, MAE minimizes penalties imposed by outliers. Although this itself offers benefits due to the skewed nature of the patent data, we choose the MAE metric because it is the most common loss metric used when evaluating deep learning models.

Main results

A description of the models used in this section, the rationale for the modeling choices we make, and the performance of each of the models are described in detail below. For convenience, they are also summarized in Table 1.

A. Linear regression model

Almost all of the literature to date has used structured patent features (e.g. originality, technological classification) to predict proxies of patent value, such as forward citations and patent renewal. One of the few exceptions is Kuhn and Thompson (2019) who use word counts in the first independent patent claim as an indicator of value. All of these models share a common goal of using linear regression models to estimate the importance of various structured patent features in predicting patent value.

[13] All of the text models are run on GPU workstations and servers. We use both the *tensorflow* and *pytorch* machine learning frameworks to execute the deep learning models and use *spaCy* for tokenization and text cleaning.

The starting point of our analysis is to replicate common specifications from the patent valuation literature to demonstrate that the sample we use in the rest of our analysis behaves as expected, given what we know from prior work in this area. The first model we test follows Hall (1993) and Hall et al. (2005) and can be specified as:

$$ln(Xi + 1) = \alpha Zi + \beta Wj + \gamma j + \theta k + \lambda h + \boldsymbol{\delta} t + \varepsilon i \qquad (1)$$

We estimate equation (1) on all patents in our training sample to obtain coefficients (including those on fixed effects included in the model). We then use these coefficient estimates to predict the value of patents in the test sample. In this specification, *i* indexes the patent granted in technology subsection *k* to firm *j* in industry *h* in year *t*, the dependent variable is the value of the patent denoted as X_i, Z_i is a vector of structured patent features,[14] W_j is a vector of structured firm features,[15] γ_j is a vector of firm or industry fixed-effects (omitted from some models), θk is a vector of fixed-effects for technology subsections, λh is a vector of fixed-effects for industries, $\boldsymbol{\delta} t$ is a vector of year fixed effects for the grant year, and εi is the unmodeled error term. Each row in this regression corresponds to one patent. In Appendix B, we present summary statistics for all variables used in our estimation of Equation (1) in the training sample and in the test sample.

Table 2 presents the results from estimating equation (1) using the training sample. In Model (1), we only consider patent characteristics including year fixed effects and subsection fixed effects (for

[14] The vector *Z* includes an extensive list of structured patent features drawn from the existing literature in this area: the number of claims, the length of the first claim, the length of all claims, the number of backward citations, the length of the brief of a patent, the number and ratio of backward citations that are new knowledge source, the originality index, the number and ratio of backward citations to basic science, the duration of backward citations, the half-life of backward citations, the ratio of backward citations within 5 years, exploration, exploitation, depth, scope, breadth of backward citations, breadth of subsection, the number of self-citation, an indicator for non-self-citation, patent thicket, the affiliation of an international patent family, the number of patents in the same family, and the number of U.S. patents in the same family.

[15] The vector *W* includes an extensive list of structured firm characteristics drawn from the existing literature in this area: R&D expenditures in log, the ratio of R&D to total assets, advertising expenditure in log, the ratio of advertising to total assets, the ratio of capital expenditure to total assets, market capitalization in logarithm, market-to-book ratio, the ratio of property, plant, and equipment to total assets (asset tangibility), firm age, ROA, ROE, financial leverage, the ratio of cash flows to total debts, the ratio of cash holdings to total assets, industry concentration in sales, the number of employees, an indicator variable for missing employees, the number of inventors, the average and sum of originality scores of all patents granted to a firm, the duration and half-life of backward citations, the maximum patent-level breadth, the average number and ratio of backward citations to basic science, breadth of backward citations, the number of patents granted (raw number and scaled by total assets), the number of citations received (raw number and scaled by total assets), the patent thicket, breadth of subsections, backward citation-based originality, exploration, exploitation, scope, and depth of a firm's patent portfolio.

technology classification) in Model (1). We find that the R-squared is 16.8%, suggesting that patent characteristics explain up to 17% of the total variation in patent value in the training sample. The MAE of Model (1) for the test sample, as shown in Table 1, is 8.64. It is noteworthy that we use logarithmic values as regression dependent variables, but then exponentiate each predicted value to convert back to a USD-denominated value; thus the MAE of 8.64 suggests the predictions made by Model (1) can deviate from the real value by USD 8.64 million on average.

In Model (2), we consider four sets of fixed effects: year, subsection (for technology classification), firm (based on PERMNO), and industry fixed effects (based on Fama-French 48 industries). We find that the R-squared value of Model (2) is as high as 79.3%, suggesting that these features can explain variation in patent values in the training sample to a great extent. The R-squared value from this regression indicates that fixed-effects for the patenting organization matter a great deal for explaining patent value, and in fact can explain a significant amount of variation in investor reaction to patent announcements. The MAE of Model (2) in the test sample reaches a low of 7.58 as shown in Table 1, which is much smaller than the MAE of Model (1). This result suggests that the identity of the organization filing the patent (captured by firm and industry fixed effects) plays an important role in predicting investor reaction to the patent. In addition, the predictions made by Model (2) can deviate from the real value by USD 7.58 million on average.

Lastly, in Model (3), we take an extensive list of firm characteristics into account and only consider two sets of fixed effects: grant year fixed effects and subsection fixed effects (for technology classification). We find that the model delivers an R-squared of 79.5% in the training sample and MAE of 7.02 in the test sample. Model (4) is the full model - it includes all patent and firm characteristics and all four sets of fixed effects (year, subsection, firm, and industry fixed effects). It delivers an R-squared of 86.3% (which is the highest among all models) in the training sample and an MAE of 5.95 (which is the lowest of all models). This suggests that the prediction made by Model (4) can deviate from the real value by USD 5.95 million on average.

B. Machine learning models using the structured feature set

The next class of models we consider continues to use structured features from the patent data but it also considers machine learning models. Specifically, we test the performance of the following models, which represent some of the most commonly used supervised models: i) Ridge regression, ii) Random forest models, and iii) Gradient boosted trees. Given the use of structured information from the patent, these models are straightforward to implement when using modern statistical packages.

Table 3 reports the results from using supervised learning models to predict patent value. The first set of models use Ridge regression. Ridge regression models are a class of estimators that "shrink" outliers to the sample mean. They offer some potential advantages over the use of linear regression in a context such as this one, where the dependent variable, patent value, is highly skewed, and where there are a very large number of independent variables. Applying Ridge regression models appears to perform better at the prediction task than the linear regression models, particularly in the models that include firm fixed-effects and therefore have a large number of independent variables. The Ridge regression model with a full set of firm fixed-effects generates an MAE value of only about 4.4 in the test sample, which is a significant reduction when compared with the linear models that we have applied until this point.

The next model for which we present results is a Random Forest (RForest) model, which is an ensemble learning technique known to perform well in a variety of applications, and works by constructing a number of different regression trees that fit the model and takes the mean of the predicted regression values outputted by the different models. In our patent-based application, as well, it appears that Random Forest works relatively well when compared to other models, with the model that includes firm fixed-effects producing an MAE value of 4.02 in the test sample, which is a lower error value than any of the models used so far, including the Ridge regression model.

The final two types of supervised machine learning models used for this application are XGBoost, another ensemble learning method which relies on Gradient boosting, and a Feed-forward Neural Network

(FF Neural Network). With firm fixed-effects, these two models also perform better than linear regression but not as well as the Ridge regression model or the Random Forest models.

The key conclusion from this set of tests is that the performance of these supervised machine learning models is superior to linear regression. This is not surprising, as these models allow for interaction effects between variables (which we do not include in the linear models) and the non-parametric approaches may respond better to the distributional difficulties presented by skewed value data. Even without introducing text, some of the most commonly used supervised learning models reduce MAE by as much as 32%.

In the next section, we 1) apply deep learning methods and 2) add textual data to investigate how much further this MAE metric can be reduced.

C. Deep learning approaches

The deep learning, neural network-based models we use are constructed in the following way:

1. Structured and categorical features are mapped to vectors using an embedding layer for each feature of a specified size.
2. The embeddings generated for all of the categorical features and the numerical features are concatenated to produce a single vector for each patent.
3. This vector is then passed through a 3-layer perceptron of size 128 using a ReLU (Rectified Linear Unit) activation function to arrive at a combined representation.
4. To predict $\log(Xi+1)$, we use a linear layer which takes the combined representation and outputs a single scalar.

The steps described above are summarized in Figure 1. We also provide more technical details of our application of deep learning to patent text in Appendix C.

In Table 4, the first column indexes the model number, the second column indicates whether structured or structured plus text data has been used, the third column indicates how the patent data is converted to features, the fourth column describes the type of predictive model used in the test, and the fifth column

reports the performance metric, which is the mean absolute error (MAE) of the model when it is run on the test sample.

A modeling choice we face when introducing text into our models is which section of the patent text to focus on. Patent documents are characterized by a number of different sections of text -- abstract, claims, summary -- and they can differ in terms of the types of language they contain and in their relative importance to the protection claim.

First, we use only the brief summary descriptions, with stop words[16] removed and with the length of the text capped at 3,000 words. On average, a brief summary description had 855 words after removing stop words with a standard deviation of 1,094. At the 99th percentile, there were 4,740 words after removing stop words. To generate the text only model, we take the following steps:

1. First, we create a vocabulary of tokens that appear at least ten times each in the training corpus. In other words, tokens that appear nine times or less across the entire corpus (and are therefore likely to be unique to a single filing or small group of filings) are not included in the training set. For these other tokens (those that appear less than ten times), we create a special <UNK> token (i.e. a token that specifies an unknown word). All words are initialized with their GLoVe vector if available. GLoVe is a vector-based representation of words that retains semantic-similarity information suitable for using deep learning approaches on text (for background on GloVe vectors, see Pennington et al 2014). Otherwise, words are initialized with random values. These word embeddings are trainable and have one hundred dimensions.
2. A dropout value of 0.50 is applied to each of the word embeddings. Dropouts are a method wherein some subset of nodes in a neural network are ignored during passes through the training phase, thereby mitigating problems related to overfitting.
3. Three convolutional neural net (CNN) layers, with 100 filters each, and with kernel sizes of 2, 3, and 4 are applied in parallel, and their outputs are concatenated at each time step.

[16] "Stop words" are words such as "and", "the", and "but" that generally have little informational content and are removed from the text corpus during the pre-processing stage.

4. A MaxPool layer of kernel size 3 and stride 2 is used to reduce the number of time steps.
5. A bidirectional long short-term memory (LSTM) network with a hidden-size of 256 processes the output of these prior stages, and we use the output hidden state of the LSTM at the final time step. LSTM structure is responsible for remembering and keeping track of local-level features extracted by CNN layers throughout long text.
6. A multi-layer perceptron (dense) layer of size 256 with a ReLU (Rectified Linear Unit) activation function is used on top of the LSTM output to produce a final text representation.
7. Finally, to predict $\log(Xi+1)$, we use an additional linear layer which takes the text representation as input and outputs a single scalar.

Steps for the text-only model are summarized in Figure 2.

We test the performance of models that use the text content of the patent documents by itself, as well as in conjunction with the structured features already discussed above. The model that incorporates both patent text and structured features works like the above model for the structured data and text data independently, up until the last dense layer. The difference is that the text representation and the numerical + categorical representation are concatenated with one another into a vector and then $\log(Xi+1)$ is generated. This model, using text and structured features of the patent, is summarized in Figure 3. Lastly, we also run XGBoost, a widely used top-performing algorithm, along with manual feature-engineered text variables such as bigrams and TF-IDF (term frequency–inverse document frequency) weighted unigrams to compare against deep learning models that utilize both structured and text data.

The first row in Table 4 only uses structured features. We do not remove the top 1% of patents by value. Using neural networks with the structured features while retaining firm fixed-effects produces a large drop in the MAE metric, which falls to 5.98 in the test sample. The key difference between this neural network model and the linear regressions used above is that the neural network is not restricted to a linear combination of model features. It can generate new features based on non-linear interactions between the existing features in the model. In the second row in Table 4, we use the same neural network method but remove the top 1% of patents by value. We find that the MAE drops to 4.08 in the test sample, which is as

low as the MAE of the best model in Table 3 and is much smaller than the MAE of the best linear regression model (5.95).

Row 3 onwards in Table 4 begins to introduce textual data into the analysis, and uses a neural network to use text input with or without structured data to predict patent value. We first introduce text from the patent brief summary, which is the section of the document intended to be a brief description of the invention. Therefore, the summary section of the patent document should encapsulate many of the key differentiating features of the invention. We also utilize claims text. The claims text from a patent filing, in contrast to the summary, describes the scope of the technical protection granted by the patent. To convert the text content of the summary into features for the neural networks model, "stop words" are removed from the summary text using the spaCy library. This is a pre-fixed set of common words with little information content that are not used by the predictive algorithm. Then, we restrict the text processing model to the first 3,000 tokens (i.e. words) after removing stop words and representing each word with either GloVe or Fasttext embedding.

Rows 3-6 present models using only textual features (brief summary or claims). The MAE rises back up to the range of 6.86-9.64 depending on modeling choice. There is a significant loss of information when using text-only models. The results suggest that, for this data, the brief summary seems to have higher signal compared to the claims text in predicting the output value. We speculate that this is due to the nature of output value relying on short-term market reactions and the nature of brief summary content which may include contemporaneous or contextual information. In addition, while claims have legal implications, their information may not be as technology-relevant as that contained in the brief summary. Removing the top 1% by value also helps with performance.

Rows 7-10 present models using both structured data and text data. Model 9, which uses FastText embedding of brief summary text along with CNN-LSTM neural net achieves the best MAE of 3.26 (Row 9). FastText (Bojanowski et al., 2017) is another word embedding technique that processes words at the character level, and is thus more efficient in some respects. This suggests that text does introduce new information; using both the structured features and the patent text together is a more accurate predictive

model than using either of these alone. In these models, the brief summary seems to carry more informative signals again. When adding both brief summary and claims text, the model seems to get confused due to noise, which might be overcome with extensive fine tuning.

Lastly, rows 12 and 13 use XGBoost along with feature-engineered textual attributes to benchmark against the deep learning models that can natively handle both structured and text data. The MAE performance is 5.25 and 5.32, falling short of deep learning approaches.

The best performing model in this table is in Row 9, which uses both the categorical structured features that have been used in many patent value studies and introduces the text from the summary document. We conjecture that contextualized embedding models (e.g., BERT by Devlin et al 2018 or ELMO by Peters et al 2018) that can learn more specific linguistic structures of patent text may be applied to increase performance even further.

A second takeaway from this set of tests is that text is informative when predicting patent-based outcome features such as value. Using structured features remains critical, and provides a great deal of predictive information. The summary text, which briefly outlines the subject matter of the patent and may therefore contain the highest density of keywords has superior predictive power compared to the claims text.

Application to forward citations

The majority of prior studies on patent value rely on the number of forward citations received (Office of Technology Assessment and Forecast, 1976; Trajtenberg, 1990; Harhoff, Narin, Scherer, and Vopel, 1999). It is an important and meaningful extension to apply our methods and comparisons to the number of forward citations received by a patent. We first calculate the number of three-year forward citations of all patents in our sample.[17] We use the sample period 2003 to 2017. We then use the logarithmic value of the forward citations plus one as the dependent variable in Equation (1) and re-estimate the best

[17] For a focal patent, its number of three-year forward citations is defined as the number of subsequent patents that cite the focal patent and are granted within a three-year period since the focal patent's grant date. Since our sample ends with patents granted in 2017, we collect the latest citation data from the Patentsview database in August 2020 to maximize the number for forward citations we are able to count.

linear model (i.e., Model (4) in Table 1) to train the model using the training sample. Similar to our earlier exercise, we then use the out-of-sample set to calculate the MAE between the predicted value and real value of forward citations. We find that the linear regression delivers an MAE of 1.74 (shown in Table 5), suggesting that the prediction deviates from the actual forward citations by 1.74 citations.

Using neural networks rather than linear regression for this task, along with only the structured data, only slightly lowers this number, such that the deviation in the prediction of forward citations falls to about 1.6, which is an 8% improvement.

The prediction results from incorporating the patent text along with the structured features fall in between the two. The MAE from using these inputs into a neural network model is 1.61. As we have seen in the case of adding both the claims and brief summary reducing performance due to noise, we see again that adding the structured data and claims text performed slightly worse than using just structured data.

These results indicate that deep-learning methods perform slightly better in forecasting forward citations in comparison with traditional regression estimation but the improvement is not as large as it is when predicting value. Nevertheless, this finding highlights the possibility of applying our methods to other non-pecuniary measures of the effects of patentable innovations.

In addition, to reflect the fact that a large portion of patents do not receive forward citations, we present histograms for the distribution of forward citations and the predicted values of three methods (linear regressions, neural networks without patent text, and neural networks with patent text) in Figure 4. We find that 45% of patents receive zero forward citation. The mass of zeros cannot be matched by linear regressions and neural networks without patent text, but can be well matched by neural networks with patent text (which predicts 43% of zero forward citations). In fact, whenever a patent's forward citation is zero, the neural networks with the patent text method can correctly predict zero with a probability of 46%. This supports the information advantage of patent text from a new perspective.

Application to patents from newly-public firms

We would like to investigate whether our prediction models perform well in a context in which there may be comparatively less information about patent value. Ideally, we would investigate patent values

in privately-held firms, but unfortunately, our ability to train the prediction models on "ground truth" patent valuations is limited to contexts which are likely severely selected (e.g., patent litigation or licensing). We thus extend our analysis to focus on the patents of firms that are newly-public because an objective benchmark, market reaction, is available to us for these patents. Another motivation is to test the extent to which our main results might be driven by established firms, some of which receive hundreds or thousands of patents per year. We evaluate how the deep learning methods described above compare against existing statistical methods.

We first train our regression and deep-learning methods using all patents granted to firms that had *not* recently gone public in a recent sample period (2003-2017). We then use the trained model to predict the values of patents assigned to newly public firms. We define whether a patent is from a newly public firm in two different ways: 1) whether its grant date is within a two-year window from the firm's IPO date, or 2) whether the patent grant date is within a slightly broader three-year window from the firm's IPO date. Similar to our earlier analysis, we truncate patents in the top 1% of patent value.

When we apply the trained models to predict patent values, we find that the regression model delivers MAE values of 20.10 and 19.61 for the two- and three-year IPO windows, respectively (the first row in two panels of Table 6).[18] These numbers are higher than the MAE of 8.64 of Model (1) in Table 1 for all public firms' patents, which is reasonable because the patents of newly public firms are likely different from those of established firms.

The next two rows in Panel A of Table 6 show the results from applying a deep learning approach to the patent document information from this sample of firms. When using deep learning along with only the structured features, the MAE drops to about 14.62. Further incorporating text further lowers the MAE value to about 10.58. By comparison, this is better than linear regression models for established firms when using no firm information. The same pattern is found in Panel B of Table 6.

[18] In addition, the two values of MAE are larger than their counterparts in Table 1 because we use the new/updated patent value of Kogan et al. (2017) and a different sample period 2003-2017.

The analyses we perform in this subsection are notable for the following reasons: first, we demonstrate the possibility of applying our methods to entities that are not publicly listed, such as private firms, universities, and research labs. Second, we provide further evidence for the importance of text-based information and deep-learning methods in predicting patent value. Finally, we observe much less accurate predictions of patent values when we do not utilize firm-level information, which supports a long belief in the literature: the private values of a patent depend on the synergies from all functions of an organization/firm. It is thus very important for researchers to collect firm-related information when they attempt to evaluate the values of patents.

Conclusions

Our study applies deep learning to patent text to predict patent value as measured by forward citations and market reactions to patent announcements. Valuing patents is important for a number of reasons, ranging from understanding the value of a firm's (intangible) assets to understanding investment decisions as well as aggregate innovation rates and directions in different economic sectors and regions. There has been substantial academic interest in predicting the impact of innovation from patent information, and the results presented in this paper suggest that incorporating text into the statistical methods most frequently used to predict patent value substantially improves predictive power. Specifically, applying deep learning to patent text improves our ability to predict patent value by about 60% relative to a baseline that uses only structured features with linear regression models. About two-thirds of this improvement comes from the application of deep learning, and the remaining third from using the patent text.

This research is intended to bridge a quickly expanding literature on text analysis methods for social science research with a large and established literature on patents and innovation economics. Due in part to the rich and extensive data on patent filings and grants, patents have been used to answer dozens of questions related to innovation, corporate strategy, geography, and investment. This paper contributes to an emerging literature suggesting that the text content of patent documents, which has largely been absent from these lines of patent-based inquiry thus far, can make valuable contributions to our understanding of patents and innovation, and perhaps open up some new areas of research.

There are, of course, several limitations to our approach. One key limitation is that the measurement of patent value -- the labels we use to train our model -- is itself noisy. There are a variety of ways to think about measuring the impact of a patent, and they each have strengths and weaknesses. Our predictions are only as good as the quality of the labeled measures that we are trying to predict. Relatedly, our analytic framework rests on patent documents, but the prior literature has discussed limitations of patents as measures of innovation, such as differences in patenting propensity across fields and time. Furthermore, our approach in this paper has been to include some of the more recent innovations in natural language processing, but this field is rapidly evolving. Innovations in this field should be continuously incorporated into this domain to further improve our ability to understand drivers of innovation as reflected in patenting activity. Despite these caveats, our hope is that this and other recent efforts at the intersection of the fields of "data science" and "patents as indicators of innovation" will propel both fields forward.

References

Almeida, H., Hsu, P. H., Li, D., & Tseng, K. (forthcoming). More cash, less innovation: The effect of the American Jobs Creation Act on patent value. *Journal of Financial and Quantitative Analysis*.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Arts, S., Cassiman, B., & Gomez, J. C. (2018). Text matching to measure patent similarity. *Strategic Management Journal*, 39(1), 62-84.

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. In *Neural networks: Tricks of the trade* (pp. 437-478). Springer, Berlin, Heidelberg.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135-146.

Cockburn, I., & Griliches, Z. (1988). Industry Effects and Appropriability Measures in the Stock Market's Valuation of R&D and Patents. *The American Economic Review*, 78(2), 419-423.

Cohen, W. (2010). "Fifty years of empirical studies of innovative activity and performance," in Handbook of the Economics of Innovation Vol. 1, B. H. Hall and N. Rosenberg, eds., pp. 129-213.

Deng, Z., Lev, B., & Narin, F. (1999). Science and technology as predictors of stock performance. *Financial Analysts Journal*, 55(3), 20-32.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

Einav, L., & Levin, J. (2014). Economics in the age of big data. *Science*, *346*(6210), 1243089.

Fitzgerald, T., Balsmeier, B., Fleming, L., & Manso, G. (2019). Innovation search strategy and predictable returns. *Management Science*, forthcoming.

Galasso, A., & Schankerman, M. (2010). Patent thickets, courts, and the market for innovation. *The RAND Journal of Economics*, 41(3), 472-503.

Galasso, A., Schankerman, M., & Serrano C. J. (2013). Trading and enforcing patent rights. *The RAND Journal of Economics*, 44(2), 275-312.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-74.

Griliches, Z. (1981). Market value, R&D, and patents. *Economics Letters,* 7, 183-167.

Griliches, Z. (1990). Patents Statistics as Economic Indicators: A survey Journal of Economic Literature, 18 (4). December, 1661, 1707.

Griliches, Z. (1998). Patent statistics as economic indicators: A survey. In R&D and Productivity: The Econometric Evidence, Z. Griliches, editor, University of Chicago Press, 287-343.

Gentzkow, M., Shapiro, J. M., & Taddy, M. (2019). Measuring Group Differences in High-Dimensional Choices: Method and Application to Congressional Speech. *Econometrica*, *87*(4), 1307-1340.

Guellec, D., & de la Potterie, B. V. P. (2000). Applications, grants and the value of patent. *Economics letters*, *69*(1), 109-114.

Hall, Bronwyn H., 1993. The stock market's valuation of R&D investment in the 1980's. *American Economic Review* 83, 259-264.

Hall, B. H., Jaffe, A. B., & Trajtenberg, M. (2001). *The NBER patent citation data file: Lessons, insights and methodological tools* (No. w8498). National Bureau of Economic Research.

Hall, B. H., Jaffe, A., & Trajtenberg, M. (2005). Market value and patent citations. *RAND Journal of economics*, 16-38.

Hansen, S., McMahon, M., & Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. *The Quarterly Journal of Economics*, *133*(2), 801-870.

Harhoff, D., Narin, F., Scherer, F. M., & Vopel, K. (1999). Citation frequency and the value of patented inventions. *Review of Economics and Statistics*, *81*(3), 511-515.

Harhoff, D., Scherer, F. M., Vopel, K. (2003). Citations, family size, opposition and the value of patent rights. *Research Policy*, 32(8): 1343-1363.

Hasan, M. A., Spangler, W. S., Griffin, T., & Alba, A. (2009, June). Coa: Finding novel patents through text analysis. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1175-1184).

Hirshleifer, D., Hsu, P. H., & Li, D. (2013). Innovative efficiency and stock returns. *Journal of Financial Economics*, 107(3), 632-654.

Hirshleifer, D., Hsu, P. H., & Li, D. (2018). Innovative originality, profitability, and stock returns. *The Review of Financial Studies*, 31(7), 2553-2605.

Hoetker, G., & Agarwal, R. (2007). Death hurts, but it isn't fatal: The postexit diffusion of knowledge created by innovative companies. Academy of Management Journal, 50(2), 446-467.
Hsu, D. H., Hsu, P. H., Zhou, T., & Ziedonis, A. A. (forthcoming). Benchmarking US University Patent Value and Commercialization Efforts: A New Approach. *Research Policy*.
Jaffe, A. B., and G. de Rassenfosse. "Patent Citation Data in Social Science Research: Overview and Best Practices." Journal of the Association for Information Science and Technology, 68 (2017), 1360–1374.
Jaffe, A. B., Trajtenberg, M., & Fogarty, M. S. (2000). Knowledge spillovers and patent citations: Evidence from a survey of inventors. *American Economic Review*, 90(2), 215-218.
Jaffe, A. B., Trajtenberg, M., Henderson, R. (1993). Geographic localization of knowledge spillovers as evidenced by patent citations, *Quarterly Journal of Economics*, 108(3), 577-598.
Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., & Mikolov, T. (2016). Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.
Kamien, M. I. & Schwartz, N. L. (1975). Market structure and innovation: A survey. Journal of Economic Literature 13, 1-37.
Kelly, B., Papanikolaou, D., Seru, A., & Taddy, M. (forthcoming). Measuring technological innovation over the long run. American Economic Review Insights.
Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological innovation, resource allocation, and growth. *The Quarterly Journal of Economics*, *132*(2), 665-712.
Kortum, S., & Lerner, J. (1998, June). Stronger protection or technological revolution: what is behind the recent surge in patenting? In Carnegie-Rochester Conference Series on Public Policy (Vol. 48, pp. 247-304). North-Holland.
Kuhn, J. M. & Thompson, N. C. (2019). How to measure and draw causal inferences with patent scope. *International Journal of the Economics of Business*. 26(1), 5-38.
Lach, S., & Schankerman, M. (2008). Incentives and inventions in universities. *The RAND Journal of Economics*, 39(2), 403-433.
Lanjouw, J. O. (1998). Patent protection in the shadow of infringement: Simulation estimations of patent value. *The Review of Economic Studies*, 65(4), 671-710.
Lanjouw, J. O., Pakes, A., Putnam, J. (1998). How to count patents and value intellectual property: the uses of patent renewal and application data. *Journal of Industrial Economics*, 46(4), 405-432.
Lerner, J. (1994). The importance of patent scope: an empirical analysis. *The RAND Journal of Economics*, 319-333.
Li, G. C., Lai, R., D'Amour, A., Doolin, D. M., Sun, Y., Torvik, V. I., ... & Fleming, L. (2014). Disambiguation and co-authorship networks of the US patent inventor database (1975–2010). *Research Policy*, 43(6), 941-955.
Marco, A. C., Myers, A., Graham, S. J., D'Agostino, P., & Apple, K. (2015). The USPTO patent assignment dataset: Descriptions and analysis.
Marx, M., Strumsky, D., & Fleming, L. (2009). Mobility, skills, and the Michigan non-compete experiment. Management science, 55(6), 875-889.
Office of Technology Assessment and Forecast (1976). U.S. DEPARTMENT OF COMMERCE, PATENT, AND TRADEMARK OFFICES. Sixth Report. Washington D.C.: U.S. Government Printing Office, 1976.
Pakes, A. (1985). On patents, R & D, and the stock market rate of return. *Journal of Political Economy*, 93(2), 390-409.
Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv preprint arXiv:1802.05365.
Pennington, J., Socher, R., & Manning, C. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).
Raymond, L., Stern, S. (2019). Predicting patent impact: A machine learning approach" working paper. MIT.
Sampat, B. N., & Ziedonis, A. A. (2004). Patent citations and the economic value of patents. In Handbook of quantitative science and technology research (pp. 277-298). Springer, Dordrecht.
Schankerman, M. (1998). How valuable is patent protection? Estimates by technology field. *The RAND Journal of Economics*, 29(1), 77-107.
Siegel, D.S., Wright, M. (2015). University technology transfer offices, licensing, and startups. In: Link, A., Siegel, D, Wright, M. (Eds.), The Chicago Handbook of University Technology Transfer and Academic Entrepreneurship, pp. 1–40.
Singh, J., & Fleming, L. (2010). Lone inventors as sources of breakthroughs: Myth or reality? *Management Science*, 56(1), 41-56.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. *The Rand Journal of Economics*, 21(1), 172-187.
Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. Economics of Innovation and new technology, 5(1), 19-50.
Thursby, J. G., & Kemp, S. (2002). Growth and productive efficiency of university intellectual property licensing. *Research Policy*, 31(1), 109-124.
Thursby, J. G., & Thursby, M. C. (2002). Who is selling the ivory tower? Sources of growth in university licensing. *Management Science*, 48(1), 90-104.
Whalen, R., Lungeanu, A., DeChurch, L., & Contractor, N. (2020). Patent similarity data and innovation metrics. *Journal of Empirical Legal Studies*, 17(3), 615-639.
Ziedonis, R. H. (2004). Don't fence me in: Fragmented markets for technology and the patent acquisition strategies of firms. *Management Science*, 50(6), 804-820.

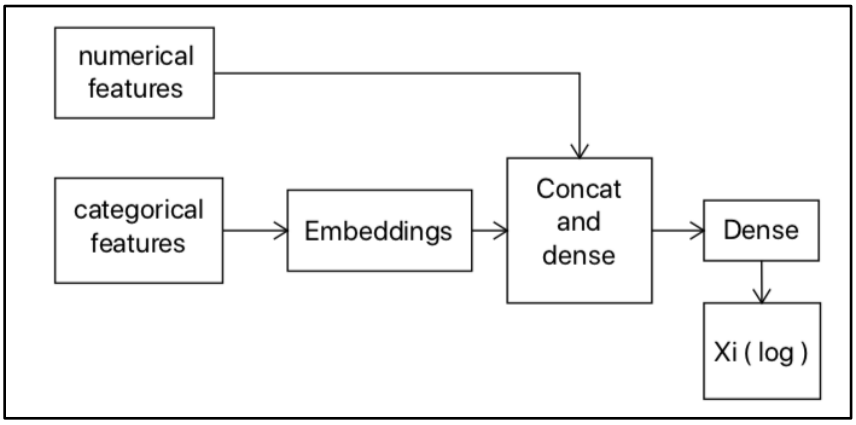

Figure 1: Structured model

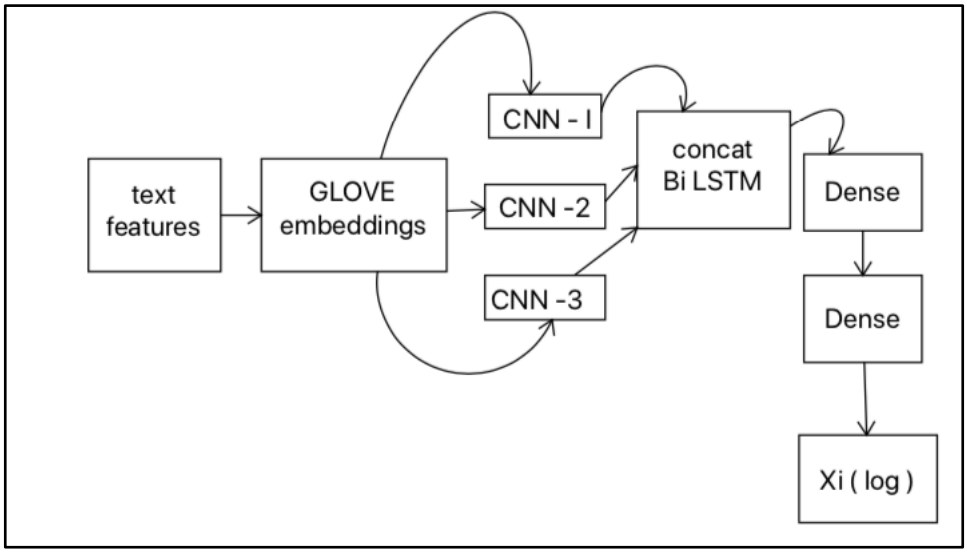

Figure 2: Text model

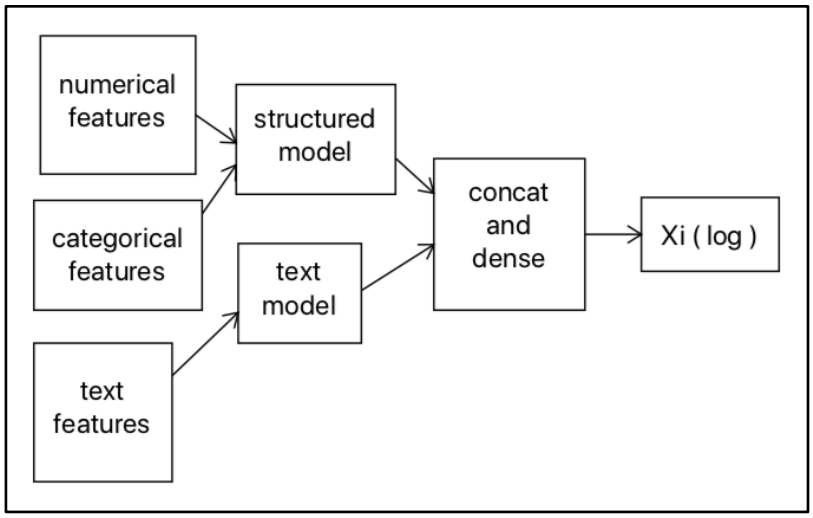

Figure 3: Structured + text model

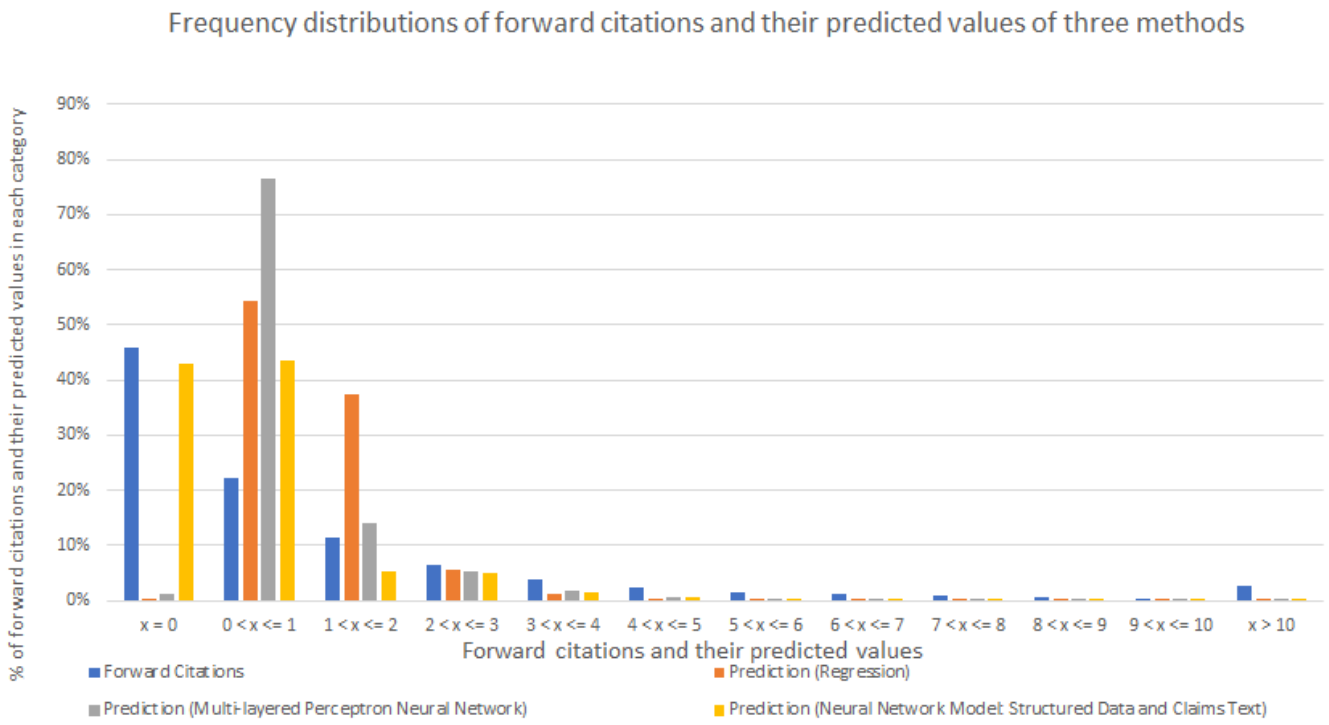

Figure 4: Frequency of the distributions of forward citations and their predicted values

This figure is the frequency distribution of the three-year forward citations and three predicted values of patents granted to public firms in the sample period 2003 to 2017. The first histogram labelled "Forward Citations" denotes the actual forward citations. The second histogram labelled "Prediction (Regression)" presents the predicted forward citations based on the linear model of Model (6) in Table 1. The third histogram labelled "Prediction (Multi-layered Perceptron Neural Network)" presents the predicted models based on neural networks with only structured data. The fourth histogram labelled "Prediction (Neural Network Model Structured Data and Claims Text)" presents the predicted models based on neural networks with structured data, claims, and brief summary.

Table 1: Summary of approaches and results

Method	Mean Absolute Error (MAE)
Linear regression model	
Model (1): Patent characteristics, and class and year FE	8.64
Model (2): Patent characteristics and all FE	7.58
Model (3): Patent characteristics, firm characteristics, and class and year FE	7.02
Model (4): Patent characteristics, firm characteristics, and all FE	5.95
ML models using structured features	
Ridge regression	4.40
Random forest	4.02
XGBoost	5.66
FF Neural Net	4.65
Best text only model	8.89
Best all features + text model	3.26

Table notes: This table summarizes key models and results using data features derived from patent filings to predict patent values as measured in Kogan et al (2017). Patents covered in our sample include those granted to U.S. public firms in 2003 to 2017. Detailed descriptions of these models and approaches are discussed in the subsequent sections.

Table 2: Linear regression results with structured features in the training sample

Model	(1)	(2)	(3)	(4)
Observations	930,618	911,838	828,058	827,668
R-squared	0.168	0.793	0.795	0.863
Patent characteristics	Yes	Yes	Yes	Yes
Firm characteristics			Yes	Yes
Grant year FE	Yes	Yes	Yes	Yes
Class (CPC) FE	Yes	Yes	Yes	Yes
Firm (PERMNO) FE		Yes		Yes
Industry FE		Yes		Yes

Table notes: This table reports linear regression results of patent features on patent value in the training sample. PERMNO indicates organization fixed-effects and they are included in columns 1, 2, and 3. Standard errors are shown in parentheses, *** p<0.01, ** p<0.05, * p<0.1

Table 3: Machine learning models using structured features

Models	Ncites	Num claims	Permno	App year	grant year	Class	**Test R^2**	**Test RMSE**	**Test MAE**
Ridge	x	x					0.00	16.66	8.91
Ridge	x	x		x	x		0.01	16.59	8.85
Ridge	x	x		x	x	x	0.07	16.11	8.35
Ridge	x	x	x	x	x		0.61	10.43	4.40
RForest	x	x					0.00	16.72	8.93
RForest	x	x		x	x		0.01	16.62	8.87
RForest	x	x		x	x	x	0.03	17.88	8.98
RForest	x	x	x	x	x		0.60	10.60	4.02
XGBoost	x	x					0.00	16.67	8.91
XGBoost	x	x		x	x		0.01	16.59	8.85
XGBoost	x	x		x	x	x	0.05	16.30	8.44
XGBoost	x	x	x	x	x		0.58	11.21	5.66
FF Neural Net	x	x					0.00	16.66	8.85
FF Neural Net	x	x		x	x		0.01	16.59	8.71
FF Neural Net	x	x		x	x	x	0.06	16.26	8.49
FF Neural Net	x	x	x	x	x		0.53	11.50	4.65

Table 4: Deep learning models with and without NLP

1	2	3	4	5	6
No.	**Features**	**Variables and data processing method**	**Model description**	**Test Set MAE**	**Time Taken to Train (approx)**
1	structured	embed categorical features	1 dense layer, 1 output single neuron	5.98	1 hour
2	structured	embed categorical features; top 1% value removed (outliers)	1 dense layer, 1 output single neuron	4.08	1 hour
3	brief summary	embedding the first 3000 tokens of brief summary using GLoVe Embedding after removing stop words	3 CNN layers, 1 bidirectional LSTM layer, 1 dense layer, 1 output single neuron	8.89	50 hours
4	claims	embedding the first 3000 tokens of claims using GLoVe Embedding after removing stop words	Same as above	9.64	50 hours
5	brief summary	embedding the first 3000 tokens of brief summary using GLoVe Embedding after removing stop words; top 1% value removed (outliers)	Same as above	6.86	50 hours
6	claims	embedding the first 3000 tokens of claims using GLoVe Embedding after removing stop words; top 1% value removed (outliers)	Same as above	7.11	50 hours
7	brief summary + structured	embed categorical features; embedding the first 3000 tokens of brief summary using GLoVe Embedding after removing stop words.	text input only: 3 CNN layers, 1 bidirectional LSTM layer, 1 dense layer; Structured input only: 1 dense layer. Concatenate the two output to feed 1 output single neuron	4.7	70 hours
8	brief summary + structured	embed categorical features; embedding the first 300 tokens of brief summary using GLoVe Embedding after removing stop words; top 1% value removed (outliers)	Same as above	3.37	70 hours
9	brief summary + structured	embed categorical features; embedding the first 3000 tokens of brief summary using fastText Embedding after removing stop words; top 1% value removed (outliers)	Same as above	**3.26**	70 hours
10	claims + structured	embed categorical features; embedding the first 3000 tokens of claims using fastText Embedding after removing stop words; top 1% value removed (outliers)	Same as above	3.62	70 hours
11	brief summary + claims + structured	embed categorical features; embedding the first 3000 tokens of claims and brief summary using fastText Embedding after removing stop words; top 1% value removed (outliers)	Same as above	3.49	90 hours
12	claims + structured	remove stopwords and special characters, weigh claims unigrams by TF-IDF and take the top 500 along with top 500 bigrams.	XGBoost	5.25	2 hours
13	brief summary + structured	remove stopwords and special characters, weigh brief summary unigrams by TF-IDF and take the top 500 along with top 500 bigrams.	XGBoost	5.32	2 hours
Table notes: For all rows, training sample size is 1 million and test sample size is 100,000.					

Table 5: Summary of results for forward citations

Method	Mean Absolute Error (MAE)
Linear Regression Model: Patent characteristics, firm characteristics, and all FE	1.736
Multi-layered Perceptron Neural Net: Using structured data only	1.604
Neural Network Model: Structured Data and Claims Text	1.613
Table notes: This table summarizes key models and results using data features derived from patent filings to predict the number for forward citations. Patents covered in our sample include those granted to U.S. public firms in 2003 to 2017.	

Table 6: Summary of results for patents from firms that have recently IPOed

Panel A: Patents granted within 2 years of the IPO date

Method	Mean Absolute Error (MAE)
Linear Regression Model: Patent characteristics, and class and year FE	20.10
Multi-layered Perceptron Neural Net: Using structured data only	14.62
Neural Network Model: Structured Data and Claims Text	10.58
Table notes: This table summarizes key models and results using data features derived from patent filings to predict the value of newly IPO firms' patents. Patents covered in our sample include those granted to U.S. public firms in 2003 to 2017.	

Panel B: Patents granted within 3 years of the IPO date

Method	Mean Absolute Error (MAE)
Linear Regression Model: Patent characteristics, and class and year FE	19.61
Multi-layered Perceptron Neural Net: Using structured data only	16.13
Neural Network Model: Structured Data and Claims Text	12.89
Table notes: This table summarizes key models and results using data features derived from patent filings to predict the value of newly IPO firms' patents. Patents covered in our sample include those granted to U.S. public firms in 2003 to 2017.	

Appendix A: Variable Definitions

Patent-level variables	**Definition**
Number of claims	Number of claims listed in a patent document
Length of the first claim	Number of words contained in the first claim
Length of all claims	Number of words contained in all claims in a patent document
Number of backward citations	Number of patents listed in the reference of a patent document
Length of the brief	Number of words in the brief summary text of a patent document
Number of new backward citations	The number of backward citations that are cited by a patent granted to a firm and that have never been cited in prior patents granted to the firm
Ratio of new backward citations	Number of new backward citations scaled by the number of all backward citations made by all patents granted to the firm
Originality index	The Herfindahl-Hirschman Index (HHI) of the distribution of the subsections of prior patents in the reference list (backward citations) of a patent document
Number of backward citations to basic science	Number of backward citations to non-patent documents made by a patent
Ratio of backward citations to basic science	Number of backward citations to basic science scaled by the number of all backward citations made by a patent
Duration of backward citations	The average duration of all backward citations made by one patent. The duration of a backward citation is defined as the grant year of the patent minus the grant year of the backward citation/referenced patent
Half-life of backward citations	The median duration of all backward citations made by one patent. The duration of a backward citation is defined as the grant year of the patent minus the grant year of the backward citation/referenced patent
Ratio of backward citations within 5 years	Number of backward citations with duration shorter than or equal to five years scaled by the number of all backward citations made by a patent
Breadth	Number of unique subsections in a patent document
Exploration	An indicator variable that equals one if the ratio of new knowledge sources to all backward citations is larger than or equal to 80% and zero otherwise. A new knowledge source denotes a backward citation that is neither a prior patent owned by the same patent assignee nor a backward citation made by prior patents owned by the same patent assignee.
Exploitation	An indicator variable that equals one if the ratio of old knowledge sources to all backward citations is larger than or equal to 80% and zero otherwise. An old knowledge source denotes a backward citation that is not a new knowledge source
Scope	The ratio of new knowledge sources to all backward citations. A new knowledge source denotes a backward citation that is neither a prior patent owned by the

	same patent assignee nor a backward citation made by prior patents owned by the same patent assignee.
Depth	The ratio of repeated knowledge sources to all backward citations. A repeated knowledge source denotes a backward citation that is either a prior patent owned by the same patent assignee or a backward citation made by prior patents owned by the same patent assignee.
Breadth of backward citations	Number of unique three-digit subsections (including the primary one and all secondary ones) of all referred patents (backward citations) of a patent
Number of self-citations	Number of backward citations that are prior patents owned by the same patent assignee
Indicator for no-self-citation	An indicator variable that equals one if none of a patent's backward citations is a self-citation, and zero otherwise
Patent thicket	The Herfindahl-Hirschman Index (HHI) of the distribution of the ownership of prior patents in the reference list (backward citations) of a patent document
Affiliation of an international patent family	An indicator variable that equals one if the patent belongs to an international patent family (i.e., having an international family ID in the PATSTAT database), and zero otherwise
Number of patents in the same family	Number of all patents covered in an international patent family
Number of U.S. patents in the same family	Number of U.S. patents covered in an international patent family
Firm-level variables	**Definition**
R&D expenditures in log	The log value of one plus a firm's annual R&D expenditures in a year
Ratio of R&D to total assets	The ratio of a firm's annual R&D expenditures in a year to its total assets in the same year
Advertising expenditures in log	The log value of one plus a firm's annual advertising expenditures in a year
Ratio of advertising to total assets	The ratio of a firm's annual advertising expenditures in a year to its total assets in the same year
The ratio of capital expenditure to total assets	The ratio of a firm's annual capital expenditures in a year to its total assets in the same year
Market capitalization in log	The log value of a firm's market capitalization (i.e., stock prices times all shares outstanding)
Market-to-book ratio	The ratio of market capitalization to total book equity of shareholders
Asset tangibility	The ratio of property, plant, and equipment to total assets
Firm age	The age of a firm starting from its first time appearance in the Compustat
ROA	A firm's net income scaled by lagged total assets (in the last year)

ROE	A firm's net income scaled by lagged total book equity of shareholders (in the last year)
Financial leverage	A firm's total debt (short-term debt and long-term debt) scaled by its total assets
Ratio of cash flows to total debts	A firm's cash flows (income before depreciation minus total depreciation) scaled by total debt
Ratio of cash holdings to total assets	A firm's cash and equivalents scaled by its total assets
Industry concentration in sales	The Herfindahl-Hirschman Index (HHI) of the sales of all firms in the same Fama-French industry (defined as 48 industries)
Number of employees	The number of employees (missing value is set as zero)
Indicator for missing employees	An indicator variable that equals one if a firm does not report the number of employees in a year, and zero otherwise
Number of inventors	Number of unique patent inventors who appear in all patents granted to a firm in the most recent five years
Number of inventors per patent	Number of co-inventors per patent of a firm in the most recent five years
Originality index (firm)	Herfindahl-Hirschman Index (HHI) of the distribution of the subsections of prior patents cited by all patents granted to a firm in a year
Average originality score	The average of originality scores of all patents granted to a firm in a year
Sum of originality scores	The sum of originality scores of all patents granted to a firm in a year
Duration of backward citations	The average duration of backward citations of all patents granted to a firm in the most recent five years
Half-life of backward citations	The average half-life of backward citations of all patents granted to a firm in the most recent five years
Maximum patent-level breadth	The maximum of patent-level subsections coverage across all patents granted to a firm in the most recent five years. A patent's subsection coverage is the number of unique subsections to which it is assigned
Average number of backward citations to basic science	The average of the numbers of backward citations to non-patent documents of all patents granted to a firm in the most recent five years
Average ratio of backward citations to basic science	The average ratio of backward citations to basic science scaled by the number of all backward citations of all patents granted to a firm in the most recent five years
Number of patents	Number of patents granted to a firm in the most recent five years
Number of patents scaled by total assets	Number of patents scaled by the firm's total assets
Number of citations received	Number of citations made by patents that are granted in the most recent five years and refer to any patent granted to a firm

Number of citations received scaled by total assets	Number of citations received scaled by the firm's total assets
Patent thicket (firm)	The Herfindahl-Hirschman Index (HHI) of the distribution of the ownership of prior patents in the reference list (backward citations) of all patents granted to a firm in the most recent five years
Breadth (firm)	Number of unique subsections of all patents granted to a firm in the most recent five years
Breadth of backward citations (firm)	Number of unique three-digit subsections (including the primary one and all secondary ones) of all backward citations made by all patents granted to a firm in the most recent five years
Exploration (firm)	Average of patent-level exploration of all patents granted to a firm in the most recent five years
Exploitation (firm)	Average of patent-level exploitation of all patents granted to a firm in the most recent five years
Scope (firm)	The average patent-level scope of all patents granted to a firm in the most recent five years
Depth (firm)	The average patent-level depth of all patents granted to a firm in the most recent five years
Average breadth of backward citations	The average of patent-level breadth of backward citations of all patents granted to a firm in the most recent five years

Appendix B: Summary Statistics

Panel A: Patent-level variables	Training Sample N = 1,110,333 patents		Test Sample N=90,000 patents	
	Mean	**St.Dev**	**Mean**	**St.Dev**
Patent value	9.46	17.05	9.53	17.09
Breadth of backward citations	2.46	2.00	2.47	2.02
Number of patents in the same family	4.39	7.56	4.40	8.04
Number of backward citations	11.98	23.46	12.10	24.08
Number of new backward citations	5.87	9.60	5.87	9.52
Scope	0.63	0.36	0.63	0.36
Depth	1.84	11.44	1.84	10.88
Number of claims	16.41	13.07	16.41	13.08
Number of backward citations to basic science	3.21	13.91	3.23	13.69
Breadth	1.58	0.85	1.58	0.84
Patent thicket	0.77	0.34	0.77	0.34
Duration of backward citations	7.05	3.77	7.06	3.78
Exploitation	0.19	0.39	0.19	0.39
Length of the brief	8521.41	10935.05	8501.21	10105.70
Length of all claims	6475.91	4572.49	6488.97	4581.13
Length of the first claim	149.49	88.47	149.97	92.33
Originality index	0.30	0.27	0.30	0.27
Number of U.S. patents in the same family	1.48	2.15	1.48	2.30
Affiliation of an international patent family	0.59	0.49	0.59	0.49
Ratio of new backward citations	0.64	0.36	0.64	0.36
Indicator for non-self-citation	0.63	0.48	0.63	0.48
Exploration	0.46	0.50	0.46	0.50
Number of self-citation	0.90	2.55	0.90	2.52
Ratio of backward citations within 5 years	0.60	0.33	0.60	0.33
Half-life of backward citations	6.55	3.90	6.56	3.91
Ratio of backward citations to basic science	0.11	0.19	0.11	0.19

Panel B: Firm-level variables	Training Sample N = 1,110,333 patents		Test Sample N=90,000 patents	
	Mean	**St.Dev**	**Mean**	**St.Dev**
Industry concentration in sales	0.11	0.09	0.11	0.09
Number of patents scaled by total assets	0.17	0.36	0.17	0.37
R&D expenditures in log	6.14	2.14	6.15	2.14
Advertising expenditures in log	3.10	3.12	3.09	3.12
Market capitalization in log	15.26	2.39	15.27	2.40
Market-to-book ratio	2.90	11.53	2.91	13.47
Financial leverage	0.21	0.15	0.21	0.15
The ratio of capital expenditure to total assets	0.07	0.05	0.07	0.05
ROA	0.05	0.14	0.05	0.13
Firm age	25.28	12.99	25.29	13.04
Ratio of cash flows to total debts	39.11	833.71	38.57	831.07
Number of employees	104.20	122.80	104.63	122.79
Originality index (firm)	0.30	0.27	0.30	0.27
Average ratio of backward citations to basic science	0.11	0.10	0.11	0.10

Average breadth of backward citations	2.30	0.92	2.30	0.94
Exploration (firm)	0.49	0.21	0.49	0.21
Exploitation (firm)	0.18	0.13	0.18	0.13
Maximum patent-level breadth	5.81	1.75	5.82	1.74
Number of inventors	2878.20	3987.12	2891.34	3994.34
Duration of backward citations	6.66	2.56	6.65	2.56
Patent thicket (firm)	0.94	0.10	0.94	0.10
Scope (firm)	0.65	0.16	0.65	0.16
Depth (firm)	1.60	3.36	1.61	3.41
Indicator for missing employees	0.04	0.19	0.04	0.19
Number of citations received scaled by total assets	1.16	2.76	1.18	2.99
Number of patents	3219.87	4317.10	3232.15	4330.16
Number of citations received	31370.12	56573.81	31632.03	57250.68
The ratio of R&D to total assets	0.07	0.07	0.07	0.07
Ratio of advertising to total assets	0.01	0.03	0.01	0.03
Asset tangibility	0.27	0.16	0.27	0.16
ROE	0.11	1.50	0.12	0.62
Ratio of cash holdings to total assets	0.16	0.15	0.16	0.15
Average originality score	0.29	0.11	0.29	0.11
Sum of originality scores	245.85	350.71	247.01	351.53
Average number of backward citations to basic science	5.20	7.11	5.23	7.43
Breadth (firm)	55.87	30.68	56.07	30.65
Breadth of backward citations (firm)	86.72	37.34	86.92	37.30
Number of inventors per patent	0.99	0.37	0.99	0.37
Half-life of backward citations	5.69	2.23	5.68	2.22

Appendix C: Technical appendix on the application of deep learning to patent text

This section provides technical details on how deep learning models were applied to the patent text to predict their economic and technological impact. This section is intended for:

i) readers who want details on the methods used in the paper and
ii) researchers who may be interested in implementing similar patent text learning architectures for their own research purposes.

Overview of workflow. The patent data were obtained from archival sources and then cleaned and prepared for analysis. The unstructured text data are converted to numerical representations through the application of "word embeddings" and then appended to the vector of structured features to form a single numerical representation for each patent. These patent vectors, which represent the different patent documents, are then passed to deep-learning models that were implemented by calling APIs from two popular open source frameworks, 'tensorflow' and 'pytorch'. The hyperparameters specifying the neural network architecture are then tuned to minimize prediction error for the training data set with validation set. Finally, the trained deep learning model is applied to the out-of-sample data to obtain predicted values of interest. The next sections provide details for each of these steps.

1. **Data sources**

Predicting patent value (measured by market value and forward citations) requires three sources of data.

i) Patent data

We use patent data from 1976-2010. The patent data were obtained from the USPTO Patents View database (https://www.patentsview.org/download/).

ii) Market value of patents

For each of the patents in the data set, we obtained matching estimates of economic value (Kogan et al 2017) from the companion website to the paper. Kogan and co-authors compute economic values for patents by studying market reactions to patent announcements made by public firms. Although more recent data have been made available by the authors using the same methods (extended through 2019), we use the data set that extends to patents filed through 2010 in order to match the data that we use to that used in the published paper, as that sample has been most extensively analyzed by subsequent work.

iii) Forward citations

Forward citations are the number of citations from later patents that cite the focal patent. They are often viewed as an indicator of the impact of a patent on the direction of future technological innovation. Forward citations are available for download through the NBER Patents database.

2. Preparation of the data

Feature generation is described below. The dependent variables (market value, forward citations) are winsorized at 1% before being used for model training.

3. Deep learning frameworks

Several open source platforms are widely used in deep learning applications. Some of the most popular of these include *pyTorch*, *TensorFlow + Keras*.[19] These platforms are accessible through the publication of API's (application programming interfaces) that can be invoked through Python and other languages. These frameworks provide abstractions that reduce the amount of code required to implement deep learning architectures.

Our analysis uses (through Python) the *tensorflow* (developed by Google) and *pytorch* (developed by Facebook) machine learning frameworks. Both platforms have grown rapidly in importance and usage in the last several years. Although the capabilities of these frameworks have arguably converged in recent years, the key differences are in how computational and architectural processing works (static vs dynamic computational graphs) and in the available support (codes and packages) for particular algorithms. We would expect the output to be similar when using either of these models.

4. Generating the input data for deep learning

Deep learning architectures require data to be passed to neural networks in numerical form. In other words, all of the available patent input features, both the structured data and the unstructured text data, must first be converted into sequences of numbers that represent each of the patent documents.

Text data

Converting the unstructured textual data for use in deep learning applications requires application of a *word embedding* that converts language input into a numerical representation of the text. Word embeddings convert a text corpus into numbers that can serve as inputs to nodes in a neural network.

There are a variety of word embeddings frequently used to develop language models. In this paper, we use two popular models: i) Global Vectors for Word Representation (GloVe, from Stanford) (Pennington et al 2014) and ii) fasttext (from Facebook) (Joulin et al 2016). Future research might investigate incorporating contextualized word embeddings such as the BERT and GTP families.

Before applying these word embeddings, we first apply the spaCy package for tokenization (i.e. converting text documents into words and phrases) and cleaning the text.

Structured patent features

The structured features used in the prediction task (grant date, backward citations, etc.) are already in numerical form. For each patent, this set of features forms a vector that was appended to the vector derived from the text data for each patent.

5. Tuning of key hyperparameters

After the data are made suitable for input to the deep learning engine, there are a number of "hyperparameters" that require tuning in order to optimize application performance. These hyperparameters – such as the learning rate, number of epochs, number of hidden layers, hidden units, and choice of activation function specify the structure and learning characteristics of the network. Different datasets require different hyperparameters that are found through application to validation set. For our data and application, our choice of hyperparameters is shown below.

[19] There are numerous articles tracing the evolution of deep learning frameworks. For example, see https://medium.com/@ODSC/5-deep-learning-frameworks-to-consider-for-2020-183e6c12cba9.

Hyperparameter	Value
Learning Rate	0.001-0.01
Number of Epochs	20-50
Hidden Layers	2-4
Hidden Units	10-50
Activation Functions	ReLU

6. Computational architecture

The storage and computational demands of the deep learning analysis are substantial. The deep learning models analyzed in this paper were executed on our own GPU workstations and servers consisting of 4 GTX 1080 TI or Titan xp and 128 GB of ram. Parallelization was carried out by running different parameter estimation on each of the GPUs to accelerate grid search.

For analyses with these types of computational demands, another common approach is to use cloud computing solutions, such as those offered by Google Cloud or Amazon Web Services (AWS). A free but limited GPU is accessible through Google Colab. The scalability of these solutions allows trading off cost with computational processing time.

7. Caveats on fitting the model

Unlike shallow models, deep learning architectures can be difficult to train. Even simple deep learning models may cease to learn or fail if model parameters are not well tuned (e.g. see Bengio 2012).

8. Predicting out-of-sample values

The final step in this process is to apply the trained model to the out-of-sample observations to obtain predictions. This step is relatively straightforward. A source of occasional errors is that it is important to ensure that the input data vectors for the out-of-sample observations match those used to train the model.