From Text to Verdict: Predicting IP Litigation Outcomes with Machine Learning

Haolin Li, Anthony Bellotti, Xiuping Hua, Wei Huang, Haisheng Yang

This draft: December 21, 2025

Abstract

This study evaluates the effectiveness of various machine learning models in predicting the outcomes of intellectual property (IP) litigation, leveraging data from China Judgement Online, which contains over 140 million lawsuits spanning 2010 to 2021. We hypothesize that the analysis of textual content should be contextualized with the company characteristics. Our findings indicate that the Structural Topic Model (STM), which integrates both financial and textual data, significantly enhances the accuracy of predictive models compared to those relying on a single input type. The STM exhibits strong predictive performance across a broad range of cases involving both plaintiffs and defendants, and across different IP types – including copyrights, trademarks, and patents. It is also more effective in predicting IP litigation outcomes of cross-border U.S. or European companies. Using the STM framework, we are able to identify five leading thematic topics associated the win/loss outcome of IP litigation in both full sample and the cross-border subsample, illustrating the STM's practical feasibility for predicting the IP litigation outcome. Winning IP litigation is positively associated with a plaintiff's return on assets (ROA) and higher levels of innovation. In asset pricing tests, an Instrumented Principal Component Analysis (IPCA) that incorporates IP litigation text features outperforms three-factor model of Liu, Stambaugh, and Yuan (2019). These findings underscore the importance of robust legal protections for IP in fostering innovation and bolstering corporate financial performance.

JEL Classification: O30; O32; O34; G17

Key Words: Machine learning; intellectual property litigation; financial impact of IP litigation; textual analysis; structural topic model

---

Li (HAOLIN.LI@nottingham.edu.cn) is from School of Computer Science, The University of Nottingham Ningbo China, Bellotti (Anthony-Graham.Bellotti@nottingham.edu.cn) is School of Computer Science, The University of Nottingham Ningbo China, Hua (Xiuping.HUA@nottingham.edu.cn) is from Nottingham University Business School, The University of Nottingham Ningbo China, Huang (weih@hawaii.edu) is from Shidler College of Business, University of Hawaii at Manoa, USA, Yang (yhaish@mail.sysu.edu.cn) is from Lingnan (University) College, Sun Yat-sen University, Guangzhou, China. We thank Kevin Tseng (discussant), Lin William Cong (session chair), and the conference participants at 2025 China International Conference in Finance (CICF) for helpful comments and suggestions. Any errors are our own.

From Text to Verdict: Predicting IP Litigation Outcomes with Machine Learning

This draft: December 21, 2025

Abstract

This study evaluates the effectiveness of various machine learning models in predicting the outcomes of intellectual property (IP) litigation, leveraging data from China Judgement Online, which contains over 140 million lawsuits spanning 2010 to 2021. We hypothesize that the analysis of textual content should be contextualized with the company characteristics. Our findings indicate that the Structural Topic Model (STM), which integrates both financial and textual data, significantly enhances the accuracy of predictive models compared to those relying on a single input type. The STM exhibits strong predictive performance across a broad range of cases involving both plaintiffs and defendants, and across different IP types – including copyrights, trademarks, and patents. It is also more effective in predicting IP litigation outcomes of cross-border U.S. or European companies. Using the STM framework, we are able to identify five leading thematic topics associated the win/loss outcome of IP litigation in both full sample and the cross-border subsample, illustrating the STM's practical feasibility for predicting the IP litigation outcome. Winning IP litigation is positively associated with a plaintiff's return on assets (ROA) and higher levels of innovation. In asset pricing tests, an Instrumented Principal Component Analysis (IPCA) that incorporates IP litigation text features outperforms three-factor model of Liu, Stambaugh, and Yuan (2019). These findings underscore the importance of robust legal protections for IP in fostering innovation and bolstering corporate financial performance.

JEL Classification: O30; O32; O34; G17

Key Words: Machine learning; intellectual property litigation; financial impact of IP litigation; textual analysis; structural topic model

1. Introduction

Intellectual property (IP) litigation has become increasingly important for publicly listed companies, as IP litigation affects a company's valuation and innovation. For example, the breadth of intellectual protection such as patent scope significantly affects valuations (Lerner, 1994). Patent litigation negatively affects investment and lower costs of patent litigation for defendants increase innovation. (Mezzanotti, 2021). Firms sued for patent infringement exhibit significantly higher expected stock returns in the following year since investors could over-discount alleged infringers' stock prices (Bereskin et al. 2023). Therefore, accurately predicting the outcomes of IP litigation is of great importance for risk management, financial planning, and long-term strategic decision-making (Branting et al., 2021).

In this study, we aim to enhance the accuracy of predicting the outcomes of IP litigation by leveraging text-based machine learning techniques that are better aligned with the complexities of such legal proceedings. Using a more accurate machine learning approach, we then investigate the impact of IP litigation outcome on firms' financial and operational performance. Following the narrative asset pricing theory, we further apply Instrumented Principal Component Analysis (IPCA) to test whether the explanatory power of three-factor model of Liu, Stambaugh, and Yuan (2019) can be improved by additionally incorporating IPCA based on IP litigation texts as prior studies found IP protection affects firm value.

Recent research found that unstructured data, such as textual data often contains actionable information that can support decision-making (Phan et al.,2023). Thus, we

predict the IP litigation outcomes with both text information and financial characteristics, which bypasses the limitations of using either text or firm characteristics for prediction. Previous studies often utilizes numerical data to model and anticipate litigation results. Despite the relatively high efficacy of numerical data in performance, it is susceptible to inaccuracies because numerical data can be easily manipulated by a sophisticated deceiver (Wang and Xu, 2018). Furthermore, the intricacies of legal standards and the divergence in judicial interpretations across different jurisdictions present substantial obstacles to the development of universally applicable predictive models. These models may also fall short in capturing the unique narratives and specific details that characterize each individual litigation case, thereby potentially undermining the reliability of their forecasts.

There is a growing literature on how textual information can help to build a better prediction model in firm's performance and legal domain (Tetlock, 2007; Loughran and McDonald, 2011; Medvedeva et al., 2020; Garcia et al., 2023;). The literature on textual analysis in finance focused on news media as a starting point. Studies then use bag of words approach, which is similar to word frequency method developed by Loughran and McDonald (2011), to analyze text context. However, this method has drawn criticism for lacking effectiveness when compared to machine learning approaches (Gentzkow et al., 2019).

Textual topic models can uncover clustering patterns of various terms, allowing for the derivation of differences among sets of words. This type of information at the thematic level is not readily obtainable from mere word frequencies or financial indicators alone. Wang and Xu (2018) employed the Latent Dirichlet Allocation (LDA) model to predict automobile insurance fraud. They found that textual features within the accident descriptions in claims were indicative of potential fraud. However, LDA,

as an unsupervised model, can sometimes produce topics that are difficult to accurately interpret due to poor classification performance. This can result in imprecise topic factors, diminishing the model's predictive accuracy. Moreover, LDA-based models are often considered “black boxes,” making it challenging for users to derive an intuitive understanding or insight into the relationship between the explanatory variables (topic factors) and the dependent variables, which is 1 for winning case and 0 for losing case in IP litigation. In response to these limitations, we propose an IP litigation prediction model based on the Structural Topic Model (STM).

STM is a semi-supervised machine learning approach to topic modeling which integrates unsupervised topic modeling with supervised information through the inclusion of covariates and also the outcome variables. These covariates are observable metadata such as authorship, publication date, or news source. They influence the prevalence of topics and the specific words associated with each topic, allowing STM to leverage external information and provide a more nuanced understanding of document content compared to traditional unsupervised methods. STM is considered superior to the LDA model because it offers greater flexibility by incorporating a wide range of document-level covariates and metadata, which enables a more detailed analysis of how various factors influence document content, thus enhancing both predictive power and interpretability of the model. Consider two commonly used English phrases: “No one can touch the speed of the cheetah” versus “No one can touch the speed of a snail.” The LDA would treat the two phrases as the same meaning due to its inability to capture the thematic content. On the other hand, STM can distinguish the

difference by incorporating the thematic constraint related to the speed. It is unclear, however, whether STM approach provides a superior predictive power in IP litigation outcome relative to the LDA model. We fill the gap in this paper.

In this study, we collect lawsuit data from Chinese Judgement Online, covering the period from 2010-2021. To focus on IP related cases, we applied a keyword selection algorithm that utilized a set of seed words related to IP protection. Using the Word2vec algorithm, we ascertain the cosine similarity among words within a pre-existing corpus. This analytical process enabled us to identify and prioritize the top 50 keywords to filter and select relevant cases from the Chinese Judgement Online dataset. Subsequently, we matched the names of firms involved in the litigation with the corresponding cases and distinguish between plaintiffs and defendants. Concurrently, we integrated financial data from the China Stock Market & Accounting Research Database (CSMAR) for each firm on an annual basis, augmenting our dataset with crucial economic indicators. We employed STM to infer the thematic distribution within each case. This thematic information, coupled with the financial variables, served as the input for our machine learning models, which included logistic regression, random forest, support vector machine, XGBoost, and deep neural networks.

We initiated our analysis by executing our models with financial variables exclusively to establish our baseline for comparison. We then ran the models using solely textual features. Subsequently, we enhanced our baseline models by incorporating both financial and textual data into a unified analysis. Our findings show that the inclusion of textual features significantly enhances the predictive power of our

models. Specifically, models enriched with both financial variables and textual data outperformed the baseline models with financial variables only. These improvements stem from the fact that details in litigation descriptions often contain key information for predicting outcomes. Moreover, when comparing our approach to LDA and other methodologies that also incorporate textual data, our chosen semi-supervised STM algorithm offered a more nuanced extraction of information from textual sources. In particular, STM can incorporate factors that affect text writing to impose supervised constraints on the solution of text themes, making it more reasonable and enabling more accurate extraction of text themes (Robert et al., 2016).

Our study yields three key findings. First, incorporating textual data substantially improves the accuracy of IP litigation outcome predictions. The most effective models were those that combined financial and textual data, highlighting the value of integrating quantitative and qualitative information. Second, STM emerged as the best-performing model among those that included textual data. This indicates that when factors influencing the drafting of legal text are incorporated as a form of supervised constraint in the model, it leads to a more precise extraction of textual themes. The implication is that STM's semi-supervised learning approach is adept at discerning the subtleties within legal narratives, which are crucial for understanding the complexities of IP disputes. The significance of these findings lies in their potential to transform the way legal analytics are conducted. Third, we show that the STM model provides robust prediction across cases involving plaintiffs, defendants, and various types of IP, such as copyrights, trademarks, and patents. Furthermore, our results suggest that winning the

IP litigation is significantly associated with an increase in the plaintiff's return on assets (ROA) and higher level of innovation.

We contribute to the literature in four main aspects. First, this study makes a significant contribution to the field by presenting an effective explainable model for predicting the IP litigation outcome based on textual data. Literature often uses linear regressions to study a certain factor that has influence on litigation outcome (Aharony et al., 2015; Donelson and Hopkins, 2016). Donelson and Hopkins, (2016) studies the relationship between market-wide declines in stock prices and litigation incidence. Aharony et al. (2015) study the effect of CEO and director turnover on corporate lawsuits. However, such systems have very limited inherent explanatory capability (Branting et al., 2021). Our research develops an IP litigation outcome prediction model aimed at delivering explainable predictions, offering practical tools to enhance the efficiency of firm's decision-making processes without relying heavily on knowledge-engineering. We focus on approaches that do not require feature sets, instead utilizing textual descriptions of case facts as input.

Second, we utilize several machine learning models, which have several advantages over traditional econometric methodologies (Mai et al., 2019) for this problem domain. For example, machine learning models can capture complex patterns and nonlinear relationships in data that may be overlooked in traditional econometric models. Additionally, machine learning algorithms can effectively handle high-dimensional datasets and extract useful information. Our findings demonstrate that non-linear machine learning models such as random forest or the deep neural network yield

more accurate estimates compared to the traditional logistic regressions.

Third, we introduce an innovative application of STM in predicting IP litigation outcomes. By integrating STM into our framework, we have enhanced the model's ability to capture the thematic nuances within legal texts, which is a critical aspect of IP litigation where the phrasing and context of legal arguments can substantially influence case outcomes. To our knowledge, our study is the first application of STM to IP research in the world. In our data, we apply STM in Chinese context, which not only expands the applicability of STM across different linguistic and cultural contexts but also provides a novel analytical tool and approach for global IP research.

Fourth, we provide evidence that textual information can effectively complement traditional financial variables in our prediction task. Notably, we provide evidence that along with financial variables, textual information can significantly improve the prediction performance of IP litigation outcome. Our study contributes to the textual analysis literature by adding fresh insights into how textual features can significantly improve the predictive performance of classification models (Kriebel and Stitz, 2022; Stevenson et al., 2021).

Another significant contribution of our research is the implementation of STM (Structural Topic Model) to predict IP litigation outcome. Unlike existing STM studies, we robustly demonstrate the accuracy of the selected covariate x and the rationality of incorporating textual information through the accuracy of the prediction results. We chose to use STM combined with textual information because assessing a company's condition solely based on its litigation outcomes yields ambiguous information. This is

because litigation results are influenced by a multitude of complex factors, and relying solely on the outcomes does not provide a comprehensive understanding of key information such as the company's operational status, market competitiveness, and management capabilities. For instance, a company may lose a lawsuit, but this does not necessarily indicate poor overall market performance; it could be due to issues in specific business segments or particular legal environments. Conversely, a company may win a lawsuit but still conceal potential risks in other areas. Only by integrating textual information, company-level situations, and litigation results can we make more precise and comprehensive judgments. Textual information can provide detailed backgrounds of cases, points of contention, and legal bases, which helps to deeply understand the essence of the litigation. Company-level situations, such as financial status, market position, and industry reputation, can reflect the company's overall strength and stability from a broader perspective. By synthesizing these multidimensional pieces of information, we can more accurately assess the company's legal risks, market competitiveness, and future development potential, thereby providing more valuable references for corporate decision-making and investment analysis. Therefore, through reverse prediction, we have proven the accuracy of the selected covariates and the rationality of introducing textual information.

In sum, we incorporate firm-level control variables, such as leverage, Tobin's Q, ROA, firm size, firm age and board size as document-level metadata into the STM framework, dynamically constraining the topic generation process through a covariate-adjusted topical prevalence mechanism. This design fully leverages STM's structural modeling

capacity to endogenously link textual topic distributions with firm-level heterogeneity, thereby capturing systematic variations in latent topics across corporate characteristics within the texts in intellectual property litigation. Compared to traditional topic models like LDA, which rely on exogenous assumptions about text-metadata relationships, this approach, which incorporates firm-level control variables as document-level metadata into the STM framework, not only mitigates the risk of topic conflation caused by omitted variable bias but also provides interpretable dimensions for heterogeneity analysis.

The rest of the paper is as follows. Section 2 describes the sample selection and textual analysis methods. Section 3 discusses machine learning techniques and evaluation measures. Section 4 presents our results and findings. Finally, Section 5 concludes the paper.

2. Data and textual analysis.

2.1 Sample selection.

To compile our dataset, we began by sourcing litigation cases from the Chinese Judgement Online database[1] This platform, launched in July 2013 by the Supreme People's Court, aims to enhance judicial transparency and elevate the standard of rule of law in China. By June 2015, the platform achieved a significant milestone: all three tiers of courts (provincial, municipal, and local city levels) across 31 provinces

[1] https://wenshu.court.gov.cn/

(including autonomous regions and municipalities) as well as the Xinjiang Production and Construction Corps, had successfully implemented the online publication of effective judgment documents. This initiative resulted in a comprehensive and inclusive archive encompassing all categories of cases adjudicated by various courts. As of December 22, 2023, the China Judgment Document Network has amassed over 143 million documents, highlighting the platform's extensive reach and utility. Moreover, the network has garnered substantial attention, with approximately 10.8 billion visits recorded till December 22, 2023, underscoring its role as a pivotal resource for legal research and public access to judicial information. This rich dataset provides a robust foundation for our analysis, allowing for an in-depth examination of IP litigation within the Chinese judicial system.

Our study focuses specifically on IP litigation cases. Therefore, the first step in our data collection involved using keywords to identify cases related to IP litigation. We follow the method of Li et al. (2021) and began with 10 seed words and expanded to a total of 50 keyword dictionary using the Word2Vec technique provided by Tencent AI Lab[2]. This approach leverages a word vector model, a pivotal concept in natural language processing (NLP), which is trained using word embedding technology and a vast corpus of text data (Mikolov et al. 2013).

[2] Based on the China Judgement Online Documentation, we extract 10 seed words, which include Trade Secret Infringement, Patent Infringement, Patent Right Dispute, Patent Ownership, Ownership of Patent Application Right, Intellectual Property, Trademark Infringement, Patent Dispute, Copyright Dispute, Copyright Ownership. We generate 100 key words related to intellectual property litigation.

Utilizing this method, we identified and compiled a dataset of 816,624cases related to IP litigation. Out of this sample, there are 382,299 IP litigation cases with the causes of lawsuits. Considering the unique dynamics of litigation involving listed firms in China, we further narrowed our focus to cases that directly pertain to these entities by matching the names of plaintiffs and defendants against a database of listed firms in China. This process yielded a dataset of 14,284 cases related to listed companies. Furthermore, we limited our analysis to first-trial cases as they provide more direct information (Long and Wang, 2015) and excluded cases with empty descriptions. This refinement produced a dataset of 5,322 first-trial cases involving listed companies. Finally, we exclude cases with missing financial data, resulting in a final dataset of 4,516 cases.

Table 1 provides a summary statistic on the China Judgement Online. Panel A provides the total number of cases over the period of 2010 to 2023. It is based on entire sample of China Judgment online Document. Panel B provides the total number of IP litigation cases over the period of 2010 to 2023. The numbers are based on the 816,624 IP cases extracted from the China Judgement Online Document. Panel C reports the total number of IP litigation for categories such as copyright, trademarks, and patents. Appendix A provides some examples of cross-border IP litigation cases.

2.2 Financial variables

We selected a set of financial variables to serve as inputs in our predictive models, as well as topic prevalence variables in STM, based on two criteria: (1) they are likely

to measure a firm's performance and financial conditions that affect a firm's interest in in IP protection, and (2) these variables are limited in number to avoid overfitting of our models (Palepu, 1986). For each case, we match the financial variables available at the time of the verdict. Below is a brief description of the financial variables used in this study.

First, Mezzanotti and Simcoe (2023) found that firm size is an important determinant of IP protection. They argue that larger firms show significantly more interest than small firms in all forms of IP protection, and that firm size generally explains more variance in patenting activities than any other observable characteristic. Second, financial performance is another key aspect of firm's IP protection. Hence, we use return on total assets (ROA), Tobin's Q, and leverage to measure the financial performance of firms (Dechow et al., 2011). Third, the board of directors, responsible for decision-making in listed companies, can also affect a firm's attitude toward IP litigation. Therefore, we include the size of the Board as a financial variable.

Table 2 Panel A provides a detailed description of the variables and Panel B reports the summary statistics of the financial variables used in our model. In Panel B of Table 2, 'litigation outcome' is a binary variable, where a value of 1 indicates that the listed firm won the IP litigation lawsuit, and a value of 0 indicates a loss. The mean of the litigation outcome is 0.548, suggesting that there are 54.8% of cases categorized as winning cases. The values of the financial variables are all within a reasonable range, consistent with existing literature, and there are no outliers.

2.3 Data preprocessing and textual analysis

2.3.1 Differentiation of litigation results

The court's judgment on IP cases is often not simply a win or lose, but selectively supports some of the plaintiff's claims. To ascertain the outcome of each legal case in our dataset[3], we followed the following procedures. If a judgment includes both "compensation" and "dismissal," the decision is considered in favor of the plaintiff if the awarded compensation exceeds 10% of the claimed amount; otherwise, it is deemed a victory for the defendant (Long and Wang, 2015). When a judgement contains only "dismissal," the ruling is conclusively in the defendant's favor. Conversely, if only "compensation" is mentioned without any "dismissal," the plaintiff is considered to have won. Finally, if the case is marked as "withdrawal" or "voluntary dismissal" by the plaintiff, the outcome is automatically in the defendant's favor. As indicated in Panel B of Table 2, the mean of the litigation outcome is 0.548, suggesting that there are 54.8% of cases categorized as winning cases.

2.3.2 Pre-processing

To enhance the quality of our dataset and reduce noise, we take the following steps. First, we refined the dataset by excluding cases that lack complete factual content or failed to specify a definitive outcome[4]. Next, considering that the text description of the litigation facts appears in the form of long sentences in different structure, but the key

[3] Most cases have only one document. If the plaintiff and the defendant are both listed companies, we separate them into two cases and the outcome of the IP litigation is based on each of their perspectives.

[4] Although all cases are closed, some cases may not have definitive outcomes. For example, the outcomes might be both the plaintiff and defendant must pay the fine. We delete those cases without definitive outcome.

information of the fact description is distributed across several words in a sentence (Wang et al., 2018), we focus on analyzing a few keywords instead of the entire sentence. Chinese word segmentation refers to the segmentation of a sequence of Chinese characters into short strings of words[5] (Wu et al., 2016). So, we employed the "jieba" segmentation tool, a widely recognized method in Chinese language processing, to accurately separate and tokenize the textual content of the remaining cases. Third, we removed abbreviations, numbers and stop words (Gandhi et al., 2019)[6]. This filtering process is essential for mitigating data sparsity issues and reducing potential noise, thereby improving the dataset's integrity and reliability (Loughran and McDonald, 2014). As discussed, our final dataset contains 4,516 observations.

2.4 Three Models in Textual Analysis

2.4.1 Word Frequency Model

The word frequency (WF) model is a linguistic analysis technique that leverages the frequency of specific word types within a text to understand its content and context (Loughran and McDonald, 2011). As mentioned earlier, we employ the "Jieba" segmentation tool, which is a powerful Chinese word segmentation library. It not only divides the text into individual words but also assigns a part-of-speech (POS)[7] tag to

[5] For example, an original Chinese sentence: I love natural language processing, can be segmented as I/love/natural language/processing.

[6] Stop words in Chinese characters, such as "de", "le". are possessive particle that indicates possession, or expresses change of state, completion of action. These words do not have actual English meaning.

[7] For example, the segmentation tool assigns "run" as a verb and "apple" as a noun.

each word, categorizing them into types such as nouns, verbs, adjectives, and adverbs (El-Haj et al., 2019). This step allows us to tally the frequency of each word type within the legal text of a judgment. The rationale behind focusing on nouns, verbs, adjectives, and adverbs is that these parts of speech often carry the most semantic weight and are integral to the meaning of sentences. By calculating the proportion of each of these word types, we create a feature set that serves as the input for the word frequency model. This approach is justified as it captures the essence of the text in a structured manner, which can be algorithmically processed to identify patterns and trends.

The input of WF model, therefore, consists of the relative frequencies or proportions of these key word types, which are believed to be indicative of the text's subject matter and style. This method is not only efficient in handling large volumes of textual data but also lays a solid foundation for further analysis, such as sentiment analysis, topic modeling, or predictive modeling in the context of legal judgments. The word frequency model is simple, but often proves effective, particularly in legal text analysis since it can help in discerning the legal arguments, claims, and decisions embedded within the language used in judicial opinions.

Word frequency models mainly fail to grasp the deeper themes in documents. They just count how often words appear without looking at what the words mean. This makes it hard to spot different topics and understand how words relate to each other. In addition, these models cannot deal with the various ways words are used or the subtle changes in word usage under different conditions, limiting their analytical depth and utility.

2.4.2 Latent Dirichlet Allocation (LDA)

In the field of natural language processing (NLP), Latent Dirichlet Allocation (LDA) stands out as a seminal generative model designed to uncover thematic structures within vast textual repositories or corpora (Blei et al., 2003). LDA encapsulates each document as a vector of word frequencies, thereby converting qualitative text data into a quantitative format suitable for data mining algorithms (Yuan et al.,2016). This approach assumes that each document is a probabilistic distribution across various topics, and in symmetry, each topic is characterized by a distribution over an array of words. A key feature of LDA is the autonomy of topics from one another, attributable to the non-correlation among the elements of the random vector within the Dirichlet distribution (Blei and Lafferty, 2006).

LDA operates by employing a joint distribution to infer the conditional probabilities of latent (hidden) topics based on the observable (known) variables. In this context, the observable variables are the words present, while the latent variables represent the topics themselves. Within our methodology, the textual narratives extracted from the descriptions of litigation cases are treated as documents. LDA is then applied to these documents to distill topics that pertain to the behavior and patterns of IP litigation. These documents are amalgamations of topics related to IP, with each topic being a collection of words distributed according to their probabilities. Topics are typically manifested through a selection of words with elevated probability distributions within those topics. Upon completing the LDA process, each legal case in the dataset is assigned a probabilistic topic distribution, each document is represented as a weighted combination of latent topics. This enriches the dataset by adding a thematic layer that

abstra cts textual patterns into interpretable legal concepts. However, LDA's critical limitation lies in its assumption that topic prevalence, which is related to firm characteristics in our setting, is fixed across all documents, ignoring contextual factors (e.g corporate characteristics) that systematically influence how topics manifest in legal texts.

For instance, in patent litigation, legal documents from large enterprises may emphasize themes like "patent portfolio strategy" (reflecting their scaled technological positioning), while startups focus more on "trade secret protection" themes (embodying their defensive strategies for core assets). LDA would obscure these distinctions by blending them into a global topic mixture, resulting in the loss of firm-level heterogeneity. LDA's "one-size-fits-all" topical modeling risks conflating firm-specific legal strategy signals. By contrast, STM, as we discuss in the next section, achieves dual-dimensional deconstruction of legal text analysis and corporate entity characteristics through structured integration of firm-level data. This capability holds critical significance for predicting outcomes in intellectual property litigation, particularly in cases involving complex corporate strategies.

2.4.3 Structural Topic Model

The Structural Topic Model (STM) is a structured approach for topic modeling, enabling the topic information within documents to be expressed in a structured manner. STM incorporates a method for topic prevalence, enabling the model to effectively supervise and constrain the selection of topics. This allows STM to generate a more relevant set of topics and their corresponding keywords. Building upon the foundation

of the LDA model, STM integrates covariates (metadata) to explore the relationship between the distribution of topics and these covariates. The STM model encompasses both topic content and topic prevalence. Topic content refers to the lexical items associated with a topic, while topic prevalence denotes the association between a document and its underlying topics.

The STM, like other topic models, is a generative model for word counts. It defines a process for generating data for each document and then uses this data to estimate the most probable values for the model's parameters. The generative process starts at the top with distributions of document-topics and topic-words, which creates documents containing associated metadata (denoted by X_d in Eq (1), where d is the document index). The generative process for each document with vocabulary of size V for a STM model with K topics can be summarized as: firstly, Equation (1) draws the document-level attention to each topic (topic prevalence model) from a logistic-normal generalized linear model based on a vector of document covariates X_d, where θ_d is the document-level attention distribution vector for document d across topics. Σ is a $(K-1)\times(K-1)$ covariance matrix. Secondly, Equations (2) and (3) denote the core language model. Equation (2) describes word's topic assignment based on the document-specific distribution over topics, where $z_{d,n}$ is the topic assignment for the nth word in document d. Equation (3) conditions on the topic chosen, draw an observed word from that topic, where $w_{d,n}$ is the nth word in document d, $\beta_{d,k} = z_{d,n}$ is the word distribution for the topic assigned to $z_{d,n}$. Thirdly, Equation (4) is the topic content model. It represents the word distribution $\beta_{d,k}$ for topic k in document d,

which is influenced by the baseline word distribution m, the topic-specific deviation $\kappa_{k,v}^{(t)}$, the covariate group deviation $\kappa_{y_d,v}^{(c)}$, and the interaction between them, $\kappa_{y_d,k,v}^{(i)}$. The architecture of STM is illustrated in Figure 1, where grey circles represent input variables, and white circles denote a series of latent variables. X represents the financial variables which are leverage, Tobin's Q, ROA, firm size, firm age and board size. D represents the textual data, and y represents the litigation outcomes. Arrows pointing from one circle to another indicate that the probability distribution of the latter is contingent upon the former. The detailed steps of STM are provided in Appendix B.

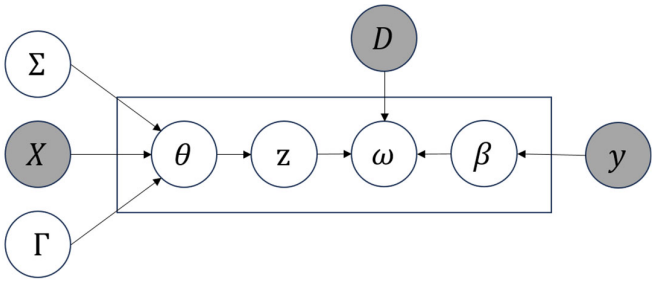

Figure 1. The technical diagram for the STM

$$\theta_d \sim LogisticNormal_{K-1}(\Gamma' X_d, \Sigma), for\ d = 1, \ldots, D \quad (1)$$

$$z_{d,n}|\theta_d \sim Multinomial_K(\theta_d), for\ n = 1, \ldots, N_d \quad (2)$$

$$\omega_{d,n}|z_{d,n}, \beta_{d,k} = z_{d,n} \sim Multinomial_V\left(\beta_{d,k} = z_{d,n}\right), for\ n = 1, \ldots, N_d \quad (3)$$

$$\beta_{y_d,k,v} = \frac{exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)}{\sum_v exp\left(m_v + \kappa_{k,v}^{(t)} + \kappa_{y_d,v}^{(c)} + \kappa_{y_d,k,v}^{(i)}\right)}, for\ v = 1, \ldots V\ and\ k = 1, \ldots, K \quad (4)$$

In summary, we hypothesize that STM is superior to LDA and word frequency models, because it offers greater flexibility by incorporating a wide range of document-level covariates and metadata, which enables a more detailed analysis of how various factors influence document content. This inclusion of covariates allows STM to

improve estimation accuracy by classifying similar firm characteristics across documents and leading to more stable and precise topic estimation. Furthermore, STM supports causal inference by examining the impact of specific covariates on topic prevalence and content, a capability that LDA, as a purely generative model without covariates, cannot provide. The ability to make such inferences is particularly valuable in social sciences and other fields where understanding the effects of specific variables on discourse is essential. Additionally, STM's interpretability is enhanced by demonstrating the relationship between covariates and topics, and it has been shown to have better predictive power, making it a more comprehensive tool for text analysis in the presence of metadata.

3. Methodology for predicting IP Litigation Outcome

In this section, we describe the three parts of our methodological approach. First, how we split our dataset into training set and out-of-sample test set. Second, we describe different machine learning approaches we employ in our prediction model. Third, we describe the measures we use to evaluate the performance of our models.

3.1 Splitting datasets

To address our research questions, we have employed the previously discussed methodology to categorize the legal cases into two distinct outcomes: wins and losses. Our process began with a random shuffling of our dataset to ensure an unbiased distribution, followed by the division of the data into separate training and testing sets. Following Doumpos et al. (2017) and Geng et al. (2015), we allocate 80% of the dataset,

selected randomly regardless of the year, to the training set, and the remaining 20% to the testing set. The efficacy of a classification model is contingent upon its predictive accuracy, which is the ability to correctly forecast the outcomes of unseen data (Espahbodi and Espahbodi, 2003). To this end, we constructed our model using the training set and subsequently validated its predictive prowess using the testing set. Panel A of Table 2 presents the variable description. Panel B report the summary statistics. The litigation outcome has a mean value of 0.548, referring to a 54.8% winning rate, Suggesting a balanced distribution of litigation outcomes.

3.2 Machine Learning Models

To perform our IP litigation outcome classification task, we use a broad range of machine learning methods. The machine learning models we use are: (1) Logistic regression (LR), (2) Random forest, (3) Support vector machine (SVM), (4) XGBoost and (5) Deep neural network (DNN). These machine learning models use textual input derived from one of the three textual analysis methods discussed in the previous section. We select these models due to their widespread use and efficacy in predictive tasks across finance and computer science, including applications like automobile insurance fraud detection (Ayo et al., 2020; Balaji et al., 2021; Wang et al., 2018). Figure 2 provides a step-by-step illustration of our process. The process begins with extracting text data from the Chinese Judgment Online Dataset, which is then preprocessed by classifying win-loss cases and selecting public listed companies. The preprocessed text is segmented using Jieba, and then split into manageable parts with stop words removed. Concurrently, financial data is extracted from the CSMAR Dataset and matched with

the text data. The text data undergoes further processing, including POS frequency analysis, LDA, and STM to derive topic distributions or POS frequency information. Finally, various machine learning algorithms such as Logistic Regression, Random Forest, Support Vector Machine, XGBoost, and Deep Neural Networks are used for model training to predict the outcomes of intellectual property litigation. This flowchart provides a comprehensive overview of the process, from data extraction and preprocessing to text analysis and model training for outcome prediction.

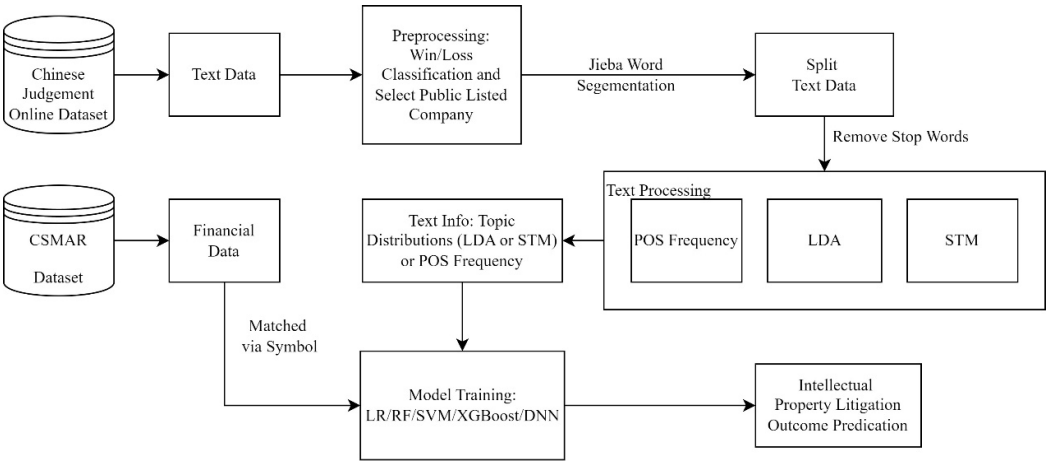

Figure 2. Flow chart of analysis

A potential challenge with our machine learning models is the risk of overfitting, which occurs when the models perform significantly better on the training set than on the testing set. In such cases, the models learn the peculiarities of the training data to an excessive degree that do not generalize well to new, unseen data. To mitigate this, we adopt multiple strategies as outlined by Katsafados et al. (2024). First, we fine-tune the hyperparameters of our models with the Grid Search algorithm using 10-fold cross-validation to minimize the problem of overfitting. More precisely, we use Grid Search algorithm to fine-tune the hyperparameters in LR, RF, SVM and XGBoost. In our Deep

Neural Network (DNN) models, we further mitigate overfitting by incorporating dropout techniques and utilizing early stopping as precautions. More details will be provided in the following sections.

3.2.1 Logistic Regression (LR)

Logistic regression (LR) is a widely used model for binary outcome prediction (Ambrose and Megginson, 1992; Fuster et al., 2022; Katsafados et al., 2024). The LR model, part of the generalized linear models' family, transforms log-odds predictions into probabilities using a sigmoid function. The underlying principle of LR is to estimate model parameters by maximizing the conditional log-likelihood of the training data. This process commonly involves stochastic gradient ascent or related methods. To avoid overfitting the training data, we introduce regularization into the log-likelihood calculation. In our analysis, we utilize L2 regularization, which adjusts the log-likelihood by subtracting the squared L2 norm of the weights vector, scaled by a hyperparameter.

3.2.2 Random Forest (RF)

We utilize the Random Forest (RF) algorithm, an ensemble method developed by Breiman (2001) as an extension of Bootstrap aggregating (Breiman, 1996), in our analysis. We generate multiple uncorrelated decision trees, each trained on a bootstrapped sample of the data with a random subset of features (Mai et al., 2019). During the prediction phase, each tree independently predicts a class, and the final output is determined by majority voting. RF is known to surpass traditional decision trees in performance, particularly in mitigating the risk of overfitting to the training

data. To further combat overfitting, we tune two critical hyperparameters of the RF model: (1) the total number of decision trees, and (2) the tree depth. Specifically, the number of trees is sampled from 100 to 300 in step sizes of 5. Additionally, an appropriate maximum depth of a tree can effectively avoid overfitting and underfitting problem (Want et al., 2018). We evaluate the maximum depth from 10 to 100.

3.2.3 Support Vector Machine (SVM)

We incorporate the Support Vector Machine (SVM) algorithm in our prediction task. SVM is a supervised machine learning model introduced by Cortes and Vapnik (1995). We construct a hyperplane that maximally separates the data points of different classes, with the goal of maximizing the margin between them. To enhance the model's predictive power and to prevent overfitting, we adjust regularization parameters and kernel parameters[8].

3.2.4 XGBoost

XGBoost (eXtreme Gradient Boosting) is an efficient and powerful implementation of the gradient boosting framework, designed to provide robust predictions by combining the outputs of multiple weak predictive models (Chen and Guestrin, 2016). It has gained popularity due to its simplicity, flexibility, and superior performance in a variety of machine learning tasks. XGBoost operates by building an ensemble of decision trees in a sequential manner, where each new tree aims to correct the errors

[8] Specifically, we set the regularization parameter C to sample values from 0.1 to 10, with a step size of 1. In addition, we set the kernel coefficient gamma to ['scale', 'auto'].

made by the preceding ones. The algorithm employs a gradient descent approach to optimize the combination of these trees, using second-order gradients (hence the name) to improve the accuracy of the model. One of the key advantages of XGBoost is its ability to handle a wide array of data types and distributions, making it a versatile choice for our analysis. Additionally, it offers various hyperparameters such as the learning rate and tree-specific parameters like maximum depth and minimum child weight, which are optimized using Grid Search.

3.2.5 Deep Neural Networks (DNN)

Deep Neural Networks (DNNs) are sophisticated computational models that emulate the way the human brain processes information. These networks are composed of layers of interconnected nodes, or neurons, which work together to analyze and learn from complex patterns in large datasets. DNNs are particularly adept at handling high-dimensional data, such as the output of text analysis models. The architecture of a DNN typically consists of an input layer, one or more hidden layers, and an output layer. Each layer contains a multitude of neurons that apply a non-linear transformation to the inputs it receives from the previous layer. By stacking these layers, DNNs can model and make predictions on tasks of varying complexity, from simple linear separations to highly abstract classifications.

Training a DNN involves the backpropagation of errors, where the network's weights are adjusted in response to the prediction error, a process that leverages the power of gradient descent. This iterative optimization allows the network to improve its performance over time, leading to highly accurate models. One of the key strengths

of DNNs is their scalability and flexibility. They can be easily scaled up by adding more layers or neurons to capture more complex features, and they are flexible enough to be applied to a wide range of problems. Additionally, advancements in hardware acceleration, such as GPUs, have made the training of DNNs more efficient, allowing them to tackle larger and more sophisticated tasks.

In our implementation of DNN for predictive analytics, we have designed a layered architecture consisting of 12 layers. The configuration commences with an initial layer of 64 neurons and subsequently expands into a series of layers with an increasing count of neurons, namely, five layers with 128 neurons each, followed by two layers with 512 neurons apiece. This expansion is then strategically followed by a compression phase, tapering down to 256, then 128 neurons, culminating with a single neuron in the final layer to yield the prediction. To improve the model's stability and ability to generalize, each layer includes a normalization step to standardize the activation values, along with a mechanism that randomly deactivates 10% of the neurons to guard against overfitting.

Moreover, to efficiently manage the training process and prevent excessive computational expenditure, we have integrated an Early Stopping mechanism. This involves reserving 20% of the training data for validation purposes and closely monitoring the validation loss (val_loss) throughout training. The training is designed to be halted prematurely after 100 consecutive cycles, or epochs, during which there is no improvement in the val_loss. Importantly, this strategy ensures that the model reverts to the state it was in before this plateau, effectively selecting the best-performing model based on the validation loss within those 100 cycles. This approach strikes a balance

between thorough training and computational efficiency, ultimately leading to a robust and optimized DNN model for our predictive task. The architecture of DNN models is shown in Figure 3. Figure 3 shows a diagram of a neural network architecture designed to process topic distributions and financial variables to predict outcomes of intellectual property litigation. The input layer receives two types of data: topic distributions derived from cases involving intellectual property litigation, and financial variables related to the financial data of the companies involved in the litigation. This data is fed into a hidden layer consisting of multiple neurons that use the ReLU activation function to introduce nonlinearity, allowing the model to learn complex patterns. Additionally, this hidden layer employs a 10% dropout technique, which randomly ignores 10% of the neurons during the training process to prevent overfitting and utilizes batch normalization to normalize the inputs to each neuron, stabilizing the learning process and enhancing the network's speed and performance. Finally, the network passes through an output sigmoid layer, which is used for binary classification tasks, outputting a probability value between 0 and 1.

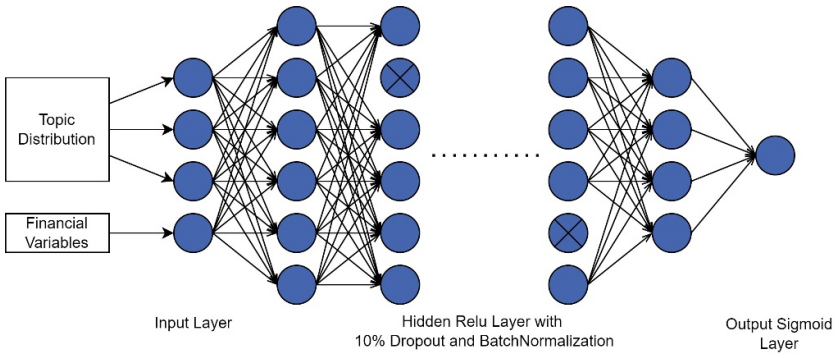

Figure 3. Architecture of DNN models

3.3 Evaluation measure

We evaluate the out of sample performance of our classification models using the Accuracy Score and Area Under the Curve (AUC), which is calculated from Receiver Operating Characteristics (ROC) curve (Hanley and McNeil, 1982; Bradley, 1997). The accuracy score is a metric for measuring the performance of a model. It represents the proportion of all samples that are correctly predicted by the model. The AUC score is computed from the receiver operating characteristic (ROC) curves. The ROC curve plots the true positive rate of the classifier on the vertical axis, and the false positive rate on the horizontal axis, as the classification threshold varies. AUC values are in the range 0 to 1. Higher AUC values imply better out-of-sample classification ability of our models.

4. Empirical results and discussion

4.1 STM results

In our paper, we use the financial variables mentioned above as the input of STM and obtain our baseline result. Following Robert et al. (2019), we choose the appropriate topic number via maximum likelihood method. In our method setting, we choose 30 as our topic number. FREX, which stands for "Frequency-Reversed Exclusivity," is a measure used to identify topic words in a model that analyzes text data. By calculating FREX (see detailed description in Appendix B), we select the keywords of the 30 topics as displayed in Table 3. The first column is the 30 topics, the second column reports the keywords of each topic. Column 3 reports the topic content

by the topic keywords. The table shows that topics 1 to 23 are related to various industries, topics 24 & 25 are related to copyright protection. Topic 26 reports publishing related topic. Topics 27 & 28 report topics related to trademark protection. Topics 29 & 30 report topics related to legal processing. We then use the factual description of the case of each firm in each year as input to train our machine learning model.

4.2 Prediction with financial variables only

As the first step in our empirical analysis, we examine the predictive power on the IP litigation outcome of our models using only financial variables as inputs. We aim to determine whether a firm's financial variables alone fully predict the outcome of IP lawsuits. The dependent variable is binary variables with 1 indicating winning the IP litigation and 0 otherwise. The explanatory variables include financial variables such as leverage, Tobin's Q, ROA, firm size, firm age and board size. The results are reported in Table 4, in which we present the Accuracy and AUC scores for our machine learning models. Accuracy score ranges from 57.0% (SVM) to 76.9% (XGBoost). Furthermore, the most frequently used model in the literature, LR, has an Accuracy score of 58.4%. Thus, XGBoost, RF, and DNN models appear to provide better predictive power than LR and SVM models.

4.3 Prediction with textual features only

In this section, the study investigates whether the factual description of the IP

litigation cases has any predictive power in our prediction task. Table 5 reports our result. Panel A and Panel B report the Accuracy and AUC score respectively for different textual analysis methods. We use the three different types of textual analysis methods, which are STM, LDA and word frequency model. Columns 1 to 5 are different machine learning methods. Our results show that textual feature also have predictive power of litigation outcomes. Furthermore, the results in Table 5 shows that STM has the best predictive power in all machine learning methods in comparison with other two textual analysis methods. This indicates that STM could better capture useful information in the text data, which meets our research expectations.

In comparison with our benchmark models in Table 4, the results are mixed. As shown in Table 5, STM has a better prediction power in every machine learning models. The fact indicates that by accurately capturing topic distribution of the litigation cases, textual information contains vital information for predicting IP litigation outcomes in China. However, the predictive performance of LDA and word frequency model surpasses that of financial data-only methods in certain machine learning applications (such as LR and SVM). Conversely, in other machine learning models, predictions based solely on financial data yield better results. This discrepancy may arise, in part, because LDA and word frequency models do not always accurately extract the pertinent information from text, potentially introducing noise into the analysis.

4.4 Prediction with both financial variables and textual features

This study jointly uses both financial variables and textual features as inputs in our

prediction model. By doing so, we want to investigate whether and to what extent textual information can effectively be combined with financial variables in predicting the IP litigation outcome. Table 6 presents the results of our analysis. Among the three different textual analyses method, STM is the best performing method among all machine learning approaches. In addition, when using DNN as the machine learning method, Accuracy using STM reaches 84.6%, and AUC reaches 83.8%. This indicates a powerful model for predicting IP litigation outcome. Second, in comparison with word frequency model and LDA model, word frequency model has slightly better predictive power than the LDA model. This is probably because LDA model as an unsupervised model, captures noise from the text data and therefore this reduces its prediction power in our studies.

In addition, in most of the models, the prediction performance using both financial variables and textual features are better than our benchmark model (using financial data only) and models using only textual features. For instance, the accuracy score of DNN is 84.6%, which is the best prediction score among our models. Our results indicate that the integration of financial and textual data proves to be the most effective for predictive modeling, as it harnesses the strengths of both types of information to create a more comprehensive and robust prediction. By combining quantitative financial data with qualitative textual data from factual description of IP litigation cases, the model accounts for the varied financial health of companies when assessing the potential outcomes of litigation. Financial data offers a quantitative snapshot of a company's stability and solvency, which are critical factors in litigation where economic resources

can significantly influence the strategy and resolution of legal disputes. Textual data from the case descriptions offers qualitative insights into the specifics of the legal claims, the nature of the IP involved, and the strategic and operational implications for the companies concerned. This dual perspective enhances the model's ability to forecast the potential results of IP litigation more accurately, providing a more reliable and informative basis for strategic decision-making.

From another perspective, as shown in Table 6, STM achieves the best prediction power among all machine learning models. For instance, in DNN model, STM has an accuracy score of 84.6%, while in LDA and word frequency model, the accuracy score are 65.8% and 77.4% respectively. The result is consistent with Table 5. STM's advantage lies in its ability to not only uncover the latent topics within textual data but also to incorporate additional structured information, such as the board information of the firm. It appears that STM better captures the subtleties and patterns within the text that are indicative of case outcomes. This method allows for a more informed extraction of textual features that are predictively valuable. Moreover, STM's flexibility in handling various types of data, including financial metrics and case-specific details, enables it to generate a richer and more accurate representation of the data, which translates to improved predictive accuracy. As a result, STM has proven to be a more robust tool for our predictive modeling needs, in the context of IP litigation.

4.5 Further Analysis and Interpretation

4.5.1 Importance of textual features

To further elucidate the critical role of textual characteristics, we have implemented the Gini impurity method (Kurt et al., 2008; Katsafados et al., 2024). In essence, this method calculates an important score for each variable within the model and is particularly utilized in tree-based models such as RF and XGBoost models. Consequently, we have determined the Gini importance scores for the top 25 features in our RF and XGBoost models. The focus on the top 25 features is deliberate, considering that the number of textual features we have significantly exceeds the financial variables. Subsequently, we have computed the aggregate of these scores for both textual features and financial variables separately. Table 7 shows the Gini importance scores for the predictive models. Upon comparing these totals, it is evident that textual features consistently emerge as more influential inputs than financial variables across all scenarios, and by a considerable margin. For instance, the test Gini score for STM in the RF model is 0.764, while the financial Gini score is only 0.082. This finding aligns with our baseline results, which has established the significance of textual features in forecasting the outcomes of IP litigation in China. Table 8 also shows that using STM provides more influence to text data over financial data. However, the results of the Gini importance score are different in WF. The text Gini importance score is 0.522 for RF model, and 0.369 for XGBoost model, while the financial Gini importance score for RF model is 0.478, and 0.631 for XGBoost model. The results indicates that by counting the word number can not accurately gain the relevant information of text.

4.5.2 Identification of important textual features in litigation outcome prediction

Employing the STM, we have effectively identified topics that carry substantial weight in the litigation outcomes for listed companies, differentiating those that are more closely associated with victory versus defeat in legal disputes. By further visualizing the thematic content extracted from legal narratives, we have introduced a layer of interpretability that significantly aids in forecasting the results of IP litigation. Table 8 reports the top 10 important topics in our prediction tasks.

Table 8 demonstrates that legal cases that are tightly linked to areas such as trademark protection, copyright, or visual content creation are often correlated with adverse litigation outcomes from the defendants' perspectives. This correlation may stem from the tendency of publicly listed companies to infringe upon these areas of intellectual property, increasing the likelihood of unfavorable outcomes in lawsuits for the defendants. For example, trademarks are important symbols for corporate brands and if there are loopholes in the protection of trademarks, such as the failure to register trademarks in time and the non-standard use of trademark, it is easy for the defendants to be complained by competitors in lawsuits, so that they are passive in trademark ownership disputes or infringement litigation. Copyright involves the creation, dissemination and use of works. In copyright litigation, if a defendant lacks reasonable management and effective proof of copyright, such as the inability to provide key evidence, it is easy to be identified as infringement or unable to safeguard its own copyright rights and interests, leading to the loss of the lawsuit. In addition, the standard for infringement judgment on visual content creation is relatively flexible. It is difficult

for defending companies to accurately assess the infringement risk of their own works, and therefore it is easy to lose the lawsuit due to insufficient evidence or inability to effectively defend. Furthermore, the theme for the industrial parts and component is linked to higher likelihood of losing litigation for the defendants. This may be because the defending companies may not apply for patents for core parts in the process of technology development. Once the infringement is involved, due to the complexity of the product and the extensiveness of the supply chain, it is difficult for defendants to prove that they are not at fault.

Our STM analysis has uncovered that the nature of litigation within certain industries can significantly improve the odds of success for the companies involved. For example, China's lighting sector, now relatively mature has seen a more measured pace of innovation and product evolution. As a result, this sector has experienced a more subdued rate of development, which in turn corresponds to a lower incidence of litigation. For defending companies, certain industry-specific topics such as "Electrical Appliance", "Furniture Industry", and "Wine Industry" sectors are also linked to a higher likelihood of winning their legal disputes. The positive correlation with these themes can be attributed to several factors. For instance, the Electrical Appliance sector's success may be due to its highly competitive and innovation-driven nature, where firms actively protect their intellectual property and are more adept at navigating the complexities of patent litigation. The Furniture Industry, with its focus on design and craftsmanship, might see more favorable outcomes because of the emphasis on original design and the clear differentiation of products, which can be easier to defend

legally. Besides, when defending companies highlight the authenticity of their claims within the legal context, they are more likely to achieve a positive outcome in court.

4.5.3 Predicting IP litigation for plaintiff, defendant as well as for different lawsuit types

After confirming the advantage of using STM over LDA and word frequency models, we apply STM prediction on subsamples to gain deeper insights into IP litigation outcomes. These predictions cover various perspectives, including outcomes for plaintiffs and defendants, as well as different types of lawsuits such as copyright, trademark, and patent cases. The results are summarized in Table 9 below.

The results indicate that when predicting the outcomes of IP litigation, the STM demonstrates good predictive performance across various perspectives, including outcomes for plaintiffs and defendants, as well as different types of IP cases (copyright, trademark, and patent). This indicates that STM can effectively capture key factors influencing litigation outcomes, including the nature of the case and the roles of the parties involved. Although STM's predictive performance is good for both plaintiff and defendant samples, the predictive effect for the plaintiff sample (Panel A) is slightly better than that for the defendant sample (Panel B). This suggests that in IP litigation, factors affecting the plaintiff's chances of winning may be more easily identified and captured by STM. The predictive effects for copyright and trademark cases are similar and both outperform the patent cases (Panels C and D outperform Panel E). This may imply that the outcomes of copyright and trademark cases are easier to predict, possibly

because these cases involve less technical complexity or have clearer legal standards and precedents. In contrast, patent cases may be more difficult to predict due to high technical complexity, variable legal standards, or strong case specificity.

4.5.4 Predicting IP litigation in cross-border cases

The paper further investigates whether the STM could demonstrate good predictive performance on IP litigations which involved cross border cases. In our sample, we have 189 cross-border cased in total, which contains 110 cases which involves the U.S entities, and 79 cases which involves the European entities. Table 10 presents the results. The results show that the STM demonstrates good predictive performance in most of the machine learning models. This indicates that STM can effectively capture key factors on IP litigations which involved cross-border cases. Furthermore, the results indicate the robustness of the STM model is not limited to domestic cases but extends to international scenarios as well. This adaptability highlights the versatility of the STM algorithm in various legal systems and jurisdictions, which is crucial for global IP litigation analysis.

We further explore the topics related to winning and losing the lawsuit with respect to the cross-border cases. The results are shown in Table 11. Thematic topics in “Photographs, Furniture, Online Shopping, Air Conditioner, and Music pieces” are more likely to be linked to losing cases for defending companies. For the furniture industry, there may be similarities in furniture design in different countries and regions. If multinational enterprises do not fully understand the intellectual property laws and

regulations of various countries, they are prone to inadvertently infringe others' patents or copyrights in the design and development. Air conditioning industry involves many complex technologies, such as compressor technology. If the patents of these technologies are not fully examined, it is easy to be sued for infringement of others' patents. In countries with high patent barriers, and the risk of losing the lawsuit is higher. For music or photographic works, the copyright management is relatively complex, and the legal provisions for the ownership and use of copyright vary greatly in different countries. Enterprises may lose the lawsuit because they are not familiar with it. Finally, when online shopping platforms engage in cross-border sales, the goods often originate from a wide range of sources with varying quality. If a multinational company fails to rigorously vet and manage its third-party merchants, it may face lawsuits for selling infringing products.

On the other hand, thematic topics in “Red wine, Toy, Software, Lighting industry, Animation and Comics” are more likely to be linked to winning cases for defendants. The red wine industry focuses on quality and brand reputation, and multinational enterprises in wine industry usually invest a lot of resources in brand building and maintenance. In IP litigation, defending companies can leverage a strong brand reputation as compelling evidence of their commitment to IP protection, making it easier for them to win the litigation. Toy products must comply with strict safety and quality standards. Multinational enterprises usually excel in product certification and quality control, which increases their chances of winning litigation. The animation industry relies heavily on creativity. Multinational animation companies often

demonstrate a strong awareness of copyright and place significant emphasis on protecting intellectual property rights of their animation works.

4.5.5. IP Litigation and financial performance

In this section, we analyze the relationship between IP litigation outcomes and the financial performance of companies. We hypothesize that winning firms will experience a positive impact on their financial performance. To explore this, we first examine the connection between IP litigation outcomes and firms' Return on Assets (ROA) and Innovation.

Table 12 presents the regression results on the relationship between IP litigation outcomes and firms' financial performance. In columns 1 & 2, ROA serves as the dependent variable, and the results indicate a significantly positive relationship between the number of wins or win ratio and ROA. The *Win_num* is the natural logarithm of the total number of win cases plus 1for a firm in a certain year. *Win_over_lose* measures the win ratio of the firm, it calculates the ratio of the total number of win cases over lose cases for a firm in a certain year. In columns 3 & 4, the dependent variable is innovation, measured by the natural logarithm of the number of patents.The findings suggest that winning IP litigation case positively relates to subsequent innovation, as reflected in the number of patents. This evidence supports the view that robust legal protection of IP promotes innovation.

4.5.6. IP Litigation and Asset Pricing

In this section, we further investigate whether the capital markets can capture

signals from firm IP litigation outcome. Following the narrative asset pricing theory (Bybee et al., 2023; Kelly et al., 2023), we applied Instrumented Principal Component Analysis (IPCA) to measure the model's explanatory power when incorporating IP litigation texts, comparing it with traditional factor models.

The IPCA model is a conditional factor pricing model that offers a comprehensive approach to incorporating a large amount of conditional information into the factor model estimation, thereby improving the model's predictive efficiency for latent factors. When constructing observable factors, existing literatures have constructed "exclusive" factors in their fields, which will lead to the problem of "factor zoo" (Cochrane, 2011). In contrast, IPCA does not rely on explicitly designated factors; instead, it constructs latent factors from a large set of features, which are then used to build the factor models.

We utilize IPCA to explore the market's reaction mechanism to firm litigation and confirmed the market's ability to recognize and absorb signals released by firm's litigation. Specifically, the IPCA model better describes systematic risk, and we expect it to perform well in explaining excess returns. We used Liu, Stambaugh, and Yuan (2019)'s three-factor model for China's A-share market as the benchmark. In particular, We use their Chinese market, size, and value factors based on China's A-share market as the benchmark CH-3 factor model[9]. We obtain other financial data from the CSMAR database. As per Kelly et al. (2023), we used the "Total R^2" of bond yield fluctuations explained by common risk factors in the panel as the evaluation criterion. We trained

[9] The CH-3 factor data are drawn from Robert F. Stambaugh's homepage in Liu et al. (2019).

the IPCA model using the 30-topic distribution and Chinese three-factor data. We first calculated the Total R^2 of traditional factor models and the IPCA model. We then investigate whether litigation outcomes (win/loss) affect the model's predictive ability. The results are shown in the Table 13.

Table 13 shows that IPCA has achieved good results. In the prediction of the whole sample and the subsamples (Win/Loss), the total R^2 value is 47.9%, 59.0% and 48.2%, respectively, which is much better than the CH-3 factor model (23.3%). The traditional three factor model performs poorly and has deficiencies in the impact on asset pricing. These traditional factors may not be able to fully capture the asset price fluctuations caused by narrative factors. When considering the text of the judgment document, it has a better prediction effect on the capital market, that is, the capital market can capture the information disclosed by enterprises involved in litigation, which will help improve prediction ability. Secondly, the study shows the response of the capital market to the judicial ruling. When the sample is divided into winning and losing cases, the IPCA model still shows better prediction ability than the CH-3 factor model. However, the R^2 of the winning sample is higher than that of the losing sample. This suggests that a company's victory in a lawsuit often conveys a positive narrative, making the signal clearer and easier to detect in the market.

5. Conclusion

In this study, we utilize several machine learning models to predict IP litigation

lawsuits' outcomes in China, with a novel approach that incorporates textual information from the factual narratives of the legal disputes into our predictive framework. Our primary innovation lies in examining the potential of linguistic elements within the lawsuit descriptions to enhance the predictive capabilities of our classification models, beyond the scope of conventional financial metrics. Furthermore, we introduce the STM for textual analysis within our study, discovering that a more refined selection of topic distributions within the textual data can lead to improved predictive accuracy. The STM proves particularly adept at handling such nuanced textual analysis, making it an optimal choice for our research context.

We harness the Word2vec algorithm coupled with a keyword search technique to compile a dataset encompassing 4,516 firm-year instances of IP legal disputes from the Chinese Judgement Online database. For the construction of our textual features, we leverage the STM, LDA, and straightforward word frequency techniques. Subsequently, these textual features, either in isolation or in conjunction with financial metrics, are utilized as inputs for our machine learning models. Our objective is to assess the extent to which the integration of textual features can enhance the predictive performance of our established benchmark models.

Our research yields compelling evidence that underscores the significance of textual information in the context of predicting outcomes for IP litigation. Indeed, the incorporation of textual data into our benchmark models has yielded superior predictive performance. Notably, the STM has emerged as the most efficacious method for textual analysis within our studies, demonstrating the most accurate predictive capabilities. To

substantiate our results, we have conducted t-tests, which confirm that our findings are statistically robust.

Moreover, we have implemented additional robustness checks to ascertain the extent to which textual features contribute to the predictive power of our models in the IP litigation outcome prediction task. These checks have quantified the importance of textual elements, reinforcing the central role they play in enhancing the accuracy of our predictive analytics. Additionally, the research reports topics related to win cases and lose cases, respectively, which provides explainable power to our findings. Our findings suggest that in the IP litigation outcome prediction task, by using STM method and integrate with financial variables are most informative approach. This is in line with our intuition that textual information is meaningful for this task.

This paper documents that the STM model exhibits strong predictive performance across diverse litigation cases involving both plaintiffs and defendants, as well as different intellectual property types such as copyrights, trademarks, and patents. In addition, the STM model also provides better prediction for IP litigation involving cross-border U.S. or European companies. Using the STM models, we are able to identify five leading thematic topics linked to the win/loss outcome of IP litigation for full sample as well as for the cross-border subsample. This suggests the practical feasibility of the STM model in predicting the IP litigation outcome. Furthermore, we show that winning IP litigation is significantly associated with a plaintiff's return on assets (ROA) and higher levels of innovation. In additional asset pricing tests, an Instrumented Principal Component Analysis (IPCA) that incorporates IP litigation texts

perform better than the traditional factor models.

The findings of this study offer significant insights for companies in the management and strategic formulation of IP litigation. Firstly, companies should recognize the importance of textual data from case descriptions and financial data in predicting litigation outcomes. This implies that, in the preparation phase of legal proceedings, in addition to traditional legal analysis, a comprehensive assessment that includes text analysis and financial status should also be incorporated. Moreover, these models can enable companies to better understand the risks and opportunities in the litigation process, thereby making more informed decisions in resource allocation, strategic planning, and long-term decision-making. Thirdly, this study also demonstrates that the combination of legal expertise and advanced data analysis technology can provide companies with more accurate predictions of litigation outcomes, helping them maintain a competitive edge in the context of fierce market competition. In addition, by demonstrating the value of integrating diverse data types and the effectiveness of sophisticated algorithms like STM, we are paving the way for more nuanced and accurate predictive models in the field of IP law.

References

Aharony, J., Liu, C., & Yawson, A. (2015). Corporate litigation and executive turnover. *Journal of Corporate Finance*, *34*, 268-292.

Ambrose, B. W., & Megginson, W. L. (1992). The role of asset structure, ownership structure, and takeover defenses in determining acquisition likelihood. *Journal of Financial and Quantitative Analysis*, *27*(4), 575-589.

Ayo, F. E., Folorunso, O., Ibharalu, F. T., & Osinuga, I. A. (2020). Machine learning techniques for hate speech classification of twitter data: State-of-the-art, future challenges and research directions. *Computer Science Review*, *38*, 100311.

Balaji, T. K., Annavarapu, C. S. R., & Bablani, A. (2021). Machine learning algorithms for social media analysis: A survey. *Computer Science Review*, *40*, 100395.

Bybee, L., Kelly, B., and Su, Y., 2023. "Narrative Asset Pricing: Interpretable Systematic Risk Factors from News Text", The Review of Financial Studies, 36 (2), 4759-4787.

Bereskin, F., Hsu, P. H., Latham, W., & Wang, H. (2023). So Sue Me! The cross section of stock returns related to patent infringement allegations. *Journal of Banking & Finance*, *148*, 106740.

Blei, D., & Lafferty, J. (2006). Correlated topic models. *Advances in Neural Information Processing Systems*, *18*, 147.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993-1022.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145-1159.

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

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123-140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5-32.

Chen, S., Zhang, Y., Song, B., Du, X., & Guizani, M. (2022). An intelligent government complaint prediction approach. *Big Data Research*, *30*, 100336.

Chien, C. V. (2011). Predicting patent litigation. *Tex. L. Rev.*, *90*, 283.

Cochrane, J. H., 2011, "Presidential Address: Discount Rates", The Journal of Finance, 66 (4), 1047-1108.

Cutler, J., Davis, A. K., & Peterson, K. (2019). Disclosure and the outcome of securities litigation. *Review of Accounting Studies*, *24*, 230-263.

Dechow, P. M., Ge, W., Larson, C. R., & Sloan, R. G. (2011). Predicting material accounting misstatements. *Contemporary Accounting Research*, *28*(1), 17-82.

Donelson, D. C., & Hopkins, J. J. (2016). Large market declines and securities litigation: Implications for disclosing adverse earnings news. *Management Science*, *62*(11), 3183-3198.

Doumpos, M., Andriosopoulos, K., Galariotis, E., Makridou, G., & Zopounidis, C. (2017). Corporate failure prediction in the European energy sector: A multicriteria approach and the effect of country characteristics. *European Journal of Operational Research*, *262*(1), 347-360.

El-Haj, M., Rayson, P., Walker, M., Young, S., & Simaki, V. (2019). In search of meaning: Lessons, resources and next steps for computational analysis of financial discourse. *Journal of Business Finance & Accounting*, *46*(3-4), 265-306.

Espahbodi, H., & Espahbodi, P. (2003). Binary choice models and corporate takeover. *Journal of Banking & Finance*, *27*(4), 549-574.

Fuster, A., Goldsmith-Pinkham, P., Ramadorai, T., & Walther, A. (2022). Predictably unequal? The effects of machine learning on credit markets. *The Journal of Finance*, *77*(1), 5-47.

Gao, W., & Chou, J. (2015). Innovation efficiency, global diversification, and firm value. *Journal of Corporate Finance*, *30*, 278-298.

Gandhi, P., Loughran, T., & McDonald, B. (2019). Using annual report sentiment as a proxy for financial distress in US banks. *Journal of Behavioral Finance*, *20*(4), 424-436.

Garcia, D., Hu, X., & Rohrer, M. (2023). The colour of finance words. *Journal of Financial Economics*, *147*(3), 525-549.

Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research*, *241*(1), 236-247.

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535-574.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, *143*(1), 29-36.

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

Katsafados, A. G., Leledakis, G. N., Pyrgiotakis, E. G., Androutsopoulos, I., & Fergadiotis, M. (2024). Machine learning in bank merger prediction: A text-based approach. *European Journal of Operational Research*, *312*(2), 783-797.

Kelly, B., Palhares, D., and Pruitt, S., 2023, "Modeling Corporate Bond Returns", The Journal of Finance, 78 (4), 1967-2008.

Kriebel, J., & Stitz, L. (2022). Credit default prediction from user-generated text in peer-to-peer lending using deep learning. *European Journal of Operational Research*, *302*(1), 309-323.

Kurt, I., Ture, M., & Kurum, A. T. (2008). Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease. *Expert Systems with Applications*, *34*(1), 366-374.

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring corporate culture using machine learning. *The Review of Financial Studies*, *34*(7), 3265-3315.

Liu, J., Stambaugh, R. F., and Yuan, Y., 2019, "Size and Value in China", Journal of Financial Economics, 134 (1), 48-69.

Liu, Q., Wu, H., Ye, Y., Zhao, H., Liu, C., & Du, D. (2018). Patent Litigation Prediction: A Convolutional Tensor Factorization Approach. In Proceedings, International Joint Conference on Artificial Intelligence (*IJCAI)* (pp. 5052-5059).

Long, C. X., & Wang, J. (2015). Judicial local protectionism in China: An empirical study of IP cases. *International Review of Law and Economics*, *42*, 48-59.

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35-65.

Loughran, T., & McDonald, B. (2014). Measuring readability in financial disclosures. *The Journal of Finance*, *69*(4), 1643-1671.

Mai, F., Tian, S., Lee, C., & Ma, L. (2019). Deep learning models for bankruptcy prediction using textual disclosures. *European Journal of Operational Research*, *274*(2), 743-758.

Manso, G. (2011). Motivating innovation. *The Journal of Finance*, *66*(5), 1823-1860.

Medvedeva, M., Vols, M., & Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*(2), 237-266.

Mezzanotti, F. (2021). Roadblock to innovation: The role of patent litigation in corporate R&D. *Management Science*, *67*(12), 7362-7390.

Mezzanotti, F., & Simcoe, T. (2023). *Innovation and appropriability: Revisiting the role of intellectual property* (No. w31428). *National Bureau of Economic Research*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, (26),1188–1196

Palepu, K. G. (1986). Predicting takeover targets: A methodological and empirical

analysis. *Journal of Accounting and Economics*, *8*(1), 3-35.

Phan, M., De Caigny, A., & Coussement, K. (2023). A decision support framework to incorporate textual data for early student dropout prediction in higher education. *Decision Support Systems*, *168*, 113940.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research*, *295*(2), 758-771.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139-1168.

Wang, Y., & Xu, W. (2018). Leveraging deep learning with LDA-based text analytics to detect automobile insurance fraud. *Decision Support Systems*, *105*, 87-95.

Wu, F., Huang, Y., Song, Y., & Liu, S. (2016). Towards building a high-quality microblog-specific Chinese sentiment lexicon. *Decision Support Systems*, *87*, 39-49.

Yuan, H., Lau, R. Y., & Xu, W. (2016). The determinants of crowdfunding success: A semantic text analytics approach. *Decision Support Systems*, *91*, 67-76.

Appendix A. Examples of cross-border IP litigation cases

1. The plaintiff Otto Enm Company claims that the plaintiff is a limited partnership registered in Texas, USA, and owns the copyright of the MDaemon series of email service system software, which has become one of the most popular email server software in the world. The defendant Norn Company, without the plaintiff's authorization and permission, copied and commercially used the plaintiff's MDaemon 13.0.1 software (hereinafter referred to as the software in question), infringing on the plaintiff's computer software copyright. Requesting the defendant to compensate the plaintiff for economic losses and reasonable expenses paid by the plaintiff to stop the infringement, totaling 150000 yuan. The defendant, Norn Company, argued that the defendant had downloaded and used the software in question from the internet, but had no subjective fault and should not be deemed as an infringement; The plaintiff's claim for compensation from the defendant for economic losses and reasonable expenses incurred by the plaintiff to stop the infringement, totaling 150000 yuan, cannot be established and should be rejected in accordance with the law.

2.Tibet Yuedu Ji Company has sued Apple Inc., seeking compensation for economic losses and legal expenses amounting to 72,186 yuan, on the grounds that Apple failed to fulfill its audit obligations, leading to an unauthorized app in its App Store providing online reading services for the book "Xiao Ran Dream: Seventh Anniversary Revised Collection," for which Tibet Yuedu Ji Company holds the information network dissemination rights. Apple argues that it is merely a hardware manufacturer, that the

App Store is operated by other companies in China, and that the App Store cannot review content on third-party servers, thus it should not bear the responsibility for the infringement. Additionally, Apple believes that Tibet Yuedu Ji Company's evidence is insufficient to prove its rights to the work, and that the App Store has provided the developer's identity information without being responsible for auditing the developer's qualifications. The fees collected are for technical services, not for direct economic benefits obtained from the distribution of the work.

Appendix B. STM key steps

(1) Parameter Estimation. This paper conducts word segmentation on the factual description of the litigation lawsuit texts without stop words and constructs a term frequency matrix, denoted as *W*. By inputting the number of topics *K*, the topic prevalence variable *X*, and the topic content variable *Y*, the text topics are analyzed. The selection of the number of topics requires manually defining different quantities of *K*. The final determination of the number of topics is made through a held-out likelihood test, where a higher held-out likelihood indicates a better model fit. In this study, the held-out likelihood is maximized when the number of topics is set to 30.

(2) Model Interpretation. This paper interprets the topic word distribution parameter (β) and the article topic distribution from the STM model to understand the text topic content of enterprises involved in intellectual property litigation. Firstly, the interpretation of the topic word distribution is conducted. Parameter β is a $K \times V$ matrix. Robert et al. (2019) identify topic words through FREX, which considers both traditional exclusivity and semantic coherence indicators. Exclusivity refers to whether different topics are significantly distinct, reflected in the types and weights of words that appear in different topics. Semantic coherence refers to whether the semantics of the words within a topic are coherent, reflected in whether high-weight words in a topic often appear adjacent to each other. A good topic classification model should yield relatively high exclusivity and semantic coherence, also reflected in a higher FREX indicator. The topic keywords extracted through the FREX indicator are representative of the main content of the topic and ensure a certain difference in keywords across

different topics; their definition is as follows:

$$FREX_{k,v} = \left(\frac{e}{ECDF\left(\beta_{k,v}/\Sigma_{j=1}^{k}\beta_{j,v}\right)} + \frac{1-e}{ECDF\left(\beta_{k,v}\right)} \right)^{-1}$$

In the context, the subscript k denotes the given number of the factual description of the litigation lawsuit's topics, and v represents the term frequency matrix of all texts. The parameter β, which controls the distribution of words within topics in Figure 2, is related to the content judgment result Y. When interpreting the connotation, it is the weighted average of the parameters for wins β and losses β, with the weight being the proportion of non-compliant samples. The parameter e represents the topic exclusivity, which Roberts et al. (2016) use with a value of 0.7; the function ECDF(·) represents the empirical cumulative distribution function.

Furthermore, the article topic parameter θ is a $D \times K$ matrix, where its elements represent the proportion of topic k in the dth cases' text. Similarly, by comparing the differences and significance in the topic proportions between winning and losing cases, this paper can derive topic factors that encompass the quality of textual information. In further factor analysis, γ reflects the relationship between the feature variable X and θ, and can be estimated through the following equation:

$$\theta_{d,k} = \alpha + \sum_{n=1}^{N} \gamma_{n,k}\, x_{d,n} + \varepsilon_{i,t}$$

Appendix C Statistical significance test

The research provides supportive evidence that the inclusion of textual features substantially improves the performance of our benchmark models. We also find that STM model achieves the best predictive power in predicting IP litigation outcome due to its advantages in structured information. However, it is important to test the consistency of these results, by including statistical significance tests to validate metric gains. To do so, we conduct t-test for our results. Firstly, we use the t test to check the significance difference of our result and random classification, which assigns a random prediction on each observations. The result is shown in Appendix Table 1. The results show that almost all the p values are statistically significant at the 1% level. The results indicate that our model has significant differences between random classification results.

Appendix Table 1 t-test for difference between machine learning and random classification

	STM	LDA	Word Frequency
LR	(p<0.0001) ***	(p<0.0001) ***	(p<0.0001) ***
RF	(p<0.0001) ***	(p<0.0001) ***	(p<0.0001) ***
SVM	(p<0.0001) ***	(p<0.0001) ***	(p<0.0001) ***
Xgboost	(p<0.0001) ***	(p<0.0001) ***	(p=0.0720) *
DNN	(p<0.0001) ***	(p=0.0144) **	(p<0.0001) ***

The table reports the significance difference of our result and random classification. *, **, and *** denote statistical significance at the 10%, 5%, and 1% levels, respectively.

Table 1. The total number of cases each year and IP lawsuits cases number

Panel A. The total number of litigation cases each year (based on raw data of judgement documents)

2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
197765	220110	397788	1328504	6402566	9719058	12489487
2017	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
16192329	18706723	22221917	22654256	16315756	8834524	543798

Panel B. The total number of IP litigation cases each year

2010	**2011**	**2012**	**2013**	**2014**	**2015**	**2016**
872	1463	5293	9516	28151	37192	59747
2017	**2018**	**2019**	**2020**	**2021**	**2022**	**2023**
82800	104605	151446	184570	99793	50408	768

Panel C. Breakdown of IP lawsuit cases by types

		Listed firms only:		
Types	Total_num (including individuals)	Total	Plaintiff	Defendant
Copyright	192,312	1,424	884	540
Trademark	66,012	1,728	1,672	56
Patent	60,757	1,187	941	246
Other	63,218	177	127	50

This table shows numbers of cases for Chinese Judgement Online for each year and IP lawsuit cases. Panel A provides the total number of cases over the period of 2010 to 2023. The total number of the cases on Chinese judgement online is 136,224,581. Panel B reports the total number of IP litigation for categories such as copyright, trademarks, patent and other types of categories. 'Other' includes contract disputes and other situations where the case type cannot be accurately identified.

Table 2 Variable description and summary statistics

Panel A. Financial variable description	
Variable	Description
Leverage	Leverage ratio, defined as book value of debt divided by book value of total assets measured at the end of year t;
TOBINQ	Market-to-book ratio
ROA	Return on assets ratio, defined as operating income before depreciation divided by total assets, measured at the end of year t
FirmSize	The natural logarithm of the book value of total assets
FirmAge	The total year since the firm started
Boardsize	The number of board members

Panel B. Summary statistics of financial variables and litigation outcome.					
Variable	Obs	Mean	Std. Dev.	Min	Max
Leverage	4516	1.954	4.108	-4.287	64.698
TOBINQ	4516	2.605	2.046	0.774	18.611
ROA	4516	0.038	0.044	-0.441	0.301
FirmSize	4516	9.840	0.568	7.834	12.351
FirmAge	4516	1.201	0.170	0.699	1.732
Boardsize	4516	8.155	1.786	5.000	17.000
Litigation outcome	4516	0.548	0.498	0.000	1.000

The table provides variable descriptions of our key variables and their summary statistics. Panel A is the variable description of the key variables and Panel B reports their summary statistics.

Table 3 STM topics

Topics	Topic	Topic keywords
Topic 1	Wine industry	‘Wuliangye’ (五粮液)(Chinese Wine brand), Yibin(宜宾) (City), Wine(白酒)
Topic 2	Design	Design(设计), Shape(形状), Pattern(图案)
Topic 3	Airline Industry	Aviation(航空), Airline(航线), ‘Chunqiu’(airline company)(春秋)
Topic 4	Visual Content Creation	Photography(摄影), Picture(图片), photographic plate(底片)
Topic 5	Beds	Quilt(被子), Bedspread(床罩), Bedsheet(床单)
Topic 6	Software industry	Microsoft Corporation(微软公司), Microsoft(微软), Software(软件)
Topic 7	High-tech industry	Circuit(电路), Chip(芯片), Controller(控制器)
Topic 8	Comic industry	Alpha Group(奥飞), Comic(动漫), Cartoon Character(卡通人物)
Topic 9	Lighting industry	Oupu(欧普) (Lightng brand), Lighting(照明), Module(模组)
Topic 10	Stationery industry	Chenguang (晨光)(stationery brand in China), Pen(笔), Stationery(文具)
Topic 11	Electrical appliance	Europa(欧派), Electric(电器), Gree(格力)
Topic 12	Furniture industry	Home Furnishings(家居用品), Furnished(家居), Furniture(家具)
Topic 13	Clothing industry	Clothing(服装), Yagor(雅戈尔)(Clothing brand), Nine Shepherds(九牧王)
Topic 14	Wine industry	‘Gujing’(古井), ‘Yuanjiang’(原浆), ‘Gongjiu’(贡酒)(Wine brand)
Topic 15	Industrial parts industry	Screw(螺片), gyroscope(陀螺), base plate(底板)
Topic 16	Food	Ejiao(阿胶)(Food name in China), ‘Haitian’(海天), Shandong(山东)
Topic 17	Toy industry	Appearance Design(外观设计), Toys(玩具), ‘Audi’(奥迪) (Toy brand)
Topic 18	Gold industry	Jewelry(珠宝), Gold(黄金), Precious(贵重)
Topic 19	Furniture industry	‘Oupai’(欧派)(Furniture brand), Kitchen Cabinet(橱柜), Shenyang(沈阳)
Topic 20	Logistics industry	Suning Tesco(苏宁易购), Express Delivery(快递), Cloud shop(云商)
Topic 21	Online shopping	Taobao(淘宝), Alipay(支付宝), Jingdong(京东)
Topic 22	Children goods	Hands on ability(动手), fun(趣味性), children's clothing(童装)
Topic 23	Games industry	Games(游戏), Mobile Games(手机游戏), Online Games(游戏网)
Topic 24	Copyright	Artwork(艺术作品), Copyright Owner(著作权人), Copyright(版权)
Topic 25	Copyright	Musical Work(音乐作品), Picture(图片), Copyright(版权)
Topic 26	Publishing	Book(图书), Publishing(出版), Publishing House(出版社)
Topic 27	Trademark protection	Trademark(注册商标), Trademark Law(商标法), Exclusive Right(专用权)

Topic 28	Trademark authority	Committee(委员会), Trademark Office(商标局), General Administration(总局)
Topic 29	Authenticity	Legitimacy(合法性), No Objections(无异议), Authenticity(真实性)
Topic 30	Legal Proceedings	Withdrawal of Suit(撤诉), Withdrawal(撤回), Dispute(纠纷)

The table reports the keywords of the 30 topics. The first column is the 30 topics, the second column reports the keywords of each topic. Column 3 reports the topic content by the topic keywords.

Table 4 Out of sample predictability performance using only financial variables.

	LR	RF	SVM	XGBoost	DNN
Accuracy score	58.4%	75.9%	57.0%	76.9%	73.8%
AUC score	0.558	0.755	0.547	0.764	0.734

The table presents results of out of sample performance using only financial variables. Financial variables include leverage, Tobin's Q, ROA, firm size, firm age and board size. The paper applies LR, RF, SVM, XGboost and DNN to predict the outcome. Accuracy score and AUC score are used as the evaluation method.

Table 5 Out of sample predictability performance using only textual features.

	LR	RF	SVM	XGBoost	DNN
Panel A. Accuracy score					
STM	72.5%	77.9%	76.3%	79.0%	79.7%
LDA	65.3%	68.6%	63.1%	71.0%	64.0%
Word frequency	66.6%	70.3%	69.7%	70.0%	70.6%
Panel B. AUC score					
STM	0.714	0.768	0.756	0.782	0.794
LDA	0.637	0.671	0.620	0.702	0.633
Word frequency	0.654	0.694	0.680	0.692	0.699

The table presents results of out of sample performance using only textual features. The paper applies three different textual analysis methods, which are STM, LDA and word frequency method. STM and LDA are based on topics, WF is based on word counts. Panel A reports the results of using Accuracy score as an evaluation method. Panel B reports the results of using AUC score as an evaluation method.

Table 6 Out of sample predictability performance using both financial variables and textual features

	LR	RF	SVM	XGBoost	DNN
Panel A. Accuracy score					
STM	73.2%	78.2%	73.5%	79.3%	84.6%
LDA	64.1%	76.5%	63.9%	78.5%	65.8%
Word frequency	67.2%	77.5%	69.8%	79.2%	77.4%
Panel B. AUC score					
STM	0.720	0.770	0.724	0.786	0.838
LDA	0.632	0.752	0.627	0.779	0.657
Word frequency	0.662	0.769	0.682	0.788	0.764

The table presents results of out of sample performance using both financial variables and textual features. The paper applies three different textual analysis methods, which are STM, LDA and word frequency method. Panel A reports the results of using Accuracy score as an evaluation method. Panel B reports the results of using AUC score as an evaluation method.

Table 7 Sum of Gini importance scores

	Text Gini	Financial Gini
STM		
RF	0.764	0.082
XGBoost	0.680	0.087
LDA		
RF	0.689	0.199
XGBoost	0.648	0.164
WF		
RF	0.522	0.478
XGBoost	0.369	0.631

The table reports the Gini importance scores for the predictive models. The higher the Gini importance score, the more significant the feature's role in improving the model's predictive power. The second column shows the Gini importance scores using text features. The third column shows the Gini importance scores using financial variables.

Table 8 Five leading topics linked to win/lose outcomes from defendants' perspectives based on STM.

Negative impact （More likely to lose）	Positive impact （More likely to win）
Trademark protection(商标保护)	Lighting industry(照明行业)
Industry for industrial parts and components(工业零部件)	Authenticity(真实性)
Copyright(版权)	Electrical appliance(电器)
Gold industry(黄金工业)	Furniture industry(家具业)
Visual content creation(视觉内容)	Wine industry(酿酒业)

The table reports the top topics related to win or lose cases. Column 'Negative impact' relates to topics more related to lose cases; Column 'Positive impact' related to cases more related to a win case.

Table 9 STM based machine learning prediction on subsamples

	LR	RF	DNN	SVM	Xgboost
Panel A. STM result for plaintiff					
STM Accuracy	80.83%	84.28%	83.03%	81.93%	81.93%
STM AUC score	0.791	0.830	0.817	0.797	0.811
Panel B. STM result for defendant					
STM Accuracy	78.65%	78.65%	78.65%	76.97%	76.97%
STM AUC score	0.790	0.792	0.791	0.773	0.772
Panel C. STM result for copyright					
STM Accuracy	81.05%	86.67%	86.67%	81.75%	83.16%
STM AUC score	0.777	0.827	0.903	0.786	0.870
Panel D. STM result for trademark					
STM Accuracy	87.86%	86.42%	88.73%	86.99%	85.26%
STM AUC score	0.882	0.867	0.891	0.875	0.854
Panel E. STM result for patent					
STM Accuracy	73.00%	74.68%	79.32%	75.11%	73.84%
STM AUC score	0.679	0.701	0.768	0.701	0.701

The table reports the STM Accuracy scores and AUC scores for various prediction models in cases involving plaintiff, defendant, and different types of litigations such as copyright, trademark, and patents.

Table 10 Out of sample predictability performance on cross border cases

	LR	RF	DNN	SVM	Xgboost
Panel A. Accuracy score					
STM	55.85%	73.94%	73.40%	57.98%	73.94%
LDA	51.02%	69.32%	48.34%	50.99%	67.47%
Word freq	57.45%	62.77%	68.62%	66.49%	64.89%
Panel B. AUC score					
STM	0.559	0.739	0.734	0.580	0.739
LDA	0.481	0.671	0.459	0.480	0.654
Word freq	0.574	0.628	0.686	0.665	0.649

This table presents results of out of sample performance on cross border cases. In our sample, we have 189 cross-border cased in total, which contains 110 cases which involves the U.S entities, and 79 cases which involves the European entities. Panel A reports the results of using Accuracy score as an evaluation method. Panel B reports the results of using AUC score as an evaluation method.

Table 11 Five leading topics linked to win/lose outcomes in cross-border cases from defendants' perspectives.

Negative impact （More likely to lose）	Positive impact （More likely to win）
Photographs (摄影作品)	Red wine (红葡萄酒)
Furniture(家具)	Toy (玩具)
Online shopping (线上购物)	Software (软件)
Air conditioner (空调)	Lightning industry(照明)
Music pieces (音乐作品)	Comics(动漫)

The table reports the top topics related to win or lose cases from defendants' perspectives of cross border cases. Column 'Negative impact' relates to topics more related to lose cases; Column 'Positive impact' related to cases more related to a win case.

Table 12 Impact of IP Litigation on firm's financial performance

	(1)	(2)	(3)	(4)
	ROA	ROA	Innovation	Innovation
Win_num (Log)	0.017***		0.137***	
	(0.007)		(0.044)	
Win_over_Lose		0.005**		0.034**
		(0.002)		(0.017)
Leverage	-0.004***	-0.004***	-0.004*	-0.004*
	(0.001)	(0.001)	(0.002)	(0.002)
SaleGrowth	0.023	0.023	-0.008***	-0.008***
	(0.024)	(0.024)	(0.002)	(0.002)
FirmSize	-0.027	-0.026	0.276***	0.277***
	(0.021)	(0.020)	(0.031)	(0.031)
FirmAge	0.020	0.020	-0.816***	-0.818***
	(0.039)	(0.039)	(0.113)	(0.113)
YEAR	YES	YES	YES	YES
Industry	YES	YES	YES	YES
R2	0.004	0.004	0.023	0.023
N	34,454	34,454	33,769	33,769

This table presents the regression results on the relationship between out-of-sample predictions of IP outcomes and financial performance of companies. In columns (1) and (2), the dependent variable is the Return on Asset (ROA). Among the expandatory variables, 'Win_num' is the natural logarithm of the total number of win cases plus 1 for a firm in a certain year. 'Win_over_lose' is the win ratio of the firm, which is the ratio of the total number of win cases over lose cases for a firm in a certain year. In columns (3) and (4), the dependent variable is innovation, which is measured by the natural logarithm of the number of patents.

Table 13 Comparison of prediction ability between IPCA model and traditional model

	Total R^2
CH-3 factor model (Liu, Stambaugh, and Yuan 2019)	0.233
IPCA（Whole sample）	0.479
IPCA（Win case sample）	0.590
IPCA（Lose case sample）	0.482

The table reports the comparison of prediction ability between IPCA model and traditional model. We use the "Total R^2" of bond yield fluctuations explained by common risk factors in the panel as the evaluation criterion.