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

Haolin Li, Anthony Bellotti, Xiuping Hua

University of Nottingham Ningbo China

Wei Huang

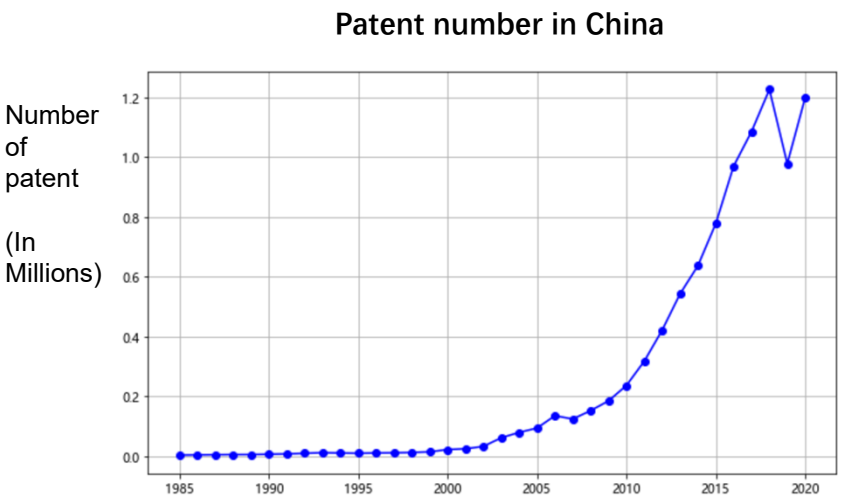
Shidler College of Business, University of Hawaii at Manoa

Haishen Yang

Lingnan (University) College, Sun Yat-sen University

2025 ASSA Poster Session

Patent number and IP litigation lawsuits in China



1.2 millions in 2020

Over 175000 in 2020

IP litigation Cases

- On June 27, 2016, Huawei sued Samsung in Quanzhou Intermediate People's Court for patent infringement, seeking a halt to the infringement and compensation of 80 million yuan for economic losses. Huawei won the litigation.

- In 2001, New Oriental's IPO was postponed due to copyright disputes with ETS and GMAC, resumed after a 2004 court ruling recognized copyright infringement without trademark violation.

- In December 2018, Aux initiated a patent dispute against Gree, claiming infringement on its "compressor" patent, which led to an initial judgment for Gree to compensate Aux with 2.25 billion yuan.

Objectives and main findings of this study

- We aim to enhance the accuracy of predicting outcomes of IP litigation by evaluating NLP text-based models (WF, LDA, STM), leveraging data from China Judgement Online, which contains over 140 million lawsuits spanning 2010 to 2021.

- Evaluates the effectiveness of various machine learning models in predicting the outcomes of intellectual property (IP) litigation.

- 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.

Objectives and main findings – cont.

- 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.
- Also more effective in predicting IP litigation outcomes of cross-border U.S. or European companies.
- Using the STM framework, we identify leading thematic topics associated with 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. In contrast, LDA is unable to achieve the result.

Objectives and main findings – cont.

- Winning IP litigation is positively associated with a firm's return on assets (ROA) and higher levels of innovation.
- In asset pricing tests, an Instrumented Principal Component Analysis (IPCA) that incorporates STM-based IP litigation text features outperforms traditional factor models.
- These findings underscore the importance of robust legal protections for IP in fostering innovation and bolstering corporate financial performance.

Prior studies in IP litigation and prediction with textual analysis

- Patent protection affects valuations (Lerner, 1994); Firms sued for patent infringement provides higher stock returns in the following year (Bereskin et al., 2023).

- Predicting the IP litigation from patent perspectives: patent picture, citation, classifications, abstract, etc. to predict the occurrence of IP litigation (Liu et al., 2018);

- Using machine learning to predict the decision of European court of human rights (Medvedeva et al., 2020); Using Latent Dirichlet Allocation (LDA) model to predict automobile insurance fraud (Wang and Xu, 2018).

Contribution of this study

- Textual information can significantly enhance the predictive performance of IP litigation outcome models when combined with traditional financial variables.

- Of the textual analysis, STM does the best to capture thematic nuances in legal text and improves the predictive power of IP litigation outcome for firms. This implication can be generalized to other languages.

- Identify leading topics linked to win/lose outcomes, which provides enterprises and investors practical tools for decision-making without extensive knowledge-engineering.

Three models in textual analysis

Word Frequency (Loughran and McDonald, 2011, JF)

- tally the frequency of each word type within the legal text of a judgment.
- Limitation: inability to capture the underlying semantic aspects of the text.

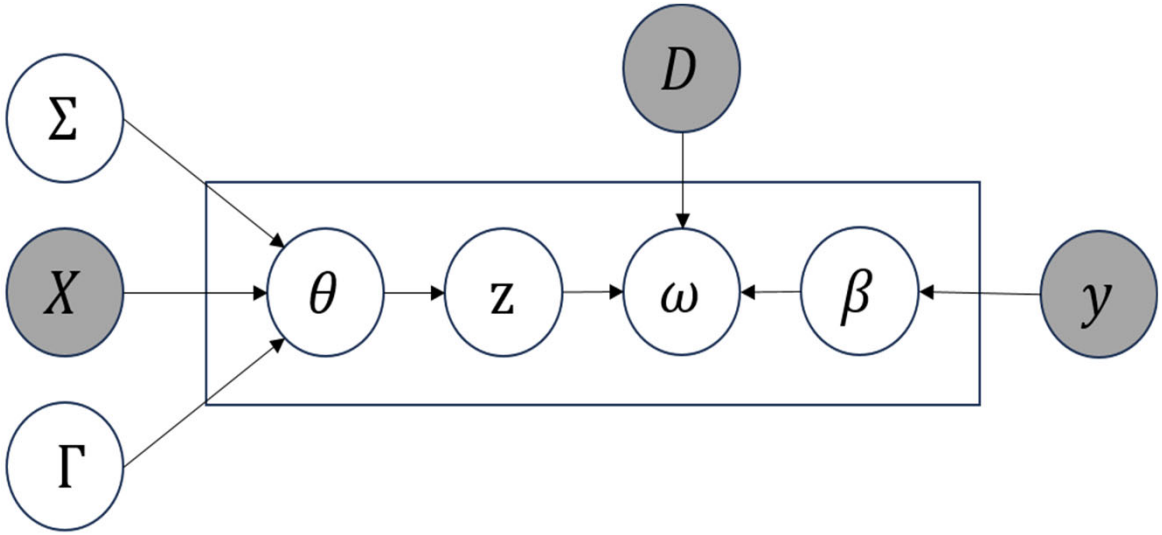
Latent Dirichlet Allocation (LDA) (Blei et al., 2003, JMLR)

- topic model designed to uncover semantic and thematic structures within text
- Limitation： assuming homogeneous firm characteristics such as firm size and cannot account for litigation outcome

Structural Topic Model (STM) (Robert et al., 2016 JASA; Chen and Mankad, 2024, MS)

- Building on LDA, STM can integrate firm characteristics and litigation outcomes in generating topic distributions.

Figure 1. Advantages of the STM (example)
Similar to Robert et al., 2016 JASA



- Grey Circles: Input variables:

- X denotes topic prevalence variables (financial variables)

- Y denotes topic content variables (litigation outcome)

- D denotes textual data (litigation document).

- White Circles:

- A series of latent variables generated by STM.

LDA vs STM

中国乒乓球队谁也赢不了
(No one can win Chinese table tennis team.)

中国足球队谁也赢不了
(Chinese football team can win no one.)

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.
- On the other hand, STM can distinguish the difference by incorporating the thematic constraint related to the speed.

Table 3 STM generated topic keywords (We specify 30 topics)

Topics	Topic keywords	Topic (Defined based on key words)
Topic 1	'Wuliangye' (五粮液)(Chinese Wine brand), Yibin(宜宾) (City), Wine(白酒)	Wine industry
Topic 2	Design(设计), Shape(形状), Pattern(图案)	Design
Topic 3	Aviation(航空), Airline(航线), 'Chunqiu'(airline company)(春秋)	Airline Industry
Topic 4	Photography(摄影), Picture(图片), photographic plate(底片)	Visual Content Creation
Topic 5	Quilt(被子), Bedspread(床罩), Bedsheet(床单)	Beds
Topic 6	Microsoft Corporation(微软公司), Microsoft(微软), Software(软件)	Software industry
Topic 7	Circuit(电路), Chip(芯片), Controller(控制器)	High-tech industry
Topic 8	Alpha Group(奥飞), Comic(动漫), Cartoon Character(卡通人物)	Comic industry
Topic 9	Oupu(欧普) (Lightng brand), Lighting(照明), Module(模组)	Lighting industry
Topic 10	Chenguang (晨光)(stationery brand in China), Pen(笔), Stationery(文具)	Stationery industry
Topic 11	Europa(欧派), Electric(电器), Gree(格力)	Electrical appliance
Topic 12	Home Furnishings(家居用品), Furnished(家居), Furniture(家具)	Furniture industry
Topic 13	Clothing(服装), Yagor(雅戈尔)(Clothing brand), Nine Shepherds(九牧王)	Clothing industry
Topic 14	'Gujing'(古井), 'Yuanjiang'(原浆), 'Gongjiu'(贡酒)(Wine brand)	Wine industry
Topic 15	Screw(螺片), gyroscope(陀螺), base plate(底板)	Industrial parts industry
Topic 16	Ejiao(阿胶)(Food name in China), 'Haitian'(海天), Shandong(山东)	Food
Topic 17	Appearance Design(外观设计), Toys(玩具), 'Audi'(奥迪) (Toy brand)	Toy industry
Topic 18	Jewelry(珠宝), Gold(黄金), Precious(贵重)	Gold industry
Topic 19	'Oupai'(欧派)(Furniture brand), Kitchen Cabinet(橱柜), Shenyang(沈阳)	Furniture industry
Topic 20	Suning Tesco(苏宁易购), Express Delivery(快递), Cloud shop(云商)	Logistics industry
Topic 21	Taobao(淘宝), Alipay(支付宝), Jingdong(京东)	Online shopping
Topic 22	Hands on ability(动手), fun(趣味性), children's clothing(童装)	Children goods
Topic 23	Games(游戏), Mobile Games(手机游戏), Online Games(游戏网)	Games industry
Topic 24	Artwork(艺术作品), Copyright Owner(著作权人), Copyright(版权)	Copyright
Topic 25	Musical Work(音乐作品), Picture(图片), Copyright(版权)	Copyright
Topic 26	Book(图书), Publishing(出版), Publishing House(出版社)	Publishing
Topic 27	Trademark(注册商标), Trademark Law(商标法), Exclusive Right(专用权)	Trademark protection
Topic 28	Committee(委员会), Trademark Office(商标局), General Administration(总局)	Trademark authority
Topic 29	Legitimacy(合法性), No Objections(无异议), Authenticity(真实性)	Authenticity
Topic 30	Withdrawal of Suit(撤诉), Withdrawal(撤回), Dispute(纠纷)	Legal Proceedings

Machine learning methods under three textual anaysis

1. **Logistic regression: Generalized linear model transforms log-odds prediction**
2. **Random forest: Ensemble of decision trees**
3. **Support Vector Machine (SVM)：Max margin classifier**
4. **XGBoost：Gradient boosting framework**
5. **Deep neural network：Multiple layers between the input and output layers**

- Training set: 80% of randomly selected dataset.
- Testing set: Remaining 20% dataset

Sample selection procedure
(From 140 million lawsuit documents)

10 Seed Words: 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

Research Framework

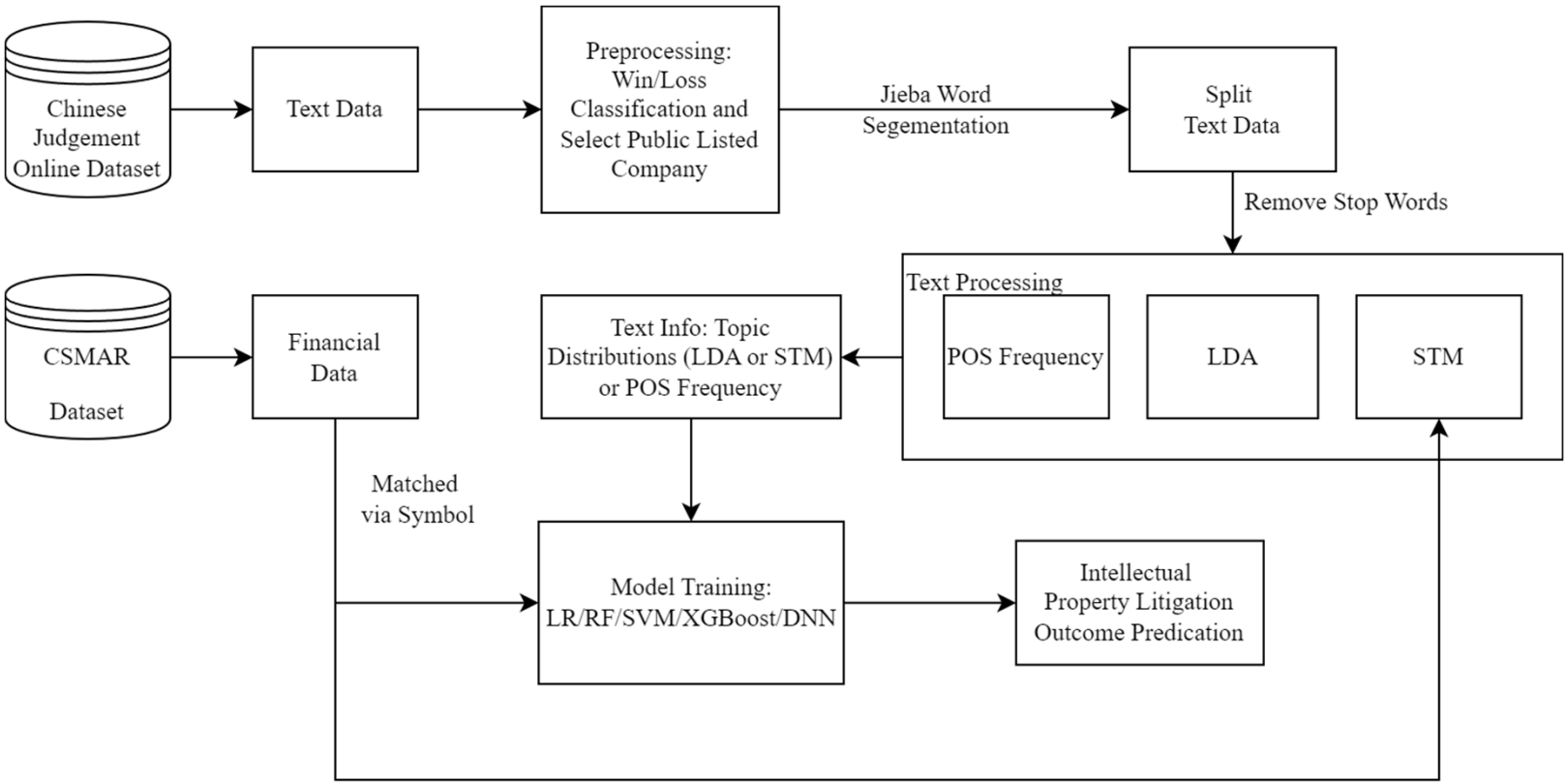


Table 1. The total number of cases each year and IP lawsuits cases number

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

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

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

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;
TOBIN's Q	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
Firm Size	The natural logarithm of the book value of total assets
Firm Age	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 (win = 1)	4516	**0.548**	0.498	0.000	1.000

Predicting Out-of-Sample IP Litigation Outcome

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

(Without textual analysis)

	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

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

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

Table 8 Five leading topics linked to win/lose outcomes.

(Based on STM model)

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)

Table 9 STM based machine learning prediction on subsamples

(STM provides the best performance for all 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

Table 10 Out of sample predictability performance on cross border cases.

(189 cases, STM outperforms)

	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

Table 11 Five leading topics linked to win/lose outcomes in cross-border cases (From domestic firms' perspective)

Negative impact (More likely to lose)	Positive impact (More likely to win)
Visual Content Creation (视觉内容创作)	Wine industry (白酒行业)
Furniture (家具)	Toy industry (玩具行业)
Online shopping (线上购物)	Software industry (软件行业)
Air conditioner (空调)	Lighting industry (照明)
Copyright (版权)	Comics (动漫行业)

Table 12 Impact of IP Litigation on firm's financial performance (Firm-Year panel data).

Win_num: Log (Firm's winning case number +1); Innovation: Log (number of patent+1)

	（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

Table 13 Comparison of prediction ability between narrative-based IPCA model and traditional model.

Instrumented Principal Component Analysis (IPCA)

（Bybee, Kelly, and Su “Narrative asset pricing” 2023, RFS ）

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

Conclusions

- Of the textual analysis, STM does the best to capture thematic nuances in legal text and improves the predictive power of IP litigation outcome for firms. This implication can be generalized to other languages.

- Identify leading topics linked to win/lose outcomes, which provides enterprises and investors practical tools for decision-making without extensive knowledge-engineering.

- IP litigation related narratives improves asset pricing.

Mahalo!
谢谢!
Thank you!
Gracias!