Innovation Novelty and Firm Performance

Group Members: CHAN Yukyee, HUANG Jianming, ZHOU Wanyue
Supervisor: Prof. LI Xin

Agenda

1. Introduction
2. Text-based Measures
3. Citation Network-based Measures
4. Econometric Analysis
5. Conclusion

1

Introduction

Patent & Firm Value

Patent is an important indicator for emphasizing the impact of innovation on firm. However, the patent novelty effects its power in generating profits.

- The uncertainty of payoff for innovation expenditure has been a concerning problem in business decisions (Schwartz 2006).
- Patents are frequently regarded as signals of an organization's innovation capacity, intellectual property, and research and development (R&D).
- Previous studies reported both positive (Useche 2014) and negative (Teece 1986) relationships between a patent and firm value.
- Breakthrough innovations can bring more revenue, so this project focuses on identifying novel patents.

Patent Novelty & Firm Value

The existing measurements for innovation novelty are mainly category-based, citation-based, and content-based.

- Category-based Measures focus on the new recombination of knowledge from different domains.
- Citation-based Measures focus on the citations a patent receives, whereas novel patents should receive more citations.
- Content-based Measures focus on the words and sentence structure of the patent description and claim section.

Overview of the Study

The project focuses on the impact of textual and network-related novelty for patents on firm's abnormal return

- Collect stock data from yahoo finance for 195 firms (2008-2019) & patent data from United States Patent and Trademark Office (USPTO) (2005-2022)
- Develop text-based measures
- Develop two network-related measures
- Analyze the impact of novelty measures on firm's performance

2

Text-based measures

Textual Novelty Detection

Textual novelty Detection can be measured using machine learning techniques, such as **"one-class classification"** novelty detection methods, where a model is established to describe "normal" data (Pimentel Marco et al., 2014).

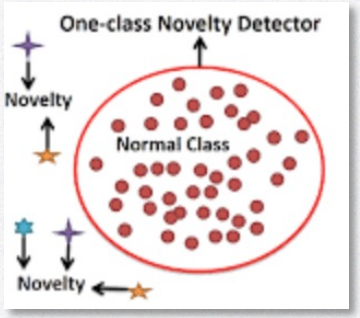


Figure 1. One-class classification

There are various novelty detection methods, including :

- Distribution-based
- Distance-based
- Classification-based

Evaluates existing methods and develop three textual novelty measures :

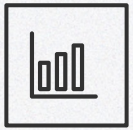
- TFIDF-based Maximum Similarity
- BERT-based Maximum Similarity
- Variational Autoencoder

Literature review

Pimentel Marco et al. (2014) reviewed novelty detection research papers in machine learning, which can be categorized into distance-based, distribution-based, and classification-based methods.

Distance -based

assume known data is clustered together and new data is farther away



Distribution -based

assume known data has its own probability distribution and can be threshold to define different classes

Classification -based

build classifiers to classify whether test data belongs to normal data

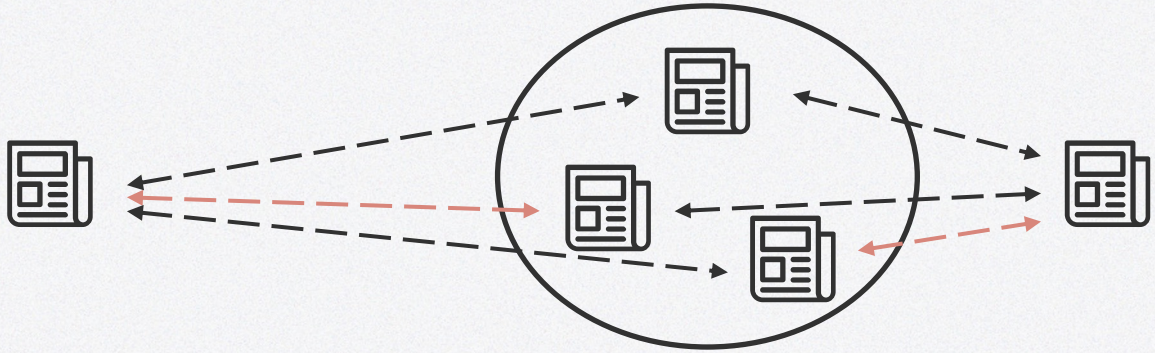
Literature review

Studies	Category	Method	Evaluation Metrix
(Shibayama et al., 2021)	Distance-based method	Q-percentile Similarity Method	Pearson Correlation
(Gerken & Moehrle, 2012)	Distance-based method	Maximum Similarity Method	Spearman's rank correlation coefficients, Recall and Precision
(Luo et al., 2022)	Distance-based method	Maximum Similarity Method with BERT	Correlation
(Hautamaki et al., 2004)	Distance-based method	KNN Graph	Receiver Operating Characteristics (ROC)
(Adarsh et al., 2021)	Classification-based method	Using Latent Dirichlet Allocation with Auto-Encoders	Precision and recall
(Mei et al., 2018)	Classification-based method	Using Semantic Clustering and Autoencoders	Pearson correlation
(Bhattarai et al., 2020)	Classification-based method	Tsetlin Machine Text Classifier	Accuracy
(Hendrycks & Kevin Gimpel, 2016)	Distribution-based method	Threshold decision in PDF	Area Under the Receiver Operating Characteristic curve (AU- ROC), and Area Under the Precision-Recall curve (AUPR)

Table 1. Literature review on textual novelty

TFIDF-based Maximum Similarity

Idea: To measure the originality of each new document, **we compare it with a set of earlier documents**, from which to calculate the textual novelty score.



TFIDF-based Maximum Similarity

Figure 2.
Sentence-BERT

TFIDF words embedding

Find the importance of a term in a set of documents

Maximum similarity

Find the most similarity between the current document and every document.

1 — 2 — 3 — 4 — 5

Data Preprocessing

Remove stop words, punctuation

Cosine similarity calculation

Calculates the similarity between each document

Novelty score calculation

Novelty Score :
$TFIDFMS = 1 - MaxSim$
i.e. , the most novel

BERT-based Maximum Similarity

Figure 2.
Sentence-BERT

1 Data Preprocessing

Remove stop words, punctuation

2 BERT fine-tune & words embedding

Getting a fixed-384 dimensional Contextualized Word Embeddings

3 Cosine similarity calculation

Calculates the similarity between each document

4 Maximum similarity

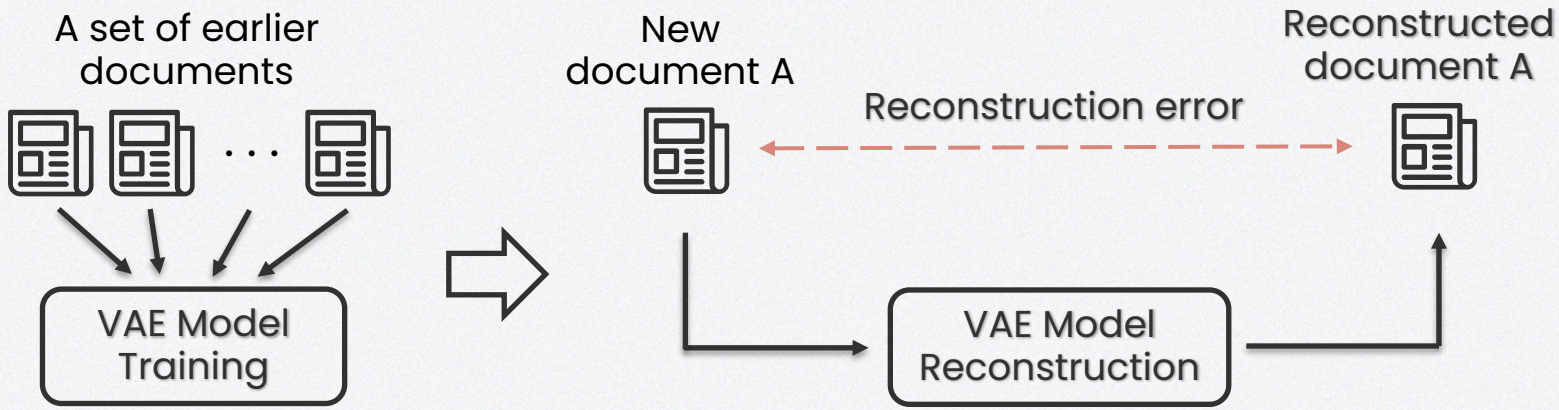
Find the most similarity between the current document and every document.

5 Novelty score calculation

Novelty Score :
$BERTMS = 1 - MaxSim$
i.e. , the most novel

Variational AutoEncoder

Variational Autoencoder (VAE) is used for dimensionality reduction and can **detect novel data based on reconstruction errors**.



Cosine similarity of the **predicted data** and **true data** is used to measure reconstruction error/loss for novelty score.

Novelty Score is defined as : $VAE_i = 1 - cosine\ similarity(p_i, a_i)$

$NoveltyScore_i =$ novelty score for the i^{th} data instance
$p_i =$ predicted data for the i^{th} data instance
$a_{i=}$ actual data for the i^{th} data instance

Variational AutoEncoder

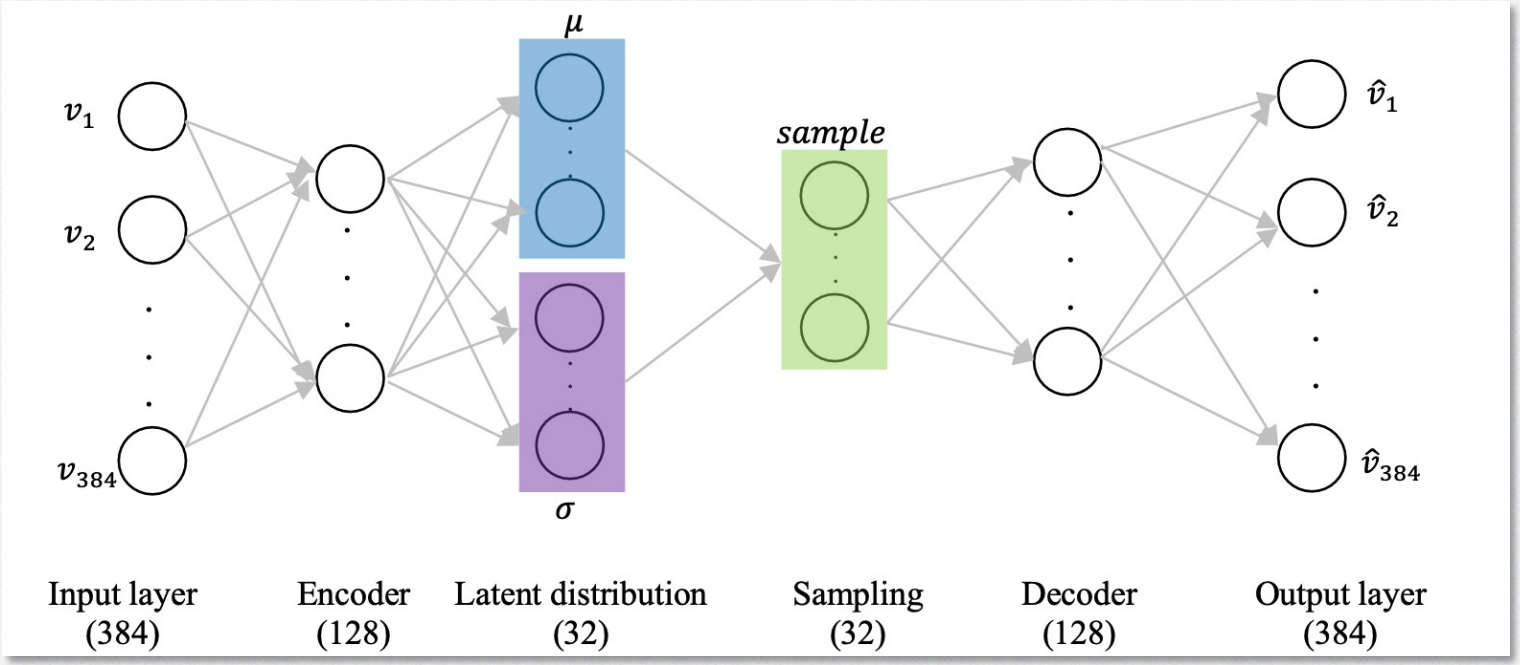


Figure 3. Architecture of Variational AutoEncoder

Measure Evaluation

The evaluation dataset used is the 20 Newsgroups Dataset, which contains 20 categories with 18,828 text documents.

The classes "alt.atheism", "comp.graphics", and "comp.os.ms- windows.misc" are considered normal, and "rec.motorcycles" is considered novel.

Training set: normal

Test set: mixed normal and novel

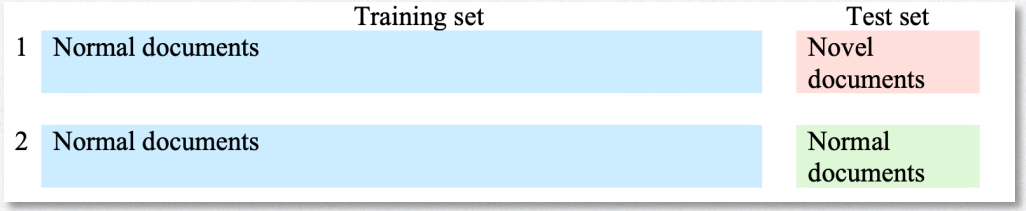


Figure 4. Evaluation Dataset Visualization

- **Pearson correlation coefficient**

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

- **Kolmogorov- Smirnov test**

$$KS = Maxmum|F(X) - F(Y)|$$

- **Jaccard coefficient**

$$Jaccard(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

- **Dice coefficient**

$$Dice(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

Performance Comparison

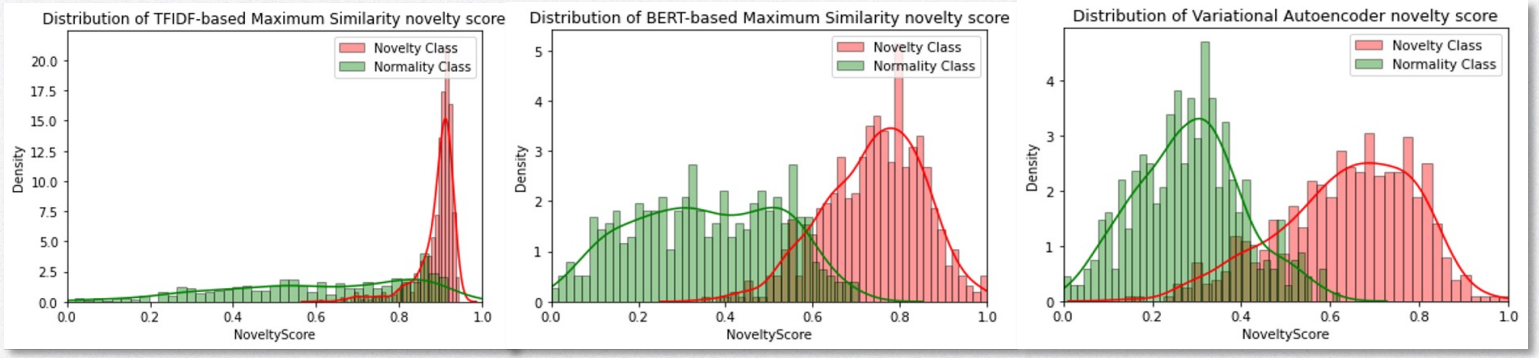


Figure 5. Novelty Score Distribution between three measures

Measures/Metrics	Pearson correlation coefficient	Kolmogorov-Smirnov test	Jaccard coefficient	Dice coefficient
TFIDF-based Maximum Similarity	0.679	0.698	0.313	0.477
Bert-based Maximum Similarity	**0.808**	**0.865**	**0.301**	**0.462**
Variational Autoencoder	0.782	0.802	0.367	0.537

Table 2. Evaluation Metrics Comparison

3

Citation Network-based Measures

Overview of Our Measure

There are various novelty detection methods, including :

- Citation-based
- Reference-based
- Network-based

To measure patent novelty, we constructed a **patent reference network** and develop seven novelty measures based on the network:

- Overlap-based measures(three types)
- PageRank-based measures(four types)

Why are we doing these measures?

- Our measure is **more current** than a cited-based measure(The latter can only evaluate a few years after the patent has been issued)

Literature review

There are three types of indicators are widely used: citation, reference, and network.

Studies	Method	Indicator
Coad et al. (2016)	Analyzing data to explore the impact of company age and innovation on company growth	Reference
Trajtenberg (1990)	Analyzing patent citations to measure the value of innovation	Reference
van Raan (2017)	A New Approach to Using Patent Citation Analysis to Map Technology-Related Research	Reference, Network
Verhoeven et al. (2016)	Developed patent-based technology novelty index	Reference, Citation
McGahan and Silverman (2006)	Analyze the impact of competitors' patents on the value of the company	Citation, Network
Verspagen (1997)	Using patent data to estimate inter-industry technology spillovers	Citation
Leydesdorff and Rafols (2009)	Created a global science map based on ISI topic categories	Network

Data Collection

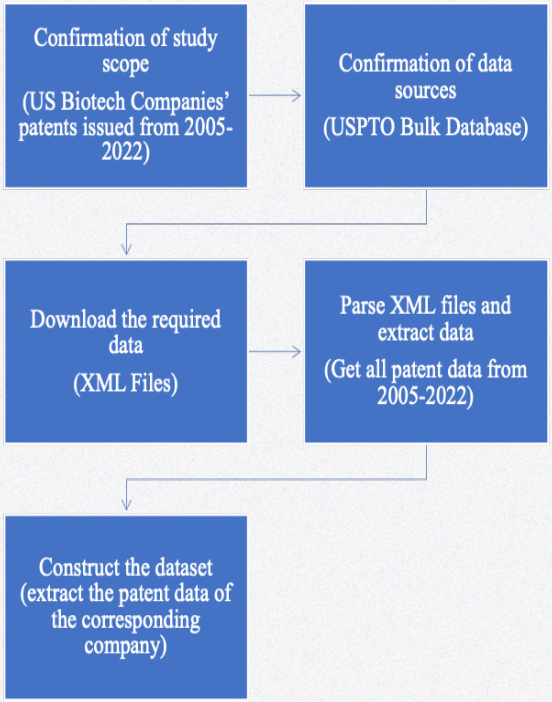


Figure 7. Database Construction

Identification of 196 biotechnology companies:

- Core Patents: Extract their patents(4878)
- First-level Citations: Extract the cited patents(9926)
- Second-level Citations: The cited of cited patents(31011)

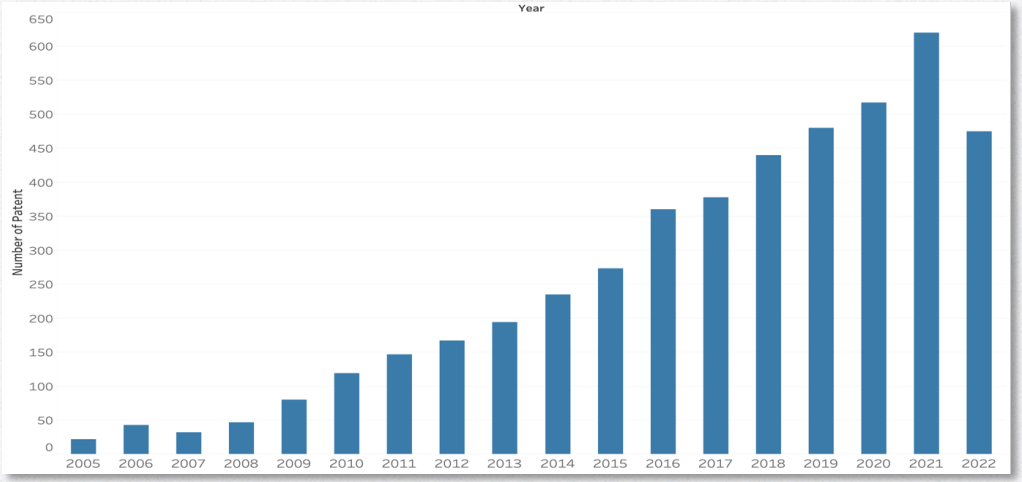


Figure 8. Annual Distribution of Core Patents

Network Construction

Node Definition:

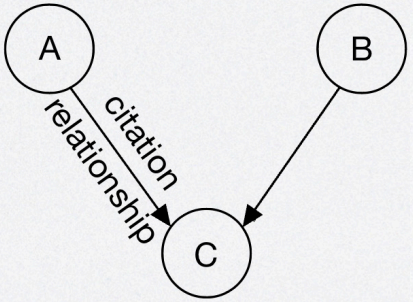
- Core Patents: Extract their patents(4878)
- First-level Citations: Extract the cited patents(9926)
- Second-level Citations: The cited of cited patents(31011)

Relation Definition:

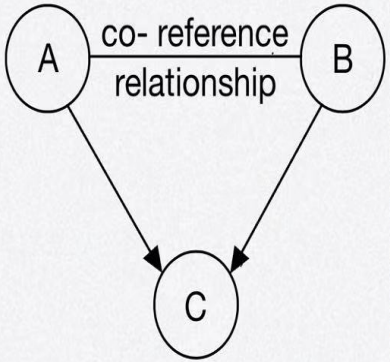
- Citation Relationship(Network1)
- Co-Reference Relationship + Citation Relationship(Network2)

Weight Definition: Overlap of the nodes at both ends of the edge

- Overlap of node's reference
- Overlap of node's classification



Network 1



Network 2

Network Statistics

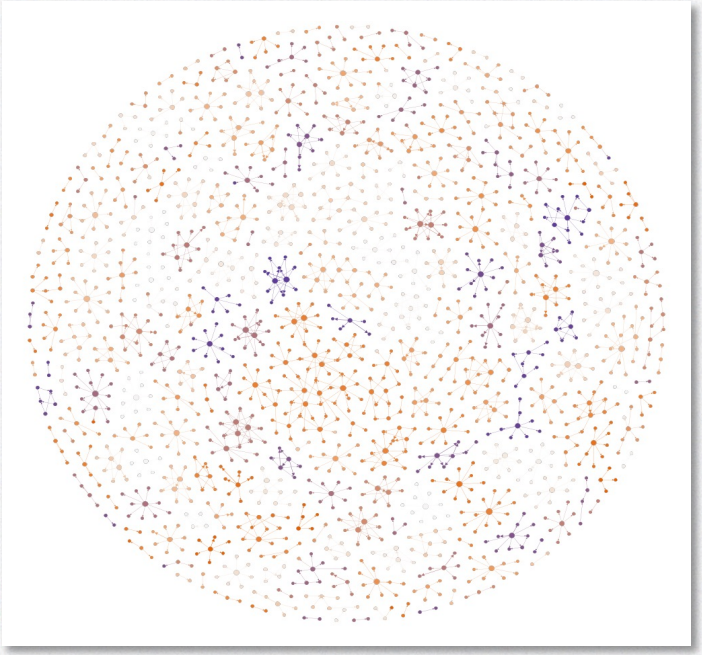


Figure 9. Network1 Visualization

Network Topology Measure	Obs.	Mean	Std. Dev.	Min.	Median	Max.
Degree	14807	3.087415	4.114356	1	1.0	141
Clustering	14807	0.021093	0.125001	0	0	1.0
Density	14807	0.000166	-	-	-	-
Degree Centrality	14807	0.000166	0.000221	0.000054	0.000054	0.00759
Betweennes s Centrality	14807	0.0001	0.001171	0	0	0.052774

Table 3. Network1 Topological Measure

The difference between network 1 and network 2 is very small, only the number of edges is different, 94297 and 97544 respectively.

Overlap

$$overlap\_score_{ij} = \frac{i_c \cap j_c}{i_c \cup j_c}$$

i, j is the set of **references /classifications**.

Compare the **target patent** with its pre-issuance patent.

The **higher** the overlap of target patent, the **more similar** to the previous patent, the **less innovative**.

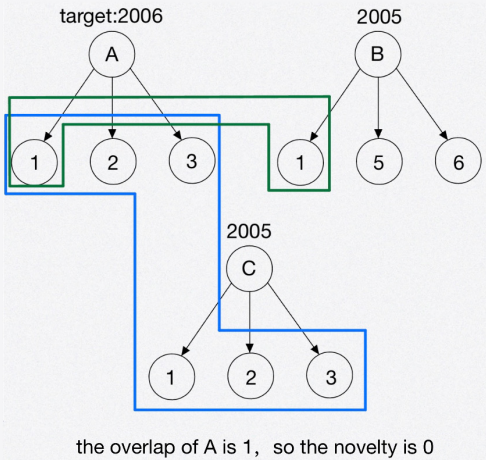


Figure 7. Overlap calculation illustration

Table 4. **Overlap Measurement**

Reference	Measurement 1
Class	Measurement 2
(Reference + Class)/2	Measurement 3

PageRank

We have two networks consisting of different edges and the edges have two different weights:

- Citation relations
- Citation and co-reference relationship
- Overlap of node's class
- Overlap of node's reference

The greater the weight of the node connection, the greater the PageRank and the higher the similarity of the node.

Table 5. **PageRank based measure**

Edge / Weighted	Reference	Class
Citation Relationships	Measurement 4	Measurement 5
Citation relationship + Co- Reference	Measurement 6	Measurement 7

Measure Evaluation

We have three **baselines: Reference Count, Classification Count, Bert** and **seven measurements**, comparing them to the **number of citations** of the patent using **Pearson coefficients** (five/ten years after issue).

$$\text{Pearson correlation coefficient} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2(y_i - \bar{y})^2}}$$

Table 6. **Overlap based measure**

Reference	Measurement 1
Class	Measurement 2
(Reference + Class)/2	Measurement 3

Table 7. **PageRank based measure**

Edge / Weighted	Reference	Class
Citation Relationships	Measurement 4	Measurement 5
Citation relationship + Co- Reference	Measurement 6	Measurement 7

Measure Evaluation

Measurement	Pearson coefficient (10 year citation)	Pearson coefficient (5 year citation)
Measurement 1	0.01571691	0.06563687
Measurement 2	0.122584928	0.05644031
Measurement 3	0.150456829	0.06999825
Measurement 4	0.0111588	0.04340325
Measurement 5	0.012481021	0.04340325
Measurement 6	0.011383788	0.04340543
Measurement 7	0.011888446	0.04340543
Classification Count	0.129365444	0.2978233
Reference Count	0.317578188	0.24148198
Bert	0.006918121	0.06186433

Table 9. Evaluation with citation

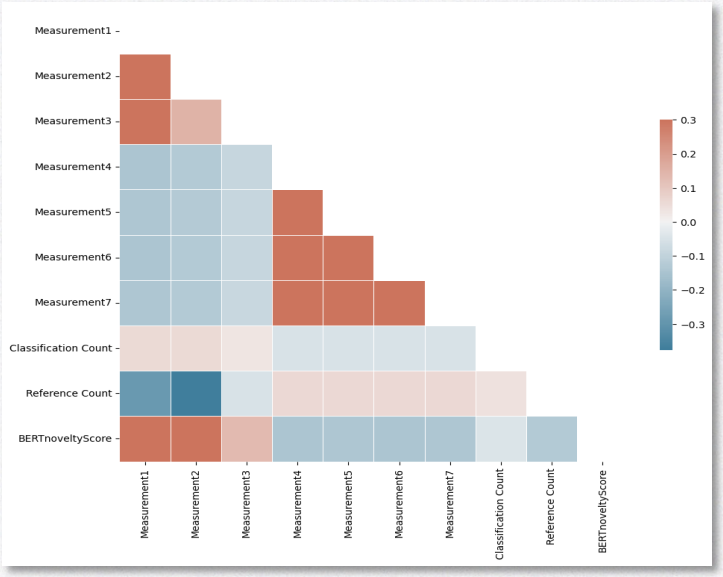


Table 10. Correlation Matrix

Measurement 3 has the highest positive correlation both in comparison with the citation 5 years later or the citation 10 years later. But the effect is worse than the baseline **Reference Count**.

4

Econometric Analysis

Model Setup

Set up a Panel Vector Autoregressive(PVAR) model examine the dynamic interaction between patent novelty measures and firm performance.

The basic model is

$$V_{i,t} = \sum_{k=1}^{K} \Phi_i^k \cdot V_{i,t-k} + Ctrl_{i,t-k} + \delta_t + f_i + \varepsilon_{i,t}$$

As an autoregression model, it considers lag(k), exogeneous control variables($Ctrl_{i,t-k}$), time-fixed effects(δ_t), individual fixed effects(f_i) and within-group bias(standard GMM estimator).

Model Setup

The dependent variable is abnormal return to measure firm equity value following Luo et al. (2013). It is the difference between stock return and abnormal stock return.

The independent variables are the weekly average of text-based measures(BertMs, TfidfMs,VaeMs) and network related measures (overlapref, overlapcls, overlapavg, pagerank).

The control variables are the weekly patent count(n), reference count(RefCnt), and a dummy variable(NoPatent) to represent whether the company publish patent in a week.

PVAR Result

	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$	$AR_{i,t}$
$AR_{i,t-1}$	-0.0350**	-0.0327**	-0.0334**	-0.0407***	-0.0309**	-0.0409***	-0.0321**	-0.0387***	-0.0292**	-0.0394***	-0.0322**
	(0.0140)	(0.0140)	(0.0140)	(0.0140)	(0.0141)	(0.0140)	(0.0141)	(0.0140)	(0.0141)	(0.0140)	(0.0141)
$TfidfMs_{i,t-1}$	0.170***					0.116***	0.139***				
	(0.0164)					(0.0201)	(0.0189)				
$BertMs_{i,t-1}$		0.137***						0.0769***	0.0412**		
		(0.0176)						(0.0226)	(0.0197)		
$VAEMs_{i,t-1}$			0.249***							0.247***	0.130***
			(0.0234)							(0.0265)	(0.0258)
$overlapcite_{i,t-1}$				0.294***		0.291***		0.295***		0.254***	
				(0.0188)		(0.0191)		(0.0190)		(0.0181)	
$overlapcls_{i,t-1}$					0.249***		0.244***		0.248***		0.252***
					(0.0176)		(0.0183)		(0.0175)		(0.0176)
$NoPatent_{i,t-1}$	0.317***	0.305***	0.362***	0.422***	0.397***	0.456***	0.445***	0.446***	0.412***	0.503***	0.467***
	(0.0174)	(0.0172)	(0.0197)	(0.0214)	(0.0219)	(0.0234)	(0.0235)	(0.0230)	(0.0226)	(0.0244)	(0.0232)
$n_{i,t-1}$	0.0392***	0.0392***	0.0494***	0.0366***	0.0507***	0.0436***	0.0584***	0.0409***	0.0516***	0.0537***	0.0572***
	(0.00776)	(0.00765)	(0.00713)	(0.00756)	(0.00788)	(0.00832)	(0.00848)	(0.00813)	(0.00818)	(0.00756)	(0.00763)
$RefCnt_{i,t-1}$	0.0415***	0.0426***	0.0331***	0.0470***	0.0190***	0.0426***	0.0173***	0.0455***	0.0214***	0.0367***	0.0195***
	(0.00567)	(0.00528)	(0.00487)	(0.00664)	(0.00544)	(0.00686)	(0.00580)	(0.00661)	(0.00553)	(0.00606)	(0.00552)
# of Obs.	33816	33816	33816	33816	33816	33816	33816	33816	33816	33816	33816
# of Firms	107	107	107	107	107	107	107	107	107	107	107

Table 11. PVAR result

Discussion

As the additional attempt on econometric analysis, I consider the quadratic form of independent variables, impulse response and exogeneous variables.

- Quadratic property: The quadratic form of independent variables is also significant in the former model.
- Impact of the lag: When the independent variables serve as the impulse and abnormal return as the response, it reaches the peak when there is one lag and the impact eliminate to zero when it reaches four lags(one month).
- Impact of media coverage: Use the google trend of the firm name for representing the media coverage as the exogeneous variable, the coefficient of most independent variables in PVAR increases.

5

Conclusion

Conclusion

This project develop text-based measures and network-based measures for patent novelty. Based on the measures, it verify that there is a positive relationship between patent novelty and firm performance.

Text-based measures highlights the importance of detecting novel concepts to improve the efficiency and performance of identifying the impact of novelty in patent texts on firm value.

Based on the patent citation network, we propose seven measures of novelty, all of which measure the novelty of a patent and can be measured at the time of issuance of the target patent, without waiting to be cited. Improved efficiency in identifying patent novelty

Future direction

The future direction is to conduct additional analysis on the patent novelty measures and econometric analysis.

- Solve endogeneity Issues: Use the patents' citations data for re-calculate the patent measures
- Construct instrument variables:
- Heterogeneity

Reference

Schwartz, J. 2006. The Five Founding Principles That Drive Innovation. The Financial Times.

Useche, D. 2014. Are Patents Signals for the Ipo Market? An Eu–Us Comparison for the Software Industry. Research Policy (43:8), pp. 1299-1311.

Teece, D.J. 1986. Profiting from Technological Innovation: Implications for Integration. Collaboration. Licensing and Public Policy. Research Policy (15:6), pp. 285-305.

Luo, X., Zhang, J., and Duan, W. 2013. Social Media and Firm Equity Value. Information Systems Research (24:1), pp. 146-163.

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using Siamese Bert-Networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language. Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP).

van Raan, A. F. (2017). Patent citations analysis and its value in research evaluation: A review and a new approach to map technology-relevant research. Journal of Data and Information Science, 2(1), 13-50.

Verhoeven, D., Bakker, J., & Veugelers, R. (2016). Measuring technological novelty with patent-based indicators. research policy, 45(3), 707-723.

Reference

Bhattarai, B., Granmo, O.-C., & Jiao, L. (2020). Measuring the novelty of natural language text using the conjunctive clauses of a Tsetlin machine text classifier. *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*.

Shibayama, S., Yin, D., & Matsumoto, K. (2021). Measuring novelty in science with word embedding. PLOS ONE, 16(7).

Coad, A., Segarra, A., & Teruel, M. (2016). Innovation and firm growth: does firm age play a role? research policy, 45(2), 387-400.

Dahlin, K. B., & Behrens, D. M. (2005). When is an invention really radical?: Defining and measuring technological radicalness. research policy, 34(5), 717-737.

Verspagen, B. (1997). Measuring intersectoral technology spillovers: estimates from the European and US patent office databases. Economic systems research, 9(1), 47-65.

Leydesdorff, L. (2007). Patent classifications as indicators of cognitive structures. Annual Meeting of the Society for the Social Studies of Science (4S), Montreal.

Reference

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. Journal of the American Society for Information Science and Technology, 60(2), 348-362.

McGahan, A. M., & Silverman, B. S. (2006). Profiting from technological innovation by others: The effect of competitor patenting on firm value. research policy, 35(8), 1222-1242.

Trajtenberg, M. (1990). A penny for your quotes: patent citations and the value of innovations. The Rand journal of economics, 172-187.

Trajtenberg, M., Henderson, R., & Jaffe, A. (1997). University versus corporate patents: A window on the basicness of invention. Economics of Innovation and new technology, 5(1), 19-50.

The End