Measuring Knowledge Capital Risk

Pedro H. Braz Vallocci [*†]

November 24, 2023

For the latest version of this paper, click here.

Abstract

This study proposes a methodology to identify firms that are vulnerable to knowledge capital-related risks, relying on a textual analysis of the risk factors disclosed in their annual reports. Further, the paper quantifies these risks through an examination of firms' concurrent return patterns.

Keywords: Innovation, firm risk factors, productivity, intangible capital, natural language processing, asset pricing

JEL classification: C43, C55, D2, E22, G11, O3

[*]Ph.D. candidate in Economics, University of California, Santa Cruz

[†]I would like to thank my advisor, Galina Hale, and Alonso Villacorta, Grace Gu, John Fernald, Michael Leung, Brenda Samaniego de la Parra, Chenyue Hu, Roberto Mauad, Bhavyaa Sharma, Ted Liu, and the UCSC Macro Workshop participants for all their valuable comments.

1 Introduction

The transition towards a service- and knowledge-based economy has been accompanied by a sharp increase in intangible assets. Knowledge capital, conceptualized as the aggregate of a firm's investments in research and development (R&D), has become an increasingly significant factor in the valuation of firms (Belo et al., 2019). Firms endeavor to foster innovation through research, aiming to generate intellectual property that may lead to breakthroughs and augment cash flows. Concurrently, firms face the challenges of competition and product obsolescence, which can symmetrically induce sudden devaluations. The episodic nature of innovation implies that firms with substantial knowledge capital are likely to experience more volatile valuations.

However, the existing literature has predominantly attempted to correlate the risk associated with knowledge capital to a firm's R&D or patent intensity (Andrei et al., 2019). Such classifications overlook the reality that the efficacy with which firms convert R&D into tangible innovations—and consequently into increased cash flows—is both industry-specific and firm-specific. Moreover, patenting activities are recognized to be unevenly distributed across industries and firm sizes (Mezzanotti and Simcoe, 2023; Li and Hall, 2020).

Within this framework, several pertinent questions arise: Can we more accurately discern knowledge-related risks by conducting textual analysis of the risk factors delineated in firms' annual reports? Furthermore, if it is possible to identify knowledge-intensive firms in this manner, does this insight impact the way market participants value these entities? In essence, is there a differential in the market's risk assessment for firms with heavy knowledge capital compared to their less knowledge-intensive counterparts? This paper proposes a methodology to identify firms that are vulnerable to knowledge capital-related risks, relying on a topic analysis (LDA) of the risk factors disclosed in their annual reports. Further, the paper quantifies these risks through an examination of firms' concurrent return patterns.

Spending on intangible assets is an investment since it reduces current consumption to increase future consumption (Corrado et al., 2009b,a). Among the components of a firm's intangible capital are knowledge capital, brand capital, and organization capital. In this context, *brand capital* is defined as a firm's accumulated expenses in advertising (Belo et al., 2019); *organization capital*, as the set of an enterprise's unique systems and processes (Eisfeldt and Papanikolaou, 2013; Bloom and Van Reenen, 2007); and *knowledge capital*, as a firm's accumulated investments in R&D.

Accumulated R&D expenses are considered an integral part of firm's capital in supply-

side models such as Belo et al. (2013), accounting for an increasing share of public firms' valuation. Along these lines, Hall (2001); McGrattan and Prescott (2001); Vitorino (2014); Eisfeldt et al. (2020); Li et al. (2014) confirm that intangible capital matters for aggregate stock market valuations and, more specifically, firm-level valuations. Corrado et al. (2009b) estimated total intangible capital in 2003 to be 3.6 trillion, half of which as scientific and non-scientific R&D capital. Crouzet and Eberly (2022) found that the omission of knowledge capital and its associated rents can explain up to 2/3 of the investment gap (the difference between marginal q and average Q) in R&D intense sectors such as Healthcare and Chemicals. On another note, R&D expenditures are considered as intermediate expenses that can spur growth in models with expanding varieties (Romer, 1990) and quality ladders (Grossman and Helpman, 1991; Atkeson and Burstein, 2019).

The riskiness of intangible assets differs systematically from tangible ones (Hansen et al., 2005). Eisfeldt and Papanikolaou (2013) point out that shareholders cannot entirely appropriate the cash flows from the key talent of the firm since the firm must always compensate key talent by its outside option. Eisfeldt et al. (2018) shows that key talent partially owns the cash flow from intangible capital in the form of equity, finding that almost 40% of compensation to high-skilled labor happens as equity-based pay. Finally, Ai et al. (2019) predicts that collateralizable assets, which do not include some categories of intangibles, provide insurance against aggregate shocks in the economy and should earn a lower expected return.

Knowledge capital's risk characteristics set it apart from other forms of intangible capital. Firms invest on research to increase their future earnings, but not all innovations come to fruition. For example, research conducted by a pharmaceutical firm can lead to successful new drugs that lead to patent rents for several years or to no result at all. The presence of innovation-driven jumps in cash flow is related to an increased empirical dispersion of Tobin's q, and also helps explain why the relatxion between Tobin's q and aggregate investment has become tighter since the mid-1990s (Andrei et al., 2019).

Besides the uncertainty of research investments, a firm is also susceptible to writing off part of its knowledge capital due to obsolescence, e.g., when it narrowly loses a patent race (Peters and Taylor, 2017). Such competitive forces lead to R&D capital depreciation. Depreciation rates vary over time and according to individual industry technological and competitive environments (Li and Hall, 2020). Besides varying between industries, it is expected that depreciation rates vary across firms in the same industry, since firms' ability to materialize research into innovation and additional cash flow is also expected to be random, varying with managerial processes, employees' skills, and regulation-related uncertainty.

Knowledge capital heavy firms also are especially vulnerable to loss of key talent. Eisfeldt and Papanikolaou (2013) find that firms are more likely to list "loss of key talent" as a risk factor in their 10-K reports when they have high organization capital. Firms vulnerable to loss of key talent are especially susceptible to immigration-related risks, e.g., the H-1B visa annual quota shortages (Peri et al., 2015).

A prominent reliance on R&D may also entail a different financial risk profile. Li (2011) find that the riskiness of a financially constrained firm increases with its R&D intensity. Additionally, young, R&D-intensive firms frequently face challenges in securing debt finance due to the unpredictable and fluctuating returns on R&D, and the potential adverse selection and moral hazard in the R&D financing market. Notably, fluctuations in the supply of equity finance were instrumental in shaping the R&D surge of the 1990s (Brown et al., 2009), pointing to a shared risk factor among these firms.

The current methods of identifying knowledge capital related risk across firms, deriving from uncertain outcomes of research, use proxies (accumulated R&D, SG&A, patent wealth) that fail to accurately describe potential volatilities in cash flows. A possible approach might involve tying knowledge capital risk to a firm's knowledge capital, normalized by its assets or another variable. However, a significant limitation is the inconsistent R&D reporting standards across industries and firms. This inconsistency is exacerbated by some firms not disclosing their R&D expenditure in annual reports. Similarly, methods that merely accumulate R&D expenses to characterize the knowledge intensity of a firm, using the perpetual inventory method, ignore that depreciation is random and unique to each firm.

Another possible, and similar, approach involves tying knowledge capital risk to a firm's total intangible assets instead, typically through indirect measures like Selling, General and Administrative expenses (SG&A). However this measure is by nature prone to considerable measurement error and includes not knowledge-related components, such as organizational capital and brand capital. This can blur the distinction between knowledge-heavy and non-knowledge-heavy firms.

Lastly, relying on patents as a measure of a firm's knowledge intensity only tells part of the story. While patents reflect the final outcomes of R&D investments, they do not account for the internal learning processes that take place within firms, thus potentially undervaluing those that invest heavily in internal knowledge development, even if they do not have a high patent output. Additionally, measures of patent activity, such as the one in Kogan et al. (2017), are not flawless indicators of the risks associated with R&D-centric firms. Large firms and firms in specific industries are more prone to protecting their intellectual property, which shows that patent distribution does not directly mirror

the exposure to knowledge capital risks (Mezzanotti and Simcoe, 2023).

Kogan et al. (2017) demonstrate that the companies' market value surges within the following three days of the filing of potentially lucrative patents, that is, patents that have the potential to amplify a firm's revenue streams while simultaneously stifling competition. Yet, one must ponder whether the stock market, outside patent-filing moments, also internalizes the inherent risks of devoting a significant part of investments to a risky innovation process.

2 Methodology and Data

In this paper, knowledge capital risk is defined as the unique form of risk that innovation-centric firms bear, particularly when their cash flow trajectories are influenced by the stochastic nature of research-induced innovations. This risk is inherent in the unpredictable timing and impact of such innovations on a firm's financial performance.

Methodologies for textual analysis, such as pattern detection, are diverse and multifaceted. They encompass a range of techniques from simple pattern matching to more complex algorithmic constructions of term sets, as well as the identification of prevailing topics within textual data. These methods are employed to extract meaningful patterns and trends from unstructured text.

The concept of knowledge capital risk, however, lacks an empirical "ground truth," a benchmark against which the accuracy of predictions or classifications can be validated, such as the VIX in Manela and Moreira (2017), used as ground truth to create a measure of news volatility dating back to 1890. Moreover, the construction of a comprehensive dictionary is complex, as the terminology can vary significantly across industries. For instance, the lexicon indicative of intellectual property concerns in pharmaceutical companies may frequently reference terms such as "clinical trials" or "regulatory compliance" within their annual reports.

In light of these complexities, my approach will leverage Latent Dirichlet Allocation (LDA) to analyze the risk factor sections within firms' 10-K filings. This method will facilitate the extraction of dominant topics that are most indicative of knowledge capital risk.

2.1 Latent Dirichlet Allocation (LDA)

In this study, I use Latent Dirichlet Allocation (LDA), a topic modeling technique, to identify latent topics within firms' self-reported risk factors in a comprehensive corpus of

121,839 firm annual reports (10-Ks), covering a timespan from 2006 to 2022.

Latent Dirichlet Allocation (LDA) is a generative statistical model that is widely employed for topic modelling within the field of natural language processing (NLP). LDA relies on the assumption that each document in a given corpus can be seen as a mixture of a certain number of latent topics, denoted as $k \in 1, ..., K$, each of which carries a particular weight, $\omega_{i1}, ..., \omega_{iK}$. Each of these topics is assigned a word probability vector, $`_k$, defining the likelihood of each word appearing under this topic (Blei et al., 2003).

Under this model, if we denote X_i as the vector of word counts with length n_i in the ith document, the word distribution in a document is modeled as a multinomial distribution. The probability of X_i can be written as:

$$X_i \sim Multinomial\left(n_i, \omega_{i1}\theta_1 + \cdots + \omega_{iK}\theta_K\right) \tag{1}$$

The output of an LDA operation is twofold: firstly, it generates a list of topics, with each topic represented as a collection of words. Secondly, it offers a weight distribution across these topics for each document, indicating the degree to which each topic is present in a given document.

It is important to note that LDA is an unsupervised learning method. This means that it operates without any predefined labels, instead learning and inferring patterns directly from the data. However, this also poses a challenge, since the topics generated might not necessarily align with any intuitive description, requiring therefore post-analysis interpretation.

2.2 Data

I retrieve the annual reports (10-Ks) for all U.S. publicly listed firms since 2013 from the Securities and Exchange Commission's EDGAR database. A 10-K is a comprehensive document that provides an overview of the company's financial performance and operations over a year, offering a detailed picture of a company's business. To ensure transparency and accuracy, laws and regulations strictly prohibit companies from making false or misleading statements in their 10-Ks. Additionally, under the Sarbanes-Oxley Act, a company's Chief Financial Officer (CFO) and Chief Executive Officer (CEO) are required to certify the accuracy of the 10-K (SEC: Office of Investor Education and Advocacy (2011)).

Item 1A of a 10-K ("Risk Factors") includes information about the most significant risks that apply to the company or its securities. I extract the item 1A information from each 10-K using XML parsing and `BeautifulSoup`, and removed supposedly less meaningful characters such as punctuation and numbers.

2.3 Filtering firms

Following Golubov and Konstantinidi (2019); Stambaugh and Yuan (2016), I filter firms by considering only ordinary common shares, traded on NYSE, AMEX, and NASDAQ exchanges; and following Stambaugh and Yuan (2016) I exclude those whose prices are less than $5 in 2016 dollars.

Reporting risk factors is mandatory for most firms; however, there are exceptions for asset-backed issuers and smaller reporting companies. Asset-backed issuers are defined as issuers whose reporting obligation arises from either the registration of an offering of asset-backed securities under the Securities Act or the registration of a class of asset-backed securities. Firms that are not required to disclose risk factors either leave Item 1A empty or write a placeholder text specifying that, due to their nature, they are not required to disclose risk factors. Consequently, there is an abnormal frequency of 10-K filings with a significantly low number of words, as depicted in Figure 13. In subsequent stages, 1A texts with an insufficient word count are discarded. The threshold adopted was 200 words. The total count of filtered firms by year is shown in Table **??**.

The next step is to convert each text in the corpus to a bag-of-words format, detailed in Appendix A.

2.4 Topic modeling

Upon having the corpus and the dictionary, I employ LDA to endogenously generate topic loadings for the entire set of documents. During the model's configuration, I designate a parameter to specify the number of topics, represented as k, and supply the model with the dictionary I previously created. Selecting an appropriate value for k is typically done *ad hoc*, with the primary consideration being interpretability, as highlighted by Gentzkow (2019).

A representative output from this modeling approach can be viewed in Figure 1.

Year	Total_1As	Filtered
2006	5685	2466
2007	6445	2714
2008	6931	2305
2009	8244	2190
2010	8122	2290
2011	8019	2356
2012	7797	2316
2013	7560	2401
2014	7560	2518
2015	7531	2528
2016	7196	2431
2017	6896	2394
2018	6804	2418
2019	6683	2404
2020	6531	2332
2021	6936	2308
2022	6899	1885

Table 1: The left column shows the count of all 10-Ks retrieved for a given year. The right column counts all the 10-Ks that obey to the following filtering criteria: 1) Ordinary common shares; 2) Membership to NYSE, AMEX, and NASDAQ; 3) Price above $ 5 in 2006 dollars; 4) Meaningful 10-K contents ($>$ 200 words)

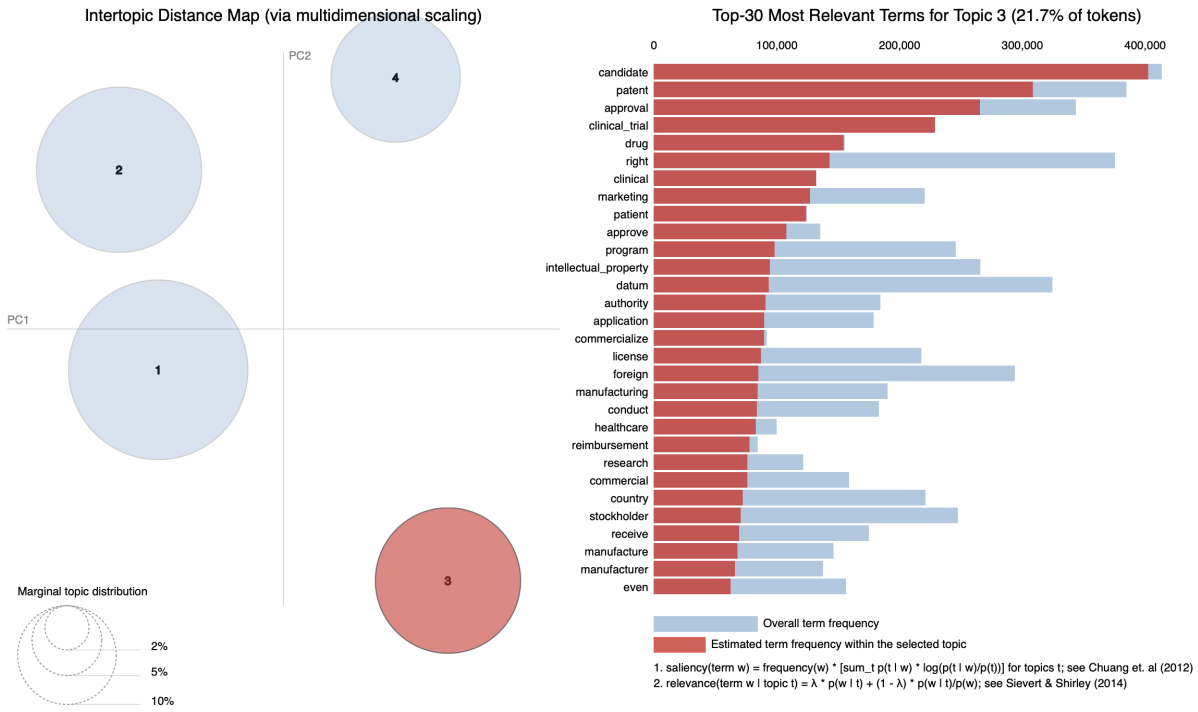


Figure 1: A graphic representation of a four-topic model on firms' risk factors since 2006.

After merging a four-topic map between firms and topic intensities to firms' identifying data, I obtain a topic map as shown in Table ??.

conm	year	CIK	topic_0	topic_1	topic_2	topic_3
BOEING CO	2015	12927	0.875	0.109	0.016	0
UNIFI INC	2022	100726	0.893	0.107	0	0
UTAH MEDICAL PRODUCTS INC	2007	706698	0.021	0.457	0	0.519
SPOK HOLDINGS INC	2010	1289945	0.356	0.643	0	0
APTARGROUP INC	2015	896622	0.791	0.146	0	0.062
OASIS PETROLEUM INC	2018	1486159	1	0	0	0
PROGRESSIVE CORP-OHIO	2011	80661	0.051	0.238	0.711	0
RENTRAK CORP	2009	800458	0.017	0.982	0	0
UNVL STAINLESS ALLOY PRODS	2015	931584	0.957	0.042	0	0
QUALCOMM INC	2015	804328	0	0.998	0	0

Table 2: The table above contains the loadings for a random sample of firm-year occurrences. $Topic_{kk}$ is $Topic_1$ above.

3 Results

In this section, I validate our text-based metric, $Topic_{kk}$, by cross-referencing it with established measures from prior literature. This assessment strengthens the reliability and relevance of our metric within the broader context of knowledge capital analysis.

Finally, I examine the implications of our findings for asset pricing, discerning correlations between knowledge capital and financial performance. This sets the stage for a deeper exploration of knowledge-intensive firms' influence on the financial landscape.

3.1 Defining $Topic_{kk}$

For every topic map, I define $Topic_{kk}$ as the topic that has the highest loading within high-tech sectors in the economy, defined as SIC codes 283, 357, 466, 367, 382, 384, 737 (Brown et al. (2009)).

In the analysis below, topic intensity is defined as the average of topic loadings for all firm-year occurrences, for each case.

Table 3 shows the average topic intensity for low- and high-tech firms for a four-topic model.

Table 3: Topic averages by hi-tech status

hi_tech	topic_0	topic_kk	topic_2	topic_3
0	0.49	0.22	0.27	0.03
1	0.1	0.58	0.02	0.3

3.2 Descriptive statistics

As depicted in Figure 2, the mean topic intensity per year demonstrates a steady rise in the usage of knowledge-capital related language in firms' risk factors since 2009. This upward trend underlines the increasing emphasis placed on knowledge capital within these organizations.

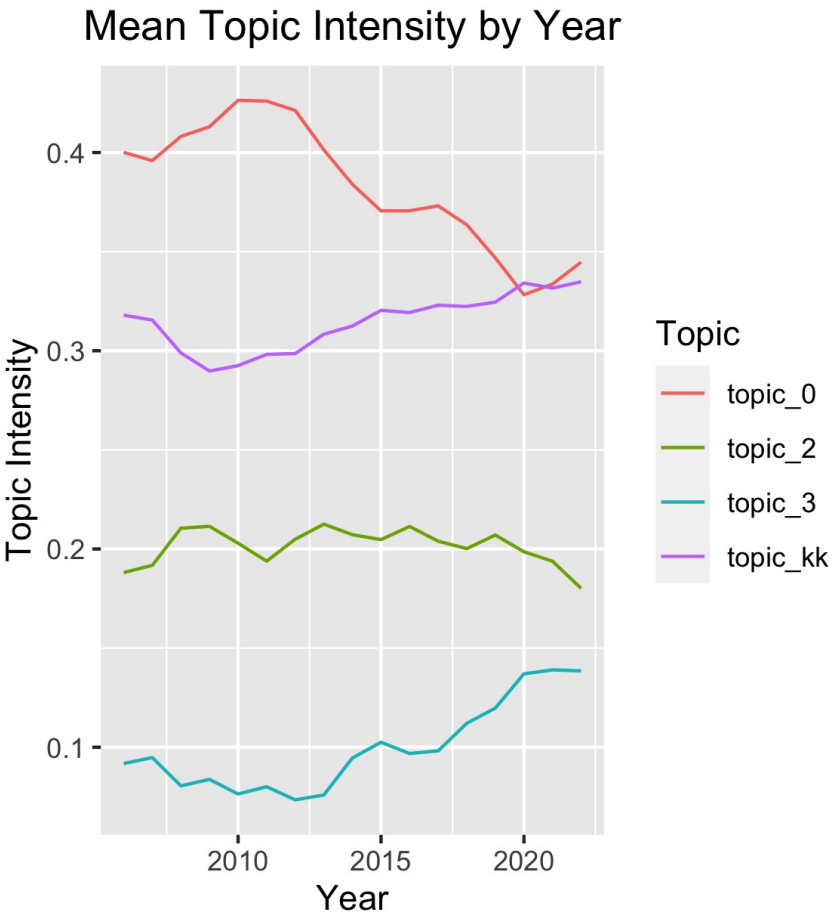


Figure 2: Mean annual topic intensity

Quartiles for $Topic_{kk}$ are shown in Figure 3. As the figure shows, knowledge capital risk is unevenly distributed among firms, with more than half of the firms having $Topic_{kk} < 0.25$.

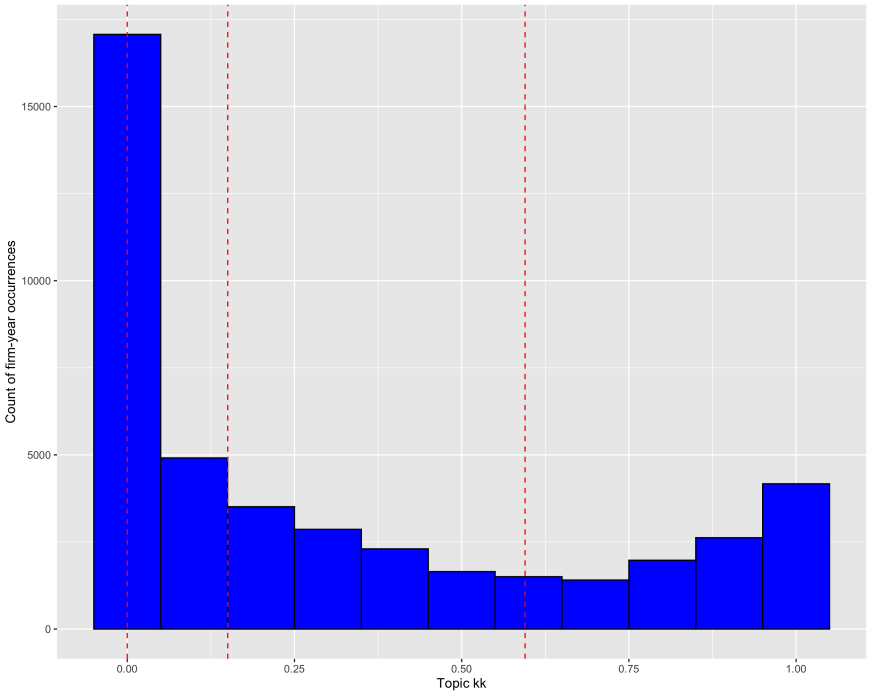


Figure 3: The histogram bars represent the frequency of $Topic_{kk}$ values, while the red dashed lines indicate the quartile dividers.

Lastly, Figure 4 illustrates the accumulated assets of firms in the upper quartile of $Topic_{kk}$, which shows a clear prevalence of firms in the Business Equipment, Chemicals, and Other sectors.

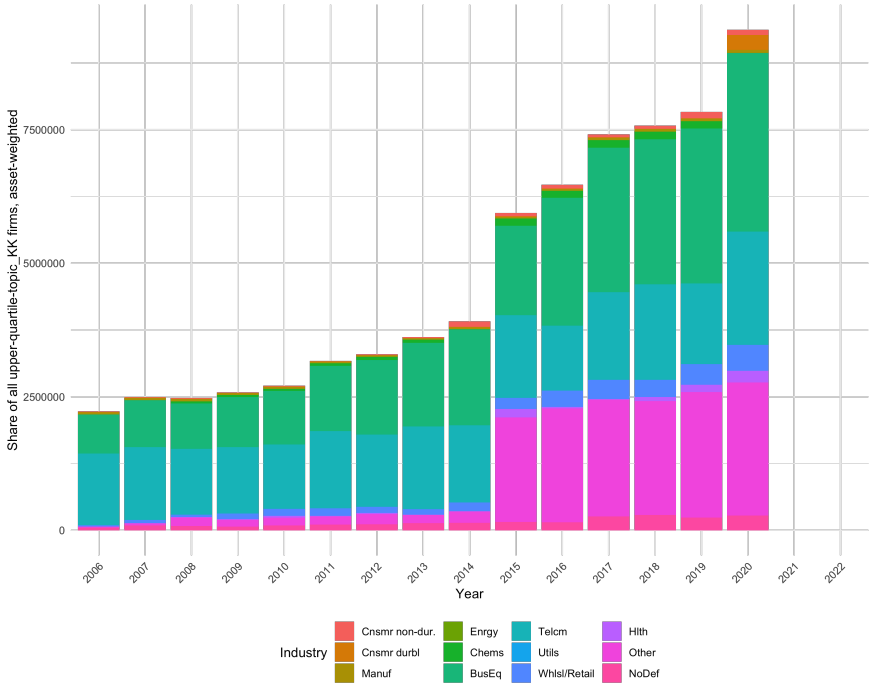


Figure 4: Accumulated Assets of Firms in the Upper Quartile of $Topic_{kk}$

3.3 Validating the text-based metric of knowledge-intensive firms

In this section, the text-based measure $Topic_{kk}$ is cross-examined with other potentially related measures drawn from existing literature.

Figure 5 presents a correlation matrix that delineates the relationship between the intensity of each topic and the average skill level of employees within a narrowly-defined industry, in accordance with the definition given by Belo et al. (2017). These authors define 'Skill' as the proportion of industry workers engaged in occupations that demand a high degree of training and preparation.

The determination of whether an occupation is high-skill is informed by the Specific Vocational Preparation (SVP) level for each occupation. This data is sourced from the 1991 edition of the Dictionary of Occupational Titles (DOT), published by the U.S. Department of Labor.

In Belo et al.'s (2017) classification, an occupation is considered high-skill if it possesses an SVP level of 7 or greater. This level implies a requirement of two or more years of preparation. Occupations failing to meet this threshold are classified as low-skill.

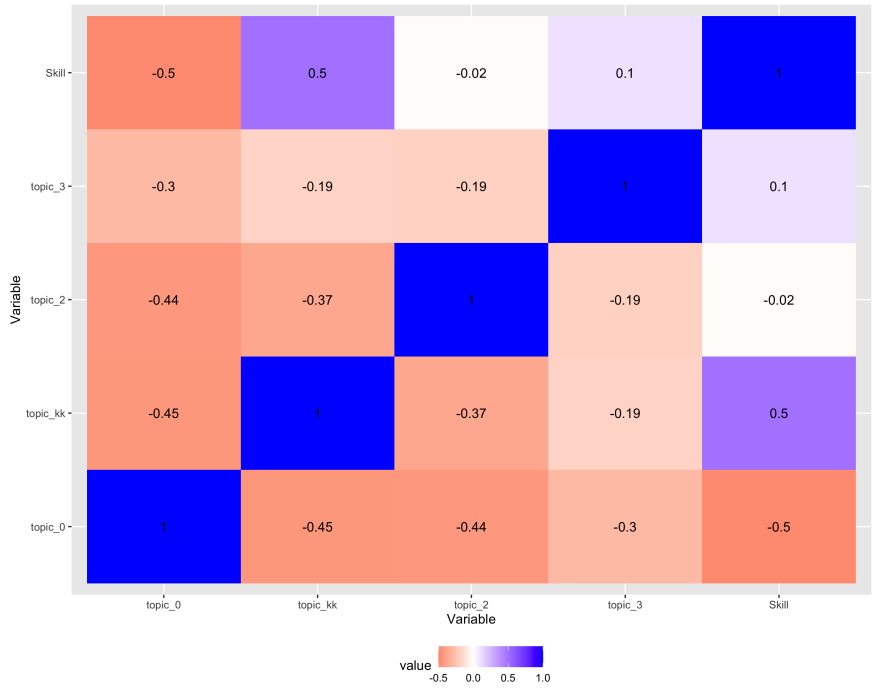


Figure 5: Correlation Matrix Showcasing the Relationship between Topic Intensity and Industry Employee Skill Level

The second component of the analysis, shown in Figure 6, computes the co-occurrences of quartiles for $Topic_{kk}$ and the accumulated patent-related market value within firms, in alignment with the methodology proposed by Kogan et al. (2017). Kogan et al. (2017) leveraged stock market data to estimate the value of individual patents filed by public corporations since 1926.

In Figure 6, the vertical axis signifies annually assigned quartiles of the ratio between Accumulated Patent Value and Total Assets. The results imply a clear correlation between elevated accumulated patent value and increased loadings of $Topic_{kk}$. A basic correlation analysis between average patent intensity and different topics is shown in Figure 7.

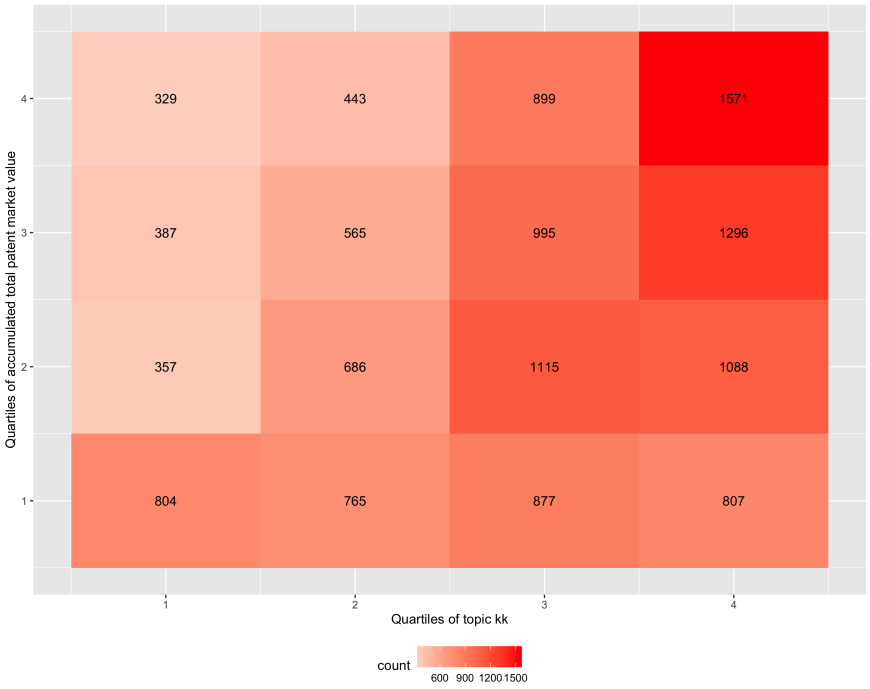


Figure 6: Quartiles of $Topic_{kk}$ vs. quartiles of accumulated total patent market value

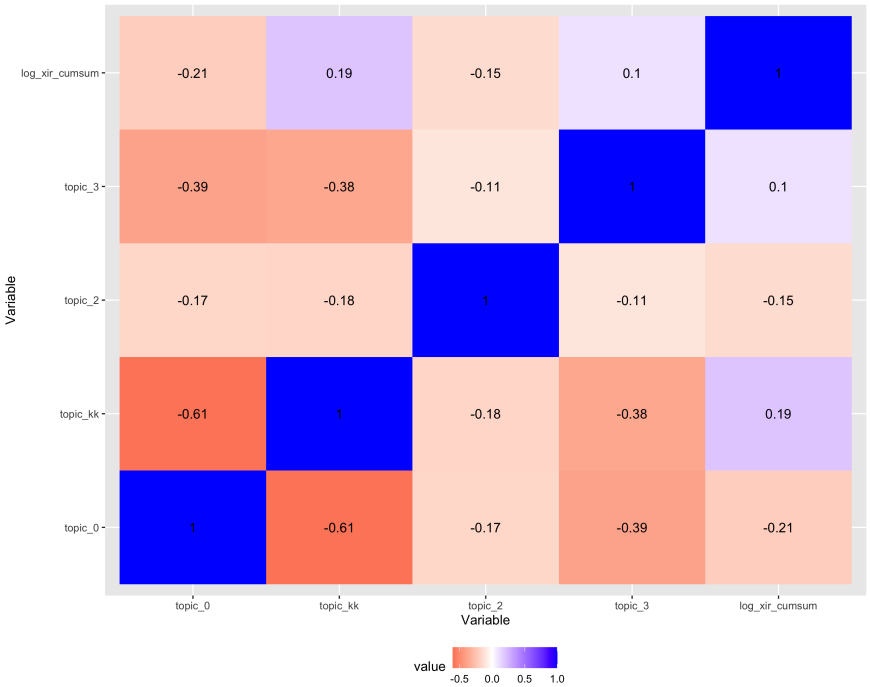


Figure 7: Correlation Matrix: Topic Intensities vs. Firms' Patent Intensities

Figure 8 demonstrates the correlation between $Topic_{kk}$ and knowledge capital, as per the definition provided by Peters and Taylor (2017). Notably, it appears that higher accu-

mulated investments in R&D correspond to increased loadings of $Topic_{kk}$.

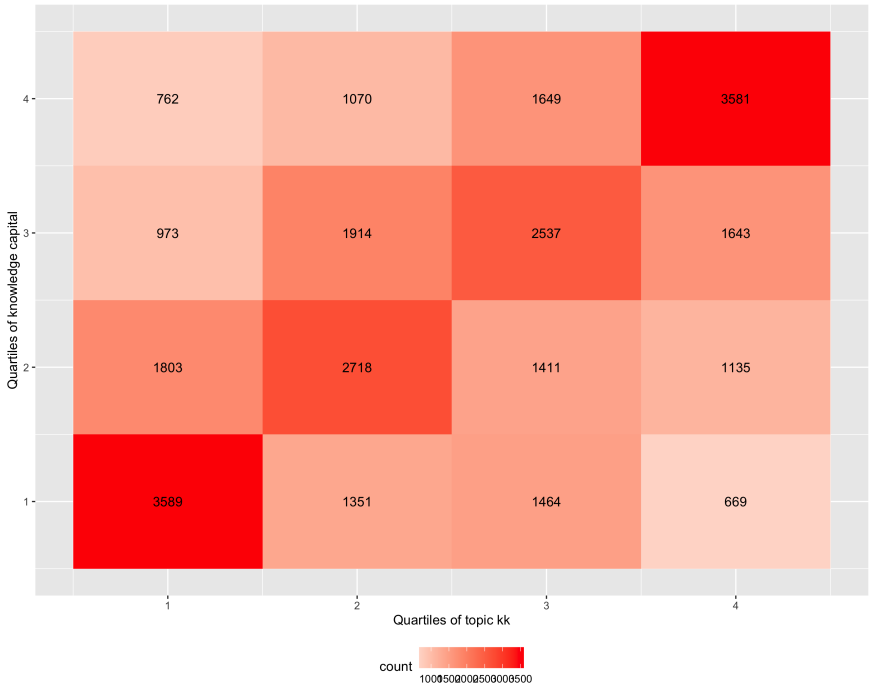


Figure 8: Correlation Between Knowledge Capital Quartiles and $Topic_{kk}$ Quartiles: Higher R&D Investments Correlate with Higher Loadings of $Topic_{kk}$

3.4 Implications for Asset Pricing

In this section, firms are matched based on their Central Index Key (CIK) and Permanent Company Number (PERMNO), facilitating a link between their annual reports and multiple other data sources. These sources encompass daily stock data (aggregated on a weekly basis), Compustat data, and metrics associated with firms' knowledge capital, their cumulative patent value, and the skill level prevalent within their respective industries as delineated in prior literature.

The investigation includes an analysis of value-weighted returns, partitioned according to diverse factors. Figure 9 illustrates value-weighted cumulative weekly returns, sorted by quartiles of $Topic_{kk}$ determined on an annual basis. The data signifies a correlation between enhanced $Topic_{kk}$ loadings and amplified weekly returns, a trend particularly prominent from 2011 onwards.

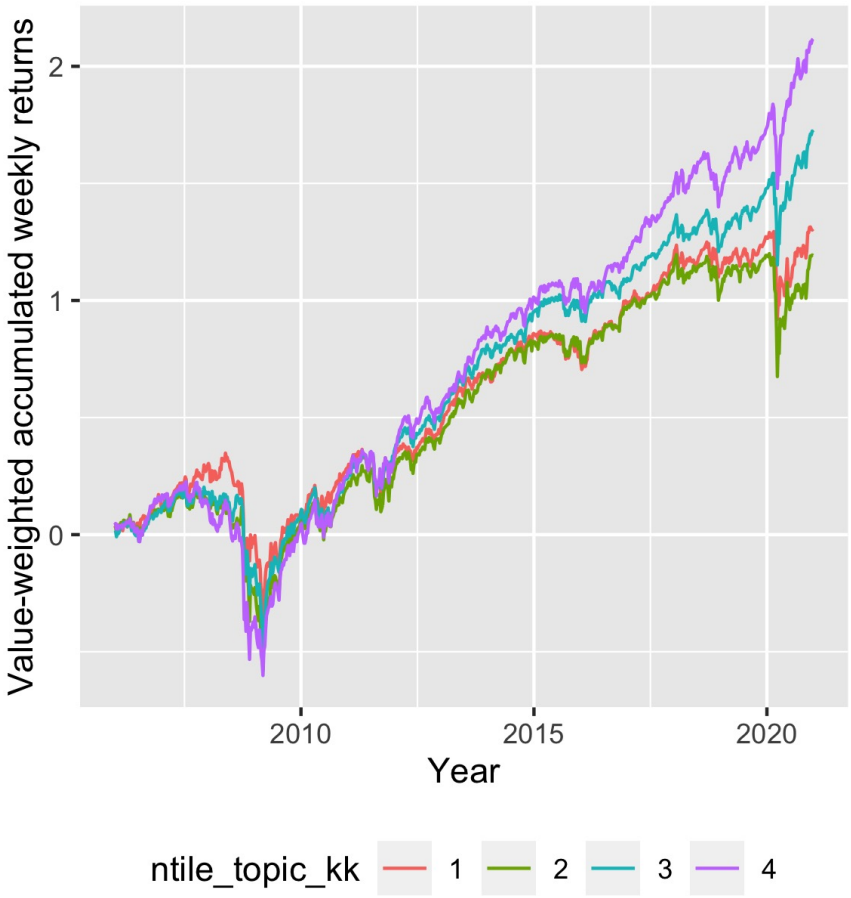


Figure 9: Value-weighted Accumulated Weekly Returns Segregated by Quartiles of $Topic_{kk}$, Showing a Positive Correlation Especially Post-2011

In Figure 10, when constructing $Topic_{kk}$ quartiles for each yearly-industry subset based on the 12-industry Fama-French classification, the resulting patterns become less interpretable.

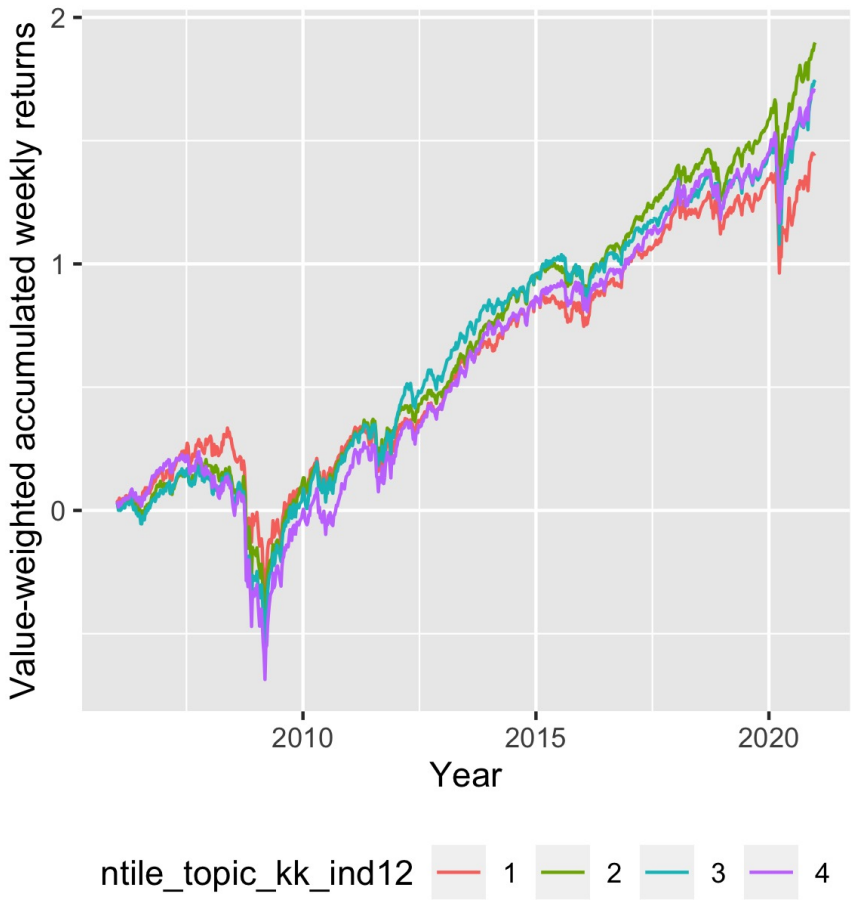


Figure 10: Value-weighted Returns Grouped by Yearly-Industry Subset Quartiles, Indicating Less Interpretable Patterns

Figure 11 displays value-weighted accumulated weekly returns, grouped by firms' maximum topic. Notably, firms with the maximum loading on $Topic_{kk}$ (topic 1) have outperformed their peers since 2008.

Figure 11: Firms with Maximum Loading on $Topic_{kk}$ Outperforming Peers Since 2008

Moving on to volatility patterns, Figure 12 presents the weekly standard deviation of returns within groups, categorized by the maximum topic. This analysis suggests that different topics are associated with varying cross-sectional volatility patterns.

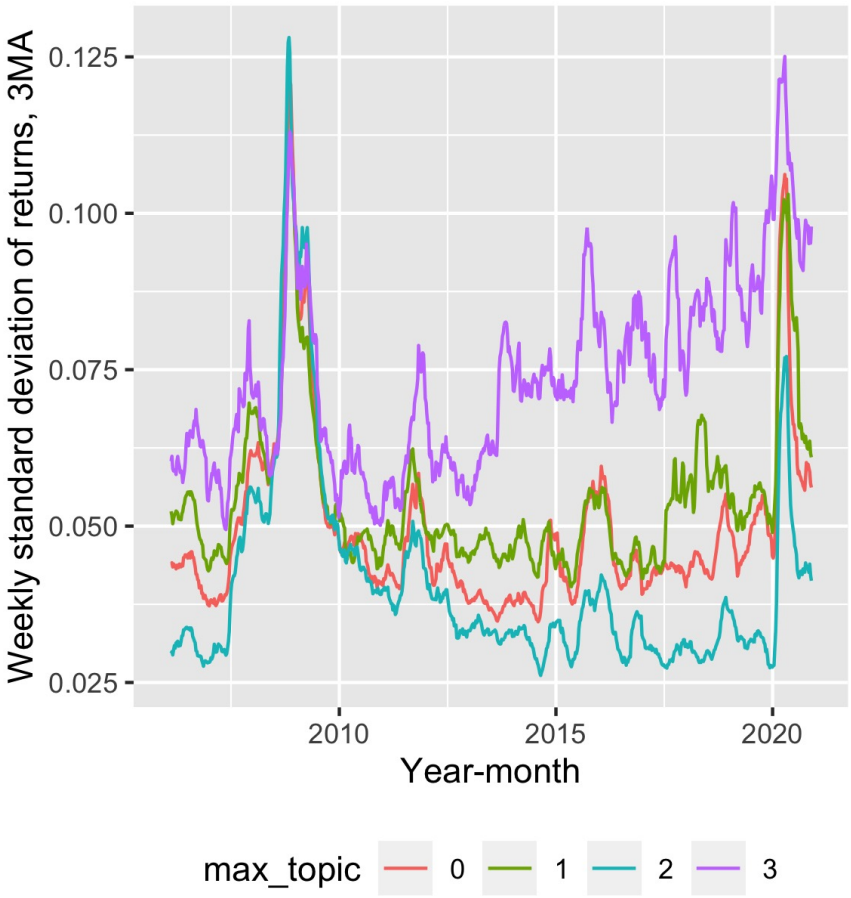


Figure 12: Weekly Standard Deviation of Returns Within Groups, Indicating Variations Across Different Topics

4 Next steps

This study is ongoing, with numerous opportunities for enhancement and further exploration. For instance, the hyperparameter selection process could be optimized, potentially through cross-validation techniques. Existing benchmarks for the choice of k, such as perplexity, are also being considered for testing.

In addition, the current methodology might evolve towards supervised models. For example, a prior for $Topic_{kk}$ could be imposed, comprising specific words such as "patents" or "intellectual property." So far, the learning process has been completely unsupervised.

Another area of interest is testing whether the derived topics can serve as factors in asset pricing models or if they can clarify any anomalies. Moreover, it would be valuable to examine whether the language associated with $Topic_{kk}$ has transformed over time. These areas of focus highlight the promising possibilities for further research and development

in this study.

References

Ai, Hengjie, Jun Li, Kai Li, and Christian Schlag (2019) "The Collateralizability Premium."

Andrei, Daniel, William Mann, and Nathalie Moyen (2019) "Why did the q theory of investment start working?."

Atkeson, Andrew and Ariel Burstein (2019) "Aggregate Implications of Innovation Policy," *J. Polit. Econ.*, 127 (6), 2625–2683.

Belo, Frederico, Vito D Gala, Juliana Salomao, and Maria Ana Vitorino (2019) "Decomposing Firm Value," *J. financ. econ.*, 143 (2), 619–639.

Belo, Frederico, Chen Xue, and Lu Zhang (2013) "A supply approach to valuation," *Rev. Financ. Stud.*, 26 (12), 3029–3067.

Blei, David M, Andrei Ng, and Michael Jordan (2003) "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*

Bloom, N and J Van Reenen (2007) "Measuring and explaining management practices across firms and countries," *Q. J. Econ.*, 122 (4), 1351–1408.

Brown, James R, Steven M Fazzari, and Bruce C Petersen (2009) "Financing innovation and growth: Cash flow, external equity, and the 1990s R&D boom," *J. Finance*, 64 (1), 151–185.

Corrado, Carol, John Haltiwanger, and Daniel Sichel (2009a) *Measuring Capital in the New Economy*: University of Chicago Press.

Corrado, Carol, Charles Hulten, and Daniel Sichel (2009b) "Intangible capital and u.S. Economic growth," *Rev. Income Wealth*, 55 (3), 661–685.

Crouzet, Nicolas and Janice Eberly (2022) "Rents and intangible capital: A Q+ framework," *Journal of Finance*.

Eisfeldt, Andrea L, Antonio Falato, and Mindy Z Xiaolan (2018) "Human Capitalists," April.

Eisfeldt, Andrea L, Edward Kim, and Dimitris Papanikolaou (2020) "Intangible Value,"Technical Report w28056, National Bureau of Economic Research.

Eisfeldt, Andrea L and Dimitris Papanikolaou (2013) "Organization Capital and the Cross-Section of Expected Returns."

Golubov, Andrey and Theodosia Konstantinidi (2019) "Where is the risk in value? Evidence from a market-to-book decomposition," *J. Finance*, 74 (6), 3135–3186.

Grossman, Gene M and Elhanan Helpman (1991) "Quality ladders in the theory of growth," *Rev. Econ. Stud.*, 58 (1), 43.

Hall, Robert E (2001) "The stock market and capital accumulation," *Am. Econ. Rev.*, 91 (5), 1185–1202.

Hansen, Lars Peter, John C Heaton, and Nan Li (2005) "Intangible risk," in *Measuring Capital in the New Economy*, 111–152: University of Chicago Press.

Kogan, L, D Papanikolaou, A Seru, and Noah Stoffman (2017) "Technological innovation, resource allocation, and growth," *The Quarterly Journal*.

Li, Dongmei (2011) "Financial Constraints, R&D Investment, and Stock Returns."

Li, Erica X N, Laura Xiaolei Liu, and Chen Xue (2014) "Intangible Assets and Cross-Sectional Stock Returns: Evidence from Structural Estimation," May.

Li, Wendy C Y and Bronwyn H Hall (2020) "Depreciation of business R&D capital," *Rev. Income Wealth*, 66 (1), 161–180.

Manela, Asaf and Alan Moreira (2017) "News implied volatility and disaster concerns," *J. financ. econ.*, 123 (1), 137–162.

McGrattan, Ellen R and Edward C Prescott (2001) "Is the Stock Market Overvalued?" January.

Mezzanotti, Filippo and Timothy Simcoe (2023) "Innovation and Appropriability: Revisiting the Role of Intellectual Property,"Technical report, National Bureau of Economic Research.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013) "Distributed Representations of Words and Phrases and their Compositionality."

Peri, G, K Shih, and C Sparber (2015) "STEM workers, H-1B visas, and productivity in US cities," *J. Labor Econ.*

Peters, Ryan H and Lucian A Taylor (2017) "Intangible capital and the investment-q relation."

Romer, Paul M (1990) "Endogenous Technological Change," *J. Polit. Econ.*, 98 (5, Part 2), S71–S102.

SEC: Office of Investor Education and Advocacy (2011) "Investor Bulletin: How to Read a 10-K."

Stambaugh, Robert F and Yu Yuan (2016) "Mispricing Factors," *Rev. Financ. Stud.*, 30 (4), 1270–1315.

Vitorino, Maria Ana (2014) "Understanding the Effect of Advertising on Stock Returns and Firm Value: Theory and Evidence from a Structural Model," *Manage. Sci.*, 60 (1), 227–245.

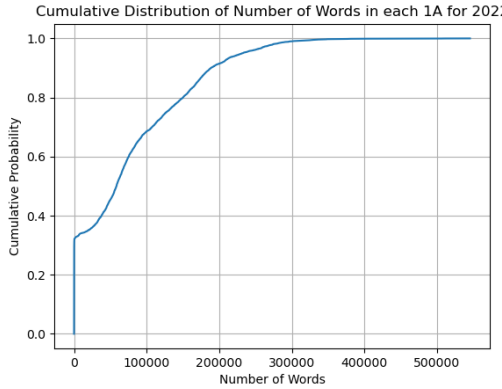
Appendix A: Text conversion to bag-of-words

After filtering the firms, I employ the `spacy` Python library to conduct lemmatization on all the refined texts. Lemmatization transforms words into their base form, ensuring consistent semantics. As an illustration, words such as "take", "took", and "taken" are standardized to "take". To do so, `spacy` leverages WordNet—a comprehensive English lexical database curated by Princeton University.

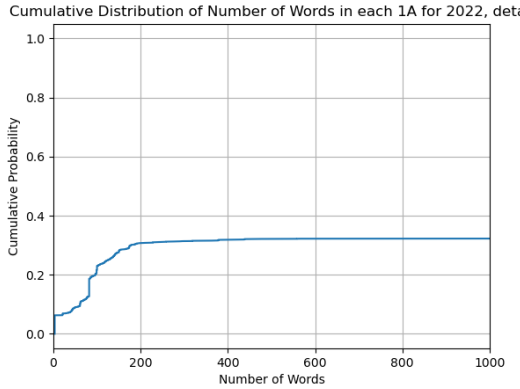
To extract significant collocations—like "patent application"—which provide richer semantic insights than individual words, this research adopts the collocation detection method outlined in Mikolov et al. (2013). This methodology yields pertinent bigrams and trigrams. A minimum occurrence threshold of 5 ensures that only the most statistically relevant combinations are incorporated into the dictionary.

This research compiles the entirety of discovered words, bigrams, and trigrams to formulate a dictionary.

Lastly, the texts are transformed into a bag-of-words model using both the dictionary and the n-gram processed texts. In this representation, each word's frequency in a document, denoted as c_{ij}, is preserved, but the sequence of words is omitted, leading to the final depiction of the corpus.



(a) Cumulative Distribution of Number of Words



(b) Cumulative Distribution of Number of Words, Zoom

Figure 13: Cumulative Distribution of Number of Words in 2022