DOI: 10.1111/acfi.13245

RESEARCH ARTICLE

ACCOUNTING & FINANCE afaanz

Using machine learning and 10-K filings to measure innovation

Essi Nousiainen | **Mikko Ranta** | **Mika Ylinen** | **Marko Järvenpää**

School of Accounting and Finance, University of Vaasa, Vaasa, Finland

Correspondence
Essi Nousiainen, University of Vaasa, P.O. Box 700, 65101 Vaasa, Finland.
Email: essi.nousiainen@uwasa.fi

Funding information
Evald ja Hilda Nissin Säätiö; OP Group Research Foundation, Grant/Award Number: 20200132

Abstract

The purpose of this paper is to develop and validate a text-based measure of innovation using latent Dirichlet allocation on a sample of 45,409 10-K filings from US listed companies. We expect that the text-based innovation measure is associated with innovation and can be used to measure innovation for companies without patents or significant research and development expenditures. The empirical results are consistent with these assumptions, but reveal that thorough initial testing is required to ensure robustness. This study extends the research on innovation measurement and company disclosures, and provides a new method for assessing innovation using company disclosures.

KEYWORDS

10-K, disclosure of innovation, innovation, text analysis, topic modelling

JEL CLASSIFICATION

M40

1 | INTRODUCTION

Corporate innovation is among the most important drivers in boosting long-term growth and the competitiveness of a firm (e.g., Bellstam et al., 2020; Chang et al., 2015; Holmstrom, 1989). Thus, research on various characteristics and determinants of innovation has gained growing interest among accounting and finance scholars (see e.g. Chenhall & Moers, 2015; He & Tian, 2018; Huang et al., 2021). Prior empirical research has used survey instruments (e.g. Bedford et al., 2019; Bisbe & Malagueño, 2009; Henri & Wouters, 2020; Moulang, 2015; Müller-Stewens et al., 2020; Nuhu et al., 2022; Ylinen & Gullkvist, 2014), patent-based proxies (e.g. Bedford et al., 2021; Cai et al., 2021; Grabner et al., 2018; Plečnik et al., 2022; Speckbacher & Wabnegg, 2020; Tang et al., 2021; Zhou & Sadeghi, 2021) and R&D expenditures (e.g. Acharya & Xu, 2017; Helling et al., 2020; Liang, 2022; Zhang et al., 2023) to measure various facets of innovation performance.

However, due to its abstract and multidimensional nature, corporate innovation is considered a challenging object to measure. Scholars have recognised that innovation indicators are context- and innovation type-specific, and especially affected by the industrial sector (Guo et al., 2019). Even though patents and new product development are useful innovation proxies among some industries, it is important to recognise that innovation encompasses more than just the development and introduction of new products and patents. Empirical evidence from Hall et al. (2013) suggests that only a small fraction, approximately 4%, of innovative firms engage in patenting activities. Similarly, Bellstam et al. (2020) found that among their sample of 703 firms from the S&P 500 spanning 1990–2010, 219 firms had no patents, and 329 firms did not engage in any research and development (R&D) activities. Therefore, relying solely on patent-based measures may overlook a wide range of innovation activities undertaken by firms. Moreover, Saidi and Žaldokas (2021) find a trade-off between patenting and trade secrets and Ciftci and Zhou (2016) find that there are more benefits of disclosing patents for industries with strong protection of intellectual property. In sum, patents and R&D expenses are only partial innovation measures (Dziallas & Blind, 2019).

Therefore, researchers are actively seeking alternative measures of innovation that would capture various dimensions of innovation beyond product introductions. To complement these more traditional survey and patent-based measures of innovation performance, there have been recent calls and attempts to use machine learning (ML) methods for creating alternative ways to measure corporate innovation (see Bellstam et al., 2020; Lu & Chesbrough, 2022; Ranta et al., 2023). One stream of research is actively implementing text-based methods for innovation measurement. By adopting a text-based measure, researchers can gain insights into the diverse innovation activities undertaken by firms, providing a more comprehensive assessment of their innovative capabilities.

We contribute to this literature by examining, whether narrative sections of 10-K filings are a suitable source of innovation measurement, with the help of ML. The study conducted by Bellstam et al. (2020) presents compelling findings that support the notion of employing text-based measures to capture a broader scope of innovative achievements beyond the conventional approach that relies on proxies such as patent counts and citations, which predominantly focus on product innovations. By adopting a more comprehensive perspective on measuring innovation, our ML approach offers novel insights and facilitates a more nuanced examination of corporate innovation performance. This approach is particularly valuable as the text-based innovation measure allows for the inclusion of all firms, including those that do not engage in patent creation or invest in R&D activities. The use of natural language processing (NLP) methods ensures that instead of manually rating financial statements for innovation, we can simultaneously generate thousands of innovation ratings based on the same quantitative standards. Using a text-based method for innovation measurement is transparent and replicable, and is not based on human judgement.

This study aims to understand how firms disclose innovative activities in their 10-K filings, and whether their innovativeness can be measured from the text in 10-K filings. Our approach is similar to Bellstam et al. (2020), who use analyst reports and NLP to construct a measure of innovation. It is a text-based measure of firm innovativeness based on latent Dirichlet allocation (LDA). The essence of the method is a way for a computer to 'learn' the topics present in a collection of documents, without knowing in advance what those topics might be. The result is a set of topics, each represented as a collection of words, and a set of documents, each represented as a mixture of these topics. However, we develop the approach of Bellstam et al. (2020) further by experimenting with different model choices and examining how robust these methods are for measuring innovation, and how the model parameters affect the outcomes. We present an alternative model of LDA that is potentially a more reliable measure of innovation when used with annual reports.

The results demonstrate that 10-K filings are suitable for measuring company innovation. Our analysis using a topic model to assess innovation proves adept at predicting a firm's

innovation outcomes based on patents. Additionally, our findings suggest that text-based measures share comparable relationships with performance metrics commonly associated with more conventional innovation proxies like patent counts. Importantly, our method demonstrates effectiveness in measuring innovation among companies that don't rely heavily on patents and in industries with lower patenting activity. Furthermore, our study underscores the significant impact of design choices in ML-based text models on their applicability in measuring innovation. We highlight that the model architecture developed by Bellstam et al. (2020) is not optimally suited for gauging innovation when applied to annual reports. However, by implementing specific modifications to the model, our approach enables its utilisation with annual reports as well.

Our paper contributes to the literature on alternative innovation measures. The availability of patenting data is limited, and not all firms patent or engage in R&D. A definite advantage of innovation measurement from 10-K filings is that they are public documents and are available for all listed companies. This measurement does not depend on patent or R&D data availability. Second, our study makes methodological contributions to NLP methods in accounting and finance research by applying LDA to companies' textual disclosures. We expect our ML approach to shed light on the topical content used in 10-K filings. Our research combines financial disclosures and NLP methods for quantitative measurement and obtaining quantitative information for complex phenomena from narrative sections of 10-K filings. Our text-based measure of innovation is accessible and extends the previous literature by using corporate disclosures to measure firm characteristics. Finally, we shed new light on the robustness of these methods and document how the efficiency of these methods is highly dependent on the source material and architectural choices of the text models.

The rest of the paper is organised as follows. Section 2 discusses the related literature and develops our research question. Section 3 introduces the research methodology and the data and sample. Section 4 discusses the empirical results, and Section 5 presents additional analyses. In the final section, the main findings and implications of the research are summarised and discussed.

2 | LITERATURE AND RESEARCH QUESTION DEVELOPMENT

Studies within business disciplines persistently seek improved methodologies for defining and assessing corporate innovation. Empirical accounting and finance research utilise innovation proxies such as survey tools, patents and R&D spending. Notably, patents and R&D expenditures are extensively employed. However, these proxies have acknowledged limitations, as highlighted in the accounting literature as well (e.g., Huang et al., 2021), potentially lacking comprehensive representation of the intricate phenomenon examined. Beyond the previously highlighted indicators, a further array of metrics comes into play. The work of Dziallas and Blind (2019) unfolds as a comprehensive exploration, culminating in the identification of 82 unique innovation indicators. Their inquiry encompasses diverse innovation dimensions, illustrating its susceptibility to multifaceted analyses. Within their investigation, the authors identify six phases that traverse the innovation process, namely strategy formulation, product definition, product concept development, validation phase, production phase and market launch and commercialisation. This paper demonstrates researchers' active pursuit of methodologies to assess the entirety of these innovation dimensions.

Previous research has produced innovative metrics aimed at achieving a more comprehensive representation of innovation. These include text-based measures of innovation, other novel ways to capture innovation and other types of innovation research that utilise text mining (Antons et al., 2020). Recent examples include: Bellstam et al. (2020), who developed a text-based innovation measure using analyst reports and topic modelling; Kogan et al. (2017), who built an

innovation measure to improve traditional patent count measures by combining patent data with stock market reactions to patent news; Mukherjee et al. (2017), who used the stock market reaction to new product announcements as an alternative innovation measure; and Cooper et al. (2022), who used the firm-specific output elasticity of R&D as an innovation measure.

The objective of this particular study is to develop an innovation measure that captures the degree of innovativeness across all stages of the innovation process by identifying text associated with innovation, rather than focusing exclusively on a specific type of innovation (Dziallas & Blind, 2019). Previous literature has demonstrated that financial report text can provide valuable information and can be used for various research purposes. We aim to extend the literature on using ML methods to analyse annual report text and innovation measurement methods. Based on the assumption that firms will communicate their innovativeness to investors and competitors, and the fact that research findings support more disclosure from innovative companies (Huang et al., 2021) and accelerated patent disclosures in highly competitive product markets (Glaeser & Landsman, 2021), we study whether financial report text can be used to measure innovation with the help of ML.

We are also interested in seeing how sensitive these text-based measures are to the initial design choices of the text model. This research aims to compare different model choices and source materials to analyse how sensitive the identification of innovativeness is to the chosen methodological approach. We evaluate which method performs best and discuss the possible reasons why that specific model is preferred for identifying innovative firms. As previous literature has mostly used proxies that measure only a specific type of innovation, like patenting activity, our goal is to formulate and analyse the suitability of text-based methods for broad identification of innovation that would also identify innovative non-patenting companies.

3 | METHODOLOGY AND DATA

3.1 | Methodology

ML proves to be an invaluable tool for extracting insights from unstructured data, typically difficult to comprehend otherwise. This feature has also attracted the interest of business research, examples of which are articles by Ahmed et al. (2023), Bao et al. (2020), Bei et al. (2021), Bertomeu et al. (2021), Ding et al. (2020), Jones and Alam (2019), as well as the recent work by Ranta and Ylinen (2023a, 2023b). These studies demonstrate how ML is able to isolate useful information from complex data, making it a promising tool to measure abstract concepts, such as innovation, from textual disclosure.

ML for textual analysis, often called natural language processing (NLP), can find patterns in text that would be impossible to detect through manual reading, and the qualitative content in financial statements can be used to garner more nuanced information about the organisation than financial statement metrics (Lewis & Young, 2019). In addition, new ML methods and increases in computing capacity enable more efficient analysis of large data masses. Thus, a significant portion of current research in accounting and finance employs ML methods for the purpose of textual analysis, examples being studies by Belloque et al. (2021), Cai et al. (2019), Clarkson et al. (2020), Garanina et al. (2021), Ylinen and Ranta (2023), Zengul et al. (2021) and Zhu et al. (2017).

A considerable body of prior research has leveraged ML methods in the analysis of 10-K texts, and the volume of new studies in this area is steadily increasing. Notable contributions include works by Basu et al. (2022), Brown et al. (2020), Buehlmaier and Whited (2018), Donovan et al. (2021), Dyer et al. (2017), Frankel et al. (2016), Hoberg and Maksimovic (2015), Kim et al. (2019) and Lehavy et al. (2011). These studies exemplify the diverse applications of ML techniques to extract valuable insights from 10-K documents. 10-K filings contain information

about firms' strategies and products, including innovations, new products and goals. The word choices for discussing these issues could signal innovativeness, for example, extensively discussing new product releases (product innovation) or new business models and strategy development (business process innovation). Thus, we see 10-K filings as a promising source of innovation measurement when combined with ML methods.

We select an unsupervised ML method for building the innovation metrics, since unsupervised methods do not require response variables. We wish to avoid using some other innovation proxy as a response variable for the text-based innovation measure because it could direct the model to measure that innovation proxy or limit the use of the measure to the availability of a specific data variable. More specifically, we use LDA (Blei et al., 2003) to construct our innovation measures. LDA is a probabilistic topic modelling method that presents the predefined number of topics as the probability distributions of words from a predefined vocabulary. The model does not label the topics. The output of the LDA model is the topics' word distributions (percentage of each word in the topic) and the intensity of each topic in each of the documents. The output can then be used to infer topic distribution for any document unseen by the model, and we use this ability to obtain a topic distribution for the innovation textbook and employ the topic distributions and topic intensities in our innovation measure.

We train eight models, with the number of topics at 15, 20, 25, 30, 35, 40, 60 and 80, to test the sensitivity of the method to its initial parameters. Different numbers of topics might affect the results in various ways. A greater number of topics affects the innovation score, resulting in lower values for individual topic intensities. A great number of topics could also result in topics that have no meaning or, conversely, generate meaningful topics that do not appear with a low number of topics. To validate the new innovation indicators, we test them on patent-based innovation indicators and firm performance variables with panel regressions.

Our approach includes two alternative LDA-based innovation measurement methods: the topic weight method and the topic distribution method. The topic weight measure, based on Bellstam et al. (2020), selects an 'innovation topic' from the model using an innovation textbook *Managing innovation* (Tidd et al., 2005). An innovation topic is chosen from all of the topics generated by each LDA model by calculating the Kullback–Leibler (KL) divergence between the word distribution of each topic and the innovation textbook. The innovation topic is then the topic with the lowest KL divergence to the innovation textbook. The final innovation measure in this method is the loading of the innovation topic for each 10-K filing.

For the topic distribution measure, we compare the topic distributions of the innovation textbook and each of the 10-K filings. We calculate the KL divergence between the topic distribution of each 10-K filing and the innovation textbook, which is the final innovation metric with this method. The more the topic distributions differ, the larger the divergence; consequently, more innovative firms should have a lower value. However, to simplify the analysis, we invert the values so that a larger value represents higher innovation.

3.2 | Data and sample

As a textual source of innovation, we use the annual reports of US companies. A definite advantage of innovation measurement from 10-K filings is that they are public documents and are available for all listed companies. Thus, the measurement does not depend on patent or R&D data availability. The same applies for using analyst reports as textual source (Bellstam et al., 2020), as they are usually not freely available. We initially train the LDA models with a sample of 45,409 SEC 10-K filings of public US companies from the years 2008–2018. The texts are pre-processed by lowercasing, removing punctuation, and converting the words into Unicode strings (tokens). Stop words, such as 'and', 'the', 'no', and other common words in English that are not important for the analysis are removed. Finally, before training the model, the number

of words in the dictionary is limited to the most common 100,000 words, excluding those that appear in more than 90% or less than four (individual) of the documents. In addition to 10-K filings, accounting data from the Thomson Reuters Eikon database and patent and citation data from Noah Stoffman's website are used in this study (Kogan et al., 2017). We use a cross-industry dataset to ensure that the innovation metric is applicable regardless of industry. Once we combine the innovation measures with the patenting data and Eikon variables, and remove missing values, the final sample size is 8734 firm-year observations. The final sample may be subject to selection bias, since larger companies are better represented in financial databases.

Descriptive statistics on both innovation measures are presented in Table 1. The topic distribution innovation measure is quite evenly distributed and the medians and means are close to each other across the different models. Since the topic distribution measure is essentially a KL-divergence score, the closer to 0 the innovation score is, the more innovative the firm should be.

TABLE 1 Descriptive statistics and correlation coefficients for the topic distribution and topic weight measures of innovation.

	No. of topics	**15**	**20**	**25**	**30**	**35**	**40**	**60**	**80**
Topic distribution	Descriptive statistics								
	Count	45,078	45,078	45,078	45,078	45,078	45,078	45,078	45,078
	Mean	4.9405	5.3686	5.7573	5.5461	5.7670	6.6942	6.6321	6.6273
	Standard deviation	2.9429	2.3003	2.1206	1.9181	1.8471	2.3111	2.0313	1.3720
	Median	4.1633	5.3233	5.7924	5.6310	5.8241	6.6351	6.6745	6.7253
	Correlations								
	15	1							
	20	0.6988	1						
	25	0.6999	0.8704	1					
	30	0.7186	0.8809	0.8715	1				
	35	0.7257	0.8292	0.8326	0.8739	1			
	40	0.8082	0.7344	0.7628	0.7720	0.7826	1		
	60	0.6667	0.8088	0.8125	0.7998	0.7871	0.7323	1	
	80	0.6717	0.7964	0.8073	0.8299	0.8385	0.7479	0.8236	1
Topic weight	Descriptive statistics								
	Count	45,078	45,078	45,078	45,078	45,078	45,078	45,078	45,078
	Mean	0.0449	0.0848	0.0647	0.0730	0.0547	0.0115	0.0530	0.0314
	Standard deviation	0.1095	0.1927	0.1708	0.1760	0.1450	0.0851	0.1384	0.1175
	Median	0.0046	0.0001	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	Correlations								
	15	1							
	20	−0.0821	1						
	25	−0.0740	0.9211	1					
	30	−0.0825	0.9580	0.9285	1				
	35	−0.0752	0.7943	0.8332	0.8358	1			
	40	0.1213	−0.0495	−0.0427	−0.0477	−0.0429	1		
	60	−0.0628	0.1048	−0.0372	0.0543	−0.0854	−0.0374	1	
	80	−0.0578	0.8238	0.8622	0.8579	0.7969	−0.0322	−0.0540	1

TABLE 2 Descriptive statistics and correlation coefficients for financial variables.

Variable	log (*Patents*)	log (*Citations*)	log (*Sales*)	log (*Assets*)	log (*Age*)	*R&D/Sales*	*Total debt/ Total capital*	*ROA*	log (*Q*)	*Beta*
Descriptive statistics										
Count	33,641	30,641	39,829	39,888	33,958	17,154	39,155	39,157	35,880	33,239
Mean	0.6239	0.7446	11.8892	13.0604	9.2090	837.91	38.5792	−2480.70	0.9889	1.9302
Standard deviation	1.3296	1.7018	3.7615	2.9064	0.6130	16,264.53	1457.41	212,590.69	0.4770	35.0934
Median	0.0000	0.0000	12.5688	13.4061	9.1883	5.5100	31.1000	2.4300	0.8545	1.2200
Correlations										
log (*Patents*)	1									
log (*Citations*)	0.8884	1								
log (*Sales*)	0.2473	0.1999	1							
log (*Assets*)	0.2288	0.1760	0.8120	1						
log (*Age*)	0.1651	0.1347	0.1984	0.1181	1					
R&D/Sales	−0.0047	0.0069	−0.1480	−0.0375	−0.0310	1				
Total debt/Total capital	−0.0047	−0.0052	0.0025	0.0091	0.0025	0.0013	1			
ROA	0.0053	0.0037	0.0368	0.0521	−0.0041	−0.0500	0.0022	1		
log (*Q*)	0.1833	0.1655	−0.1639	−0.2466	−0.0613	0.0318	−0.0751	0.1351	1	
Beta	−0.0034	−0.0074	−0.0147	−0.0233	−0.0076	0.0054	−0.0002	−0.0006	0.0242	1

Table 1 also presents the correlation coefficients between the topic-number distinguished models of the topic distribution measure. As can be seen from the table, the topic distribution measure is quite robust to changes in the predefined number of topics. The correlation coefficients stand mainly between 0.7 and 0.9. The topic weight measure ranges between 0 and 1, depending on the intensity of the 'innovation topic' in the specific 10-K filing. The mean and median, reported in the lower panel of Table 1, are generally quite low and close to 0 with the median value being 0 for most of the topic-number distinguished models. There is also high variability in the correlation coefficients with the topic weight innovation measure, ranging from negative and close to zero coefficients to the highest correlation of 0.9285. This finding does not support the consistency of the topic weight measure, and suggests that thorough testing is needed when using this approach for measuring innovation from annual reports. We give more insight into this finding in the next section.

The descriptive statistics for the patent variables and financial variables, as well as the correlation coefficients, are reported in Table 2.

4 | EMPIRICAL RESULTS

4.1 | Model comparison

4.1.1 | Patent-based innovation

We proceed by studying the association between LDA-based innovation measurement methods and patent-based innovation. We analyse two different models described in the previous section: the topic weight method and the topic distribution method. Furthermore, we evaluate

TABLE 3 The performance of the topic distribution measure of innovation on patent count and citation count.

No. of topics	15		20		25		30	
Dependent variable	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**
Explanatory variable								
Text innovation (topic dist.)	0.0274*** (0.0087)	0.0520*** (0.0118)	0.0362*** (0.0090)	0.0482*** (0.0125)	0.0368*** (0.0094)	0.0537*** (0.0130)	0.0500*** (0.0112)	0.0646*** (0.0156)
ROA	−0.0092*** (0.0006)	−0.0105*** (0.0008)	−0.0034*** (0.0003)	−0.0037*** (0.0004)	−0.0034*** (0.0003)	−0.0037*** (0.0004)	−0.0034*** (0.0003)	−0.0037*** (0.0004)
R&D/Sales	0.00002 (0.00004)	0.00003 (0.00005)	0.00001 (0.00002)	0.00002 (0.00003)	0.00001 (0.00002)	0.00002 (0.00003)	0.00001 (0.00002)	0.00002 (0.00003)
log (*Sales*)	0.0193 (0.0211)	0.0163 (0.0300)	−0.0368* (0.0197)	−0.0368 (0.0281)	−0.0365* (0.0196)	−0.0376 (0.0279)	−0.0368* (0.0197)	−0.0364 (0.0280)
log (*Assets*)	0.4874*** (0.0209)	0.5865*** (0.0294)	0.4873*** (0.0205)	0.5714*** (0.0292)	0.4852*** (0.0196)	0.5700*** (0.0289)	0.4862*** (0.02094)	0.5696*** (0.0290)
log (*Age*)	0.2999*** (0.0319)	0.2061*** (0.0414)	0.2651*** (0.0311)	0.1692*** (0.0406)	0.2668*** (0.0311)	0.1713*** (0.0406)	0.2616*** (0.0311)	0.1648*** (0.0406)
Total debt/Total capital	−0.0023*** (0.0005)	−0.0031*** (0.0007)	−0.0012*** (0.0004)	−0.0017*** (0.0005)	−0.0012*** (0.0004)	−0.0017*** (0.0005)	−0.0012*** (0.0004)	−0.0016*** (0.0005)
Beta	−0.0675*** (0.0185)	−0.0668** (0.0265)	−0.0434** (0.0178)	−0.0448* (0.0254)	−0.0438** (0.0178)	−0.0449* (0.0253)	−0.0424** (0.0178)	−0.0436* (0.0253)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	8708	8708	8708	8708	8708	8708	8708	8708
R^2	0.3907	0.3078	0.3690	0.2876	0.3690	0.2878	0.3693	0.2878

Note: The table contains panel regressions for logged patents and citations on the topic distribution measure of innovation. The columns present the topic count of the base-LDA-model for *Text innovation*. Other controls include firm-year observations for *ROA*, *R&D/Sales*, log (*Sales*), log (*Assets*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

how robust the text-based innovation metrics are by varying the number of topics in the LDA models. To assess the association between the text-based innovation measurement and patent-based innovation, we specify the following two baseline regression models:

$$\log(Patents)_{it} = \alpha + \beta_1 Text\_inn_{it} + \beta_2 ROA_{it} + \beta_3 R\&D/Sales_{it} + \beta_4 \log(Sales)_{it} + \beta_5 \log(Assets)_{it} + \beta_6 \log(Age)_{it} + \beta_7 \frac{Total\ debt}{Total\ capital}_{it} + \beta_8 Beta_{it} + \mu_j + \gamma_t + \varepsilon_{it}, \quad (1)$$

$$\log(Citations)_{it} = \alpha + \beta_1 Text\_inn_{it} + \beta_2 ROA_{it} + \beta_3 R\&D/Sales_{it} + \beta_4 \log(Sales)_{it} + \beta_5 \log(Assets)_{it} + \beta_6 \log(Age)_{it} + \beta_7 \frac{Total\ debt}{Total\ capital}_{it} + \beta_8 Beta_{it} + \mu_j + \gamma_t + \varepsilon_{it}, \quad (2)$$

where the variable definitions can be found in the Appendix (Tables A1 and A2). The explanatory variable in Equation (1) is the logged number of patents for firm *i* in year *t*. The explanatory variable in Equation (2) is the logged number of patent citations for firm *i* in year *t*. The variable of interest in both is *Text_inn*, which represents the firm-year observation of each text-based innovation measure. Included control variables are firm age, total assets, return on assets, R&D intensity, beta and net sales. A more detailed description of the control variables is presented in Appendix (Tables A1 and A2). The μ_j term represents the SIC4 industry fixed effects and γ_t the year fixed effects.

We start with the topic distribution model and the results for Equations (1) and (2) are provided in Table 3. We estimate in total eight models by varying the predefined number of topics. As we can see from the results regarding both models, the topic distribution method is strongly associated with patenting. The coefficients of text innovation are consistently statistically significant and positive.

35		**40**		**60**		**80**	
log (*Patents*)	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**
0.0927***	0.1305***	0.0696***	0.0930***	0.0671***	0.0773***	0.0820***	0.1192***
(0.0122)	(0.0160)	(0.0095)	(0.0131)	(0.0098)	(0.0128)	(0.0150)	(0.0197)
−0.0035***	−0.0038***	−0.0111***	−0.0131***	−0.0034***	−0.0037***	−0.0035***	−0.0037***
(0.0003)	(0.0003)	(0.0007)	(0.0010)	(0.0003)	(0.0003)	(0.0003)	(0.0003)
0.000004	0.00002	0.00001	0.000004	0.000003	0.00002	0.00001	0.00002
(0.00002)	(0.00003)	(0.00004)	(0.00005)	(0.00002)	(0.00003)	(0.00002)	(0.00003)
−0.0438**	−0.0473*	0.0080	0.0025	−0.0445**	−0.0439	−0.0382*	−0.0400
(0.0198)	(0.0274)	(0.0210)	(0.0306)	(0.0197)	(0.0275)	(0.0197)	(0.0275)
0.4993***	0.5893***	0.5020***	0.6073***	0.4942***	0.5772***	0.4877***	0.5736***
(0.0208)	(0.0278)	(0.0206)	(0.0296)	(0.0204)	(0.0277)	(0.0204)	(0.0277)
0.2279***	0.1589***	0.3006***	0.2127***	0.2641***	0.1686***	0.2602***	0.1617***
(0.0310)	(0.0377)	(0.0313)	(0.0408)	(0.0311)	(0.0377)	(0.0311)	(0.0377)
−0.0011***	−0.0015**	−0.0039***	−0.0054***	−0.0012***	−0.0017***	−0.0012***	−0.0016***
(0.0004)	(0.0005)	(0.0006)	(0.0008)	(0.0004)	(0.0005)	(0.0004)	(0.0005)
−0.0427**	−0.0434*	−0.0664***	−0.0748***	−0.0439**	−0.0460*	−0.0420**	−0.0422
(0.0178)	(0.0261)	(0.0185)	(0.0266)	(0.018)	(0.0261)	(0.0178)	(0.0261)
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
8708	8708	8708	8708	8708	8708	8708	8708
0.3726	0.2919	0.3938	0.3107	0.3717	0.2893	0.3702	0.2894

TABLE 4 The performance of the topic weight measure of innovation on patent count and citation count.

No. of topics	15		20		25		30	
Dependent variable	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**
Explanatory variable								
Text innovation (topic weight)	−0.4340*** (0.1210)	−0.5757*** (0.1685)	1.4248*** (0.0883)	1.9937*** (0.1233)	1.3307*** (0.0983)	1.8003*** (0.1358)	1.3748*** (0.0999)	1.8823*** (0.1394)
ROA	−0.0033*** (0.0003)	−0.0103*** (0.0008)	−0.0036*** (0.0003)	−0.0040*** (0.0004)	−0.0036*** (0.0003)	−0.0039*** (0.0004)	−0.0036*** (0.0003)	−0.0039*** (0.0004)
R&D/Sales	0.00001 (0.00002)	0.00001 (0.00005)	0.00001 (0.00002)	0.00002 (0.00003)	0.00001 (0.00002)	0.00003 (0.00003)	0.00001 (0.00002)	0.00003 (0.00003)
log (*Sales*)	0.0266 (0.0194)	0.0259 (0.0298)	−0.0387** (0.0191)	−0.0401 (0.0271)	−0.0324* (0.0193)	−0.0310 (0.0275)	−0.0328* (0.0191)	−0.0316 (0.0272)
log (*Assets*)	0.4715*** (0.0200)	0.5703*** (0.0290)	0.4902*** (0.0200)	0.5764*** (0.0282)	0.4853*** (0.0202)	0.5690*** (0.0286)	0.4874*** (0.0200)	0.5720*** (0.0282)
log (*Age*)	0.2560*** (0.0311)	0.1804*** (0.0412)	0.2474*** (0.0313)	0.1443*** (0.0408)	0.2572*** (0.0314)	0.1584*** (0.0409)	0.2483*** (0.0314)	0.1461*** (0.0409)
Total debt/Total capital	−0.0010*** (0.0004)	−0.0031*** (0.0007)	−0.0005*** (0.0004)	−0.0008*** (0.0005)	−0.0008* (0.0004)	−0.0011** (0.0005)	−0.0007* (0.0004)	−0.0010*** (0.0005)
Beta	−0.0435** (0.0173)	−0.0689*** (0.0264)	−0.0559*** (0.0175)	−0.0619** (0.0249)	−0.0582*** (0.0177)	−0.0647*** (0.0252)	−0.0536*** (0.0175)	−0.0586*** (0.0250)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	8734	8734	8708	8708	8708	8708	8708	8708
R^2	0.3699	0.3069	0.3915	0.3132	0.3849	0.3045	0.3865	0.3066

Note: The table contains panel regressions for logged patents and citations on the topic weight measure of innovation. The columns present the topic count of the base-LDA-model for *Text innovation*. Other controls include firm-year observations for *ROA*, *R&D/Sales*, log (*Sales*), log (*Assets*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

TABLE 5 The performance of the topic distribution measure of innovation on patent count and citation count in the patenting firm subsample.

No. of topics	15		20		25		30	
Dependent variable	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**
Explanatory variable								
Text innovation (topic dist.)	0.0137 (0.0101)	0.0449*** (0.0148)	0.0327*** (0.0114)	0.0469*** (0.0166)	0.0290*** (0.0110)	0.0547*** (0.0160)	0.0209 (0.0141)	0.0286 (0.0204)
ROA	−0.0077*** (0.0008)	−0.0086*** (0.0012)	−0.0077*** (0.0008)	−0.0084*** (0.0012)	−0.0077*** (0.0008)	−0.0084*** (0.0012)	−0.0077*** (0.0008)	−0.0084*** (0.0012)
R&D/Sales	0.00005* (0.00003)	0.00005 (0.00004)	0.00005* (0.00003)	0.00005 (0.00004)	0.00005* (0.00003)	0.00005 (0.00004)	0.00005* (0.00003)	0.00005 (0.00004)
log (*Sales*)	0.0746** (0.0290)	0.0702 (0.0445)	0.0695*** (0.0291)	0.0702 (0.0446)	0.0698** (0.0291)	0.0667 (0.0443)	0.0741** (0.0291)	0.0771* (0.0445)
log (*Assets*)	0.4742*** (0.0282)	0.5463*** (0.0445)	0.4774*** (0.0280)	0.5408*** (0.0442)	0.4765*** (0.0280)	0.5432*** (0.0440)	0.4728*** (0.0280)	0.5340*** (0.0441)
log (*Age*)	−0.0105 (0.0390)	−0.1245** (0.0561)	−0.0174 (0.0385)	−0.1445*** (0.0557)	−0.0154 (0.0386)	−0.1415** (0.0558)	−0.0180 (0.0387)	−0.1453*** (0.0558)
Total debt/Total capital	−0.0004 (0.0005)	−0.0006 (0.0007)	−0.0003 (0.0005)	−0.0006 (0.0007)	−0.0003 (0.0005)	−0.0005 (0.0007)	−0.0003 (0.0005)	−0.0006 (0.0007)
Beta	−0.0026 (0.0305)	0.0005 (0.0471)	−0.0004 (0.0303)	−0.0039 (0.0467)	−0.0023 (0.0302)	−0.0047 (0.0465)	−0.0031 (0.0302)	−0.0080 (0.0467)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	4401	4401	4401	4401	4401	4401	4401	4401
R^2	0.4876	0.3327	0.4885	0.3323	0.4883	0.3329	0.4876	0.3313

Note: The table contains panel regressions for logged patents and citations on the topic distribution measure of innovation for the patenting firm subsample. The columns present the topic count of the base-LDA-model for *Text innovation*. Other controls include firm-year observations for *ROA*, *R&D/Sales*, log (*Sales*), log (*Assets*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust.
*, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

35		40		60		80	
log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)
1.1724***	1.6782***	−3.9744**	−4.9659***	−0.3003**	−0.2008	1.5021***	1.9767***
(0.1076)	(0.1483)	(1.6061)	(1.8801)	(0.1378)	(0.1969)	(0.1439)	(0.1955)
−0.0037***	−0.0040***	−0.0033***	−0.0035***	−0.0035***	−0.0037***	−0.0035***	−0.0038***
(0.0003)	(0.0005)	(0.0003)	(0.0004)	(0.0003)	(0.0004)	(0.0003)	(0.0004)
0.00001	0.00003	0.00001	0.00002	0.00001	0.00002	0.00001	0.00002
(0.00002)	(0.00003)	(0.00002)	(0.00003)	(0.00002)	(0.00003)	(0.00002)	(0.00003)
−0.0272	−0.0240	−0.0269	−0.0253	−0.0257	−0.0232	−0.0318*	−0.0301
(0.0195)	(0.0276)	(0.0194)	(0.0274)	(0.0195)	(0.0276)	(0.0192)	(0.0272)
0.4841***	0.5680***	0.4718***	0.5454***	0.4743***	0.5547***	0.4808***	0.5627***
(0.0203)	(0.0287)	(0.0201)	(0.0284)	(0.0202)	(0.0286)	(0.0200)	(0.0284)
0.2543***	0.1535***	0.2612***	0.1619***	0.2661***	0.1718***	0.2486***	0.1475***
(0.0312)	(0.0406)	(0.0311)	(0.0400)	(0.0312)	(0.0407)	(0.0314)	(0.0409)
−0.0009***	−0.0012***	−0.0010***	−0.0014***	−0.0012***	−0.0017***	−0.0010***	−0.0014***
(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0004)	(0.0005)
−0.0607***	−0.0691***	−0.0438**	−0.0417*	−0.0492***	−0.0511***	−0.0559***	−0.0613***
(0.0178)	(0.0253)	(0.0173)	(0.0247)	(0.0177)	(0.0253)	(0.0177)	(0.0252)
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
8708	8708	8734	8734	8708	8708	8708	8708
0.3782	0.2988	0.3697	0.2861	0.3681	0.2864	0.3788	0.2974

35		40		60		80	
log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)
0.0586***	0.1029***	0.0179	0.0519***	0.0341***	0.0407**	0.0582***	0.0973***
(0.0151)	(0.0224)	(0.0114)	(0.0169)	(0.0114)	(0.0171)	(0.0179)	(0.0273)
−0.0078***	−0.0086***	−0.0107***	−0.0118***	−0.0077***	−0.0083***	−0.0077***	−0.0085***
(0.0008)	(0.0012)	(0.0010)	(0.0017)	(0.0008)	(0.0012)	(0.0008)	(0.0012)
0.00004**	0.00004	0.0001***	0.0001	0.00004*	0.00005	0.00005*	0.00004
(0.00003)	(0.00004)	(0.00005)	(0.0001)	(0.00002)	(0.00004)	(0.00003)	(0.00004)
0.0644**	0.0583	0.1134***	0.0864*	0.0691**	0.0719	0.0679***	0.0653
(0.0295)	(0.0451)	(0.0289)	(0.0477)	(0.0291)	(0.0447)	(0.0294)	(0.0452)
0.4845***	0.5561***	0.4617***	0.5606***	0.4774***	0.5388***	0.4789***	0.5454***
(0.0287)	(0.0450)	(0.0272)	(0.0453)	(0.0280)	(0.0442)	(0.0282)	(0.0446)
−0.0202	−0.1500***	−0.0071	−0.1219**	−0.0146	−0.1408**	−0.0169	−0.1441***
(0.0385)	(0.0556)	(0.0391)	(0.0562)	(0.0386)	(0.0558)	(0.0386)	(0.0558)
−0.0003	−0.0004	−0.0015	−0.0024**	−0.0003	−0.0006	−0.0003	−0.0005
(0.0005)	(0.0007)	(0.0007)	(0.0010)	(0.0005)	(0.0007)	(0.0005)	(0.0007)
−0.0053	−0.0105	−0.0035	−0.0085	−0.0020	−0.0072	0.0023	0.0022
(0.0302)	(0.0467)	(0.0302)	(0.0471)	(0.0301)	(0.0466)	(0.0300)	(0.0464)
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
4401	4401	4401	4401	4401	4401	4401	4401
0.4895	0.3348	0.4965	0.3399	0.4885	0.3319	0.4888	0.3333

TABLE 6 The performance of the topic weight measure of innovation on patent count and citation count in the patenting firm subsample.

No. of topics	15		20		25		30	
Dependent variable	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**	**log (*Patents*)**	**log (*Citations*)**
Explanatory variable								
Text innovation (topic weight)	−0.5728*** (0.1459)	−0.7660*** (0.2135)	0.9202*** (0.1005)	1.2617*** (0.1495)	0.7943*** (0.1030)	1.0295*** (0.1551)	0.9175*** (0.1032)	1.2452*** (0.1560)
ROA	−0.0077*** (0.0008)	−0.0082*** (0.0012)	−0.0080*** (0.0008)	−0.0088*** (0.0012)	−0.0080*** (0.0008)	−0.0087*** (0.0012)	−0.0079*** (0.0008)	−0.0087*** (0.0012)
R&D/Sales	0.00006** (0.00003)	0.00005 (0.00004)	0.00004 (0.00003)	0.00004 (0.00004)	0.00005* (0.00003)	0.00004 (0.00004)	0.00004* (0.00003)	0.00004 (0.00004)
log (*Sales*)	0.0899*** (0.0285)	0.0917** (0.0432)	0.0600** (0.0289)	0.0578 (0.0441)	0.0680** (0.0292)	0.0695 (0.0443)	0.0643** (0.0288)	0.0638 (0.0440)
log (*Assets*)	0.4560*** (0.0275)	0.5060*** (0.0428)	0.4983*** (0.0282)	0.5690*** (0.0441)	0.4883*** (0.0285)	0.5538*** (0.0444)	0.4936*** (0.0279)	0.5622*** (0.0438)
log (*Age*)	−0.0303 (0.0383)	−0.1594*** (0.0547)	−0.0319 (0.0387)	−0.1644*** (0.0563)	−0.0297 (0.0390)	−0.1602*** (0.0564)	−0.0327 (0.0389)	−0.1652*** (0.0564)
Total debt/Total capital	−0.0003 (0.0005)	−0.0004 (0.0007)	−0.00004 (0.0005)	−0.00003 (0.0007)	−0.00006 (0.0005)	−0.0002 (0.0007)	−0.000001 (0.0005)	−0.00009 (0.0007)
Beta	−0.0031 (0.0298)	−0.0027 (0.00461)	−0.0073 (0.0296)	−0.0137 (0.0459)	−0.0155 (0.0301)	−0.0242 (0.0465)	−0.0062 (0.0297)	−0.0122 (0.0460)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	4408	4408	4401	4401	4401	4401	4401	4401
R^2	0.4901	0.3292	0.5006	0.3450	0.4965	0.3396	0.4995	0.3436

Note: The table contains panel regressions for logged patents and citations on the topic weight measure of innovation for the patenting firm subsample. The columns present the topic count of the base-LDA-model for *Text innovation*. Other controls include firm-year observations for *ROA*, *R&D/Sales*, log (*Sales*), log (*Assets*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

Next, we estimate both equations for the topic weight innovation measure (Bellstam et al., 2020). The regression results can be found in Table 4. As for the topic distribution measure, we estimate eight models in total by varying the predefined number of topics. The topic weight method performs less consistently compared to the topic distribution method. The results include cases where the coefficient is either positive or negative and statistically significant, demonstrating how the topic weight model is very sensitive to the initial specifications, like the number of topics.

The topic distribution method is, according to this analysis, relatively robust at predicting patent-based innovation. However, the results of the topic weight method reveal the need for careful initial testing, when implementing sophisticated text-based methods for measuring abstract concepts, such as innovation. Bellstam et al. (2020) demonstrated relatively robust results for the topic weight model when used with analyst reports, but our results indicate that the model is very unreliable with annual reports. Analyst reports are potentially more suitable for the model. We argue that the reason could originate from the design choices behind both models. The topic weight model by nature measures one specialised form of innovation by focusing on one topic. However, the topic distribution method evaluates the strength of several topics making it more suitable for measuring many dimensions of innovation. This feature can make it more suitable for measuring innovation from more heterogeneous sources like annual reports.

4.1.2 | Patenting firm subsample

We proceed by estimating Equations (1) and (2) for a subsample consisting of patenting firms only. The criterion for this subsample is that the patent count of the firm i in year t is more

35		40		60		80	
log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)	log (*Patents*)	log (*Citations*)
0.7998*** (0.1153)	1.0620*** (0.1771)	−0.9125 (2.3115)	−1.2320 (3.1606)	−0.2100 (0.1831)	0.4151 (0.2787)	0.9685*** (0.1405)	1.1144*** (0.2170)
−0.0081*** (0.0009)	−0.0090*** (0.0012)	−0.0077*** (0.0008)	−0.0082*** (0.0012)	−0.0077*** (0.0008)	−0.0083*** (0.0012)	−0.0077*** (0.0008)	−0.0086*** (0.0012)
0.00005* (0.00004)	0.00005 (0.00004)	0.00006** (0.00003)	0.00005 (0.00004)	0.00005* (0.00003)	0.00005 (0.00004)	0.00005* (0.00003)	0.00004 (0.00004)
0.0718** (0.0296)	0.0742 (0.0444)	0.0873*** (0.0286)	0.0883** (0.0434)	0.0798*** (0.0291)	0.0809* (0.0438)	0.0692** (0.0291)	0.0723 (0.0442)
0.4856*** (0.0287)	0.5509*** (0.0444)	0.4589*** (0.0276)	0.5098*** (0.0430)	0.4684*** (0.0281)	0.5194*** (0.0434)	0.4840*** (0.0282)	0.5461*** (0.0441)
−0.0198 (0.0389)	−0.1476*** (0.0562)	−0.0269 (0.0384)	−0.1548*** (0.0549)	−0.0172 (0.0387)	−0.1399** (0.0560)	−0.0382 (0.0389)	−0.1681*** (0.0562)
−0.0001 (0.0005)	−0.0003 (0.0007)	−0.0003 (0.0005)	−0.0005 (0.0007)	−0.0004 (0.0005)	−0.0006 (0.0007)	−0.0002 (0.0005)	−0.0004 (0.0007)
−0.0188 (0.0304)	−0.0289 (0.0469)	−0.0062 (0.0300)	−0.0069 (0.0463)	−0.0104 (0.0303)	0.0051 (0.0468)	−0.0165 (0.0303)	−0.0241 (0.0468)
X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X
4401	4401	4408	4408	4401	4401	4401	4401
0.4944	0.3380	0.4882	0.3273	0.4875	0.3314	0.4949	0.3366

than 0. By focusing on patenting firms, we want to test, among other things, whether the topic weight method would be more reliable with the subset. A specific patent-type innovation should be more pronounced with patenting firms and the topic weight method might be able to capture this specific type when estimating innovation.

We start by analysing the topic distribution method. The results for Equations (1) and (2) are provided in Table 5. The results are qualitatively similar to the full sample results, and the model shows good performance especially when predicting patent citations. These results indicate that the topic distribution method is a good and consistent measure of innovation, both among patenting and non-patenting firms.

Next, we analyse the topic weight method with the patenting firm subsample. The results are provided in Table 6. The results are similar to the full sample, and the topic weight method still appears to be a relatively unreliable measure of innovation when used with 10-K filings. With the patenting firm subsample, there are significant associations for both directions, depending on the number of predefined topics. Thus, even though the subset of patenting firms might have a simpler innovation structure, there appears to be no improvement in the performance of the topic weight method for the patenting firm subsample. The reason for inconsistent results from the topic weight method could be the different nature of 10-K filings, which is not suitable for an LDA architecture that focuses on measuring one type of innovation (one topic).

4.2 | Firm performance – further analysis

We continue our analysis by assessing, how text-based innovation measures are associated with firm performance. To achieve this, we undertake panel regression analyses wherein firm-year

TABLE 7 The performance of the topic distribution measure of innovation on return on assets in the time periods $t+1$ and $t+2$.

	Dependent variable							
	ROA							
No. of topics	**15**		**20**		**25**		**30**	
Time period	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$
Explanatory variable								
Text innovation (topic dist.)	0.2536 (0.1780)	0.2431 (0.1789)	0.3707* (0.1935)	0.1524 (0.2109)	0.3699* (0.2071)	0.2421 (0.2075)	0.6284**** (0.2338)	0.5808** (0.2458)
log (*Patents*)								
R&D/Sales	−0.0115*** (0.0011)	−0.0115*** (0.0012)	−0.0047*** (0.0008)	−0.0052*** (0.0008)	−0.0048*** (0.0008)	−0.0052*** (0.0008)	−0.0047*** (0.0008)	−0.0051*** (0.0008)
log (*Sales*)	3.9726*** (0.1770)	3.8524*** (0.1713)	5.0253*** (0.2400)	4.4689*** (0.2486)	5.0068*** (0.2400)	4.4619*** (0.2499)	5.0160*** (0.2401)	4.4736*** (0.2498)
log (*Age*)	1.5214*** (0.4267)	1.5642*** (0.4102)	1.7842*** (0.4925)	1.7466*** (0.4944)	1.8041*** (0.4920)	1.7553*** (0.4940)	1.7289*** (0.4943)	1.6801*** (0.4945)
Total debt/Total capital	−0.0595*** (0.0142)	−0.0412*** (0.0135)	−0.0605*** (0.0209)	−0.0382 (0.0238)	−0.0601*** (0.0209)	−0.0378 (0.0237)	−0.0599*** (0.0208)	−0.0376 (0.0237)
Beta	−4.1923*** (0.4388)	−4.4474*** (0.4415)	−4.5478*** (0.5690)	−4.3603*** (0.5752)	−4.5501*** (0.5667)	−4.3535*** (0.5723)	−4.5250*** (0.5667)	−4.3242*** (0.5728)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	6746	6751	7404	6718	7404	6718	7404	6718
R^2	0.3186	0.3119	0.2960	0.2772	0.2959	0.2773	0.2964	0.2780

Note: The table contains panel regressions for ROA $t+1$ and ROA $t+2$ on the topic distribution measure of innovation. The columns present the topic count of the base-LDA-model for *Text innovation* and the logged patent count, representing the variables of interest. Other controls include firm-year observations for *R&D/Sales*, log (*Sales*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

observations of ROA and Tobin's Q, both 1 and 2 years ahead, are employed as dependent variables. We then compare the results of the text innovation models with the predictive power of patent counts on firm performance. We specify the following equations:

$$ROA_{it+1\ it+2} = \begin{aligned}&\alpha+\beta_1 Innovation_{it}+\beta_2\log(Patents)_{it}+\beta_3\log(Citations)_{it}+\beta_4 R\&D/Sales_{it}+\\&\beta_5\log(Sales)_{it}+\beta_6\log(Age)_{it}+\beta_7\frac{Total\ debt}{Total\ capital}_{it}+\beta_8 Beta_{it}+\mu_j+\gamma_t+\varepsilon_{it},\end{aligned} \quad (3)$$

$$\log(Tobin's\ Q)_{it+1\ it+2} = \begin{aligned}&\alpha+\beta_1 Innovation_{it}+\beta_2\log(Patents)_{it}+\beta_3\log(Citations)_{it}+\beta_4 R\&D/Sales_{it}+\\&\beta_5\log(Sales)_{it}+\beta_6\log(Age)_{it}+\beta_7\frac{Total\ debt}{Total\ capital}_{it}+\beta_8 Beta_{it}+\mu_j+\gamma_t+\varepsilon_{it}.\end{aligned} \quad (4)$$

We begin by estimating Equation (3) for the topic distribution method by using eight different innovation measures, the 15, 20, 25, 30, 35, 40, 60 and 80 topic LDA models, as previously, in addition to the patent counts for comparison. The regression results for ROA_{t+1} and ROA_{t+2} are presented in Table 7. The topic distribution measure is consistently associated with the $t+1$ return on assets. The coefficient is positive and statistically significant at the 1% level for 15, 30, 35, 40 and 80 topic models and at the 5% level for 20 and 25 topic models. Similar conclusions can be made also for the $t+2$ model, where the text innovation variables have statistically significant positive coefficients for the 15, 30, 35, 40 and 80 topic models as well. The logged patent

35		40		60		80		log (*Patents*)	
t+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2
1.2324*** (0.2649)	1.0271*** (0.2635)	0.08395*** (0.1545)	0.8930*** (0.1612)	0.3846* (0.2078)	0.1354 (0.2010)	1.3876*** (0.3311)	0.9754*** (0.3319)		
								−2.8468*** (0.2334)	−2.4316*** (0.2398)
−0.0046*** (0.0008)	−0.0050*** (0.0008)	−0.0113*** (0.0010)	−0.0116*** (0.0011)	−0.0048*** (0.0008)	−0.0052*** (0.0008)	−0.0047*** (0.0008)	−0.0051*** (0.0008)	−0.0038*** (0.0008)	−0.0043*** (0.0008)
5.0810*** (0.2412)	4.5308*** (0.2501)	4.3799*** (0.1576)	4.0082*** (0.1617)	5.0080*** (0.2400)	4.4608*** (0.2500)	5.0136*** (0.2397)	4.4692*** (0.2499)	6.1741*** (0.2982)	5.4459*** (0.3125)
1.6810*** (0.4947)	1.6614*** (0.4950)	1.8793*** (0.3658)	1.7691*** (0.3723)	1.7892*** (0.4936)	1.7499**** (0.4947)	1.6698*** (0.4928)	1.6678*** (0.4937)	2.5418*** (0.5001)	2.3859*** (0.5021)
−0.0568*** (0.0209)	−0.0352 (0.0237)	−0.0672*** (0.0123)	−0.0360*** (0.0129)	−0.0602*** (0.0209)	−0.0382 (0.0237)	−0.0582*** (0.0209)	−0.0367 (0.0237)	−0.0651*** (0.0207)	−0.0398* (0.0232)
−4.5065*** (0.5667)	−4.3097*** (0.5722)	−4.3791*** (0.4039)	−4.6162*** (0.4211)	−4.5632*** (0.5684)	−4.3659*** (0.5738)	−4.5053*** (0.5664)	−4.3197*** (0.5737)	−4.5904*** (0.5603)	−4.3802*** (0.5657)
X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
7404	6718	8188	7538	7404	6718	7404	6718	7404	6718
0.2988	0.2795	0.3509	0.3350	0.2959	0.2771	0.2981	0.2785	0.3159	0.2926

count exhibits a negative and statistically significant result for predicting ROA at $t+1$ and $t+2$ time periods, which means that it is not a good predictor of firm performance under this empirical setting. The topic distribution method outperforms the patent count in these models.

We proceed with the utilisation of Equation (4) concerning Tobin's Q, combined with topic distribution measures of innovation and the patent count proxy. The findings pertaining to the $t+1$ and $t+2$ models are presented in Table 8. In particular, our investigation reveals that a statistically significant and positive association between text innovation and Tobin's Q exists for the $t+1$ time period at the 1% level in models 40, 60 and 80, whereas in the remaining models, the coefficient fails to attain statistical significance. Conversely, during the $t+2$ time period, the relationship between text innovation and Tobin's Q becomes more pronounced. We observe statistically significant and positive associations at either the 1% or 5% level in the models featuring 15, 25, 40, 60 and 80 topics. In this empirical setting the patent count is a good predictor of Tobin's Q, and the coefficient is statistically significant and positive at 1% level for both periods.

Next, we estimate all of the previous models with the topic weight innovation measure and compare them with the patent count proxy as well. The regression results for ROA_{t+1} and ROA_{t+2} are presented in Table 9. The analysis reveals statistically significant positive coefficients for the topic weight measure in predicting ROA at the 1% level for models 20, 25, 30, 35 and 80 during the $t+1$ time period. In contrast, models 15, 40 and 60 do not exhibit statistically significant associations. Furthermore, the statistical significance levels for the ROA_{t+2} time period models align closely with those observed in the $t+1$ models. The topic

TABLE 8 The performance of the topic distribution measure of innovation on Tobin's Q in the time periods $t+1$ and $t+2$.

	Dependent variable							
	Q							
No. of topics	**15**		**20**		**25**		**30**	
Time period	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$
Explanatory variable								
Text innovation (topic dist.)	0.0146*** (0.0036)	0.0063** (0.0031)	0.0022 (0.0033)	0.0041 (0.0036)	0.0058* (0.0035)	0.0092** (0.0037)	0.0007 (0.0042)	0.0039 (0.0045)
log (*Patents*)								
R&D/Sales	0.000002 (0.00002)	0.0000003 (0.00001)	−0.000003 (0.00001)	0.000001 (0.00001)	−0.000003 (0.00001)	0.000002 (0.00001)	−0.000004 (0.00001)	0.000001 (0.00001)
log (*Sales*)	−0.0048 (0.0030)	−0.0069** (0.0030)	−0.0091*** (0.0030)	−0.0082*** (0.0031)	−0.0092*** (0.0030)	−0.0084*** (0.0030)	−0.0092*** (0.0030)	−0.0084*** (0.0031)
log (*Age*)	−0.0598** (0.0087)	−0.0529*** (0.0084)	−0.0589*** (0.0085)	−0.0567*** (0.0088)	−0.0588*** (0.0085)	−0.0565*** (0.0087)	−0.0587*** (0.0086)	−0.0570*** (0.0088)
Total debt/Total capital	0.00001 (0.0003)	0.0006** (0.0002)	0.0002 (0.0002)	0.0006** (0.0003)	0.0002 (0.0002)	0.0006** (0.0003)	0.0002 (0.0002)	0.0006** (0.0003)
Beta	−0.0047 (0.0092)	−0.0202** (0.0085)	−0.0100 (0.0086)	−0.0132 (0.0088)	−0.0097 (0.0085)	−0.0128 (0.0088)	−0.0103 (0.0085)	−0.0132 (0.0088)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	6657	7531	7401	6713	7401	6713	7401	6713
R^2	0.0132	0.0102	0.0103	0.0107	0.0107	0.0117	0.0102	0.0106

Note: The table contains panel regressions for Tobin's Q $t+1$ and Tobin's Q $t+2$ on the topic distribution measure of innovation. The columns present the topic count of the base-LDA-model for *Text innovation* and the logged patent count, representing the variables of interest. Other controls include firm-year observations for *R&D/Sales*, log (*Sales*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

weight innovation measure demonstrates consistent predictive ability for ROA across various financial models and also outperforms the patent count proxy in predicting ROA.

Finally, we estimate Equation (4) for Tobin's Q in the time periods $t+1$ and $t+2$ with the topic weight innovation measure. The regression results are presented in Table 10. The coefficients for the $t+1$ models are statistically significant and negative at the 1% or 5% level for models 20, 25, 30, 35, 40 and 80, whereas model 15 is not statistically significant and model 60 is statistically significant and positive. The $t+2$ models' results are similar to the former. The coefficients on models with 15, 25, 30, 35, 40 and 80 topics are statistically significant and negative, whereas the model with 20 topics is not statistically significant and the model with 60 topics is also in this time period statistically significant and positive. The adverse relationship observed between the 60 topic text innovation variables and Tobin's Q represents an undesired outcome, and consequently, the measurement fails to exhibit consistent predictive capacity for innovation in this empirical context. On the contrary, the patent count proxy is a consistent predictor of Tobin's Q and outperforms the topic weight method in this setting.

The regression results on the topic distribution innovation measure and firm performance are overall consistent. We can conclude that a good score with the topic distribution measure of innovation is likely followed by higher ROA and Tobin's Q in the following 2 years. Overall, the association between the variables is strongly pronounced in most cases. The patent count proxy for innovation had mixed results in the aforementioned models, since the association was positive for Tobin's Q but negative for ROA. In direct comparison, the topic distribution innovation measure proves more resolute than the patent count proxy. The results on the topic weight text innovation

35		40		60		80		log (*Patents*)	
t+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2
0.0010 (0.0046)	0.0035 (0.0048)	−0.0007 (0.0169)	0.0096 (0.0178)	0.0107*** (0.0035)	0.0137*** (0.0037)	0.0176*** (0.0058)	0.0233*** (0.0062)		
								0.0397*** (0.0040)	0.0420*** (0.0042)
−0.000004 (0.00001)	0.000001 (0.00001)	0.0001 (0.0001)	0.00001 (0.0001)	−0.000002 (0.00001)	0.000003 (0.00001)	−0.000002 (0.00001)	0.000003 (0.00001)	−0.000002 (0.00001)	0.00001 (0.00001)
−0.0093*** (0.0031)	−0.0083*** (0.0031)	−0.0530*** (0.0143)	−0.0645*** (0.0161)	−0.0091*** (0.0030)	−0.0083*** (0.0030)	−0.0091*** (0.0030)	−0.0083*** (0.0030)	−0.0255*** (0.0037)	−0.0256*** (0.0037)
−0.0586*** (0.0086)	−0.0568*** (0.0088)	−0.2910*** (0.0422)	−0.2638*** (0.0448)	−0.0593**** (0.0085)	−0.0570*** (0.0087)	−0.0605*** (0.0085)	0.0586*** (0.0087)	−0.0690*** (0.0086)	−0.0674*** (0.0088)
0.0002 (0.0002)	0.0006** (0.0003)	−0.0004 (0.0014)	0.0019 (0.0015)	0.0002 (0.0002)	0.0006** (0.0002)	0.0002 (0.0002)	0.0006** (0.0002)	0.0002 (0.0002)	0.0006** (0.0002)
−0.0103 (0.0085)	−0.0133 (0.0088)	−0.0402 (0.0421)	−0.0615 (0.0456)	−0.0098 (0.0085)	−0.01303 (0.0088)	−0.0093 (0.0085)	−0.0123 (0.0088)	−0.0100 (0.0085)	−0.0132 (0.0088)
X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
7401	6713	8188	7538	7401	6713	7401	6713	7401	6713
0.0102	0.0105	0.0120	0.0107	0.0118	0.0130	0.0119	0.0135	0.0263	0.284

method and firm performance variables are also somewhat inconsistent with even negative associations in the case of Tobin's Q. The measure is associated with firm performance for many different numbers of topics. The findings suggest that 10-K filings can serve as a reliable source for measuring innovation, particularly when using the topic distribution method. However, the effectiveness of the topic weight method was somewhat lacking. This emphasises the need for thorough testing when developing text-based innovation measures with the LDA algorithm.[1]

In summary, the assessment utilising the 10-K for innovation measurement underscores the importance of corporate communication regarding strategic decisions and innovative pursuits within these filings. The findings further emphasise the advantages that investors may derive from such disclosures, affirming the alignment between a company's stated objectives and its actions. Overall, the disclosure of innovation within these documents serves as a crucial tool for companies to depict their comprehensive innovation strategies, facilitating enhanced stakeholder involvement and transparent strategic operations.

5 | ADDITIONAL ROBUSTNESS TESTS

In order to conduct a comprehensive assessment of the efficacy of the innovation measurement method, this section undertakes a further examination of the text-based innovation

[1]We further investigate the reasons behind the inconsistency of the topic weight measure in Appendix S1.

TABLE 9 The performance of the topic weight measure of innovation on return on assets in the time periods $t+1$ and $t+2$.

	Dependent variable							
	ROA							
No. of topics	**15**		**20**		**25**		**30**	
Time period	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$
Explanatory variable								
Text innovation (topic weight)	2.4874 (1.8098)	2.4928 (1.9127)	5.7960*** (1.6496)	4.5258*** (1.5948)	4.9823*** (1.7348)	4.1241** (1.6699)	5.2859*** (1.7024)	4.1801** (1.6627)
log (*Patents*)								
R&D/Sales	−0.0050*** (0.0008)	−0.0055*** (0.0008)	−0.0048*** (0.0008)	−0.0051*** (0.0008)	−0.0048*** (0.0008)	−0.0051*** (0.0008)	−0.0048*** (0.0008)	−0.0051*** (0.0008)
log (*Sales*)	5.0211*** (0.2205)	4.6110*** (0.2339)	5.0267*** (0.2400)	4.4953*** (0.2524)	5.0385*** (0.2405)	4.5029*** (0.2521)	5.0412*** (0.2405)	4.5057*** (0.2522)
log (*Age*)	1.8740*** (0.4524)	1.7083*** (0.4558)	1.7005*** (0.4911)	1.6697*** (0.4949)	1.7419*** (0.4914)	1.6986*** (0.4955)	1.7008*** (0.4917)	1.6656*** (0.4951)
Total debt/Total capital	−0.0616*** (0.0196)	−0.0397* (0.0221)	−0.0562*** (0.0214)	−0.0359 (0.0242)	−0.0587*** (0.0212)	−0.0374 (0.0242)	−0.0578*** (0.0213)	−0.0369 (0.0242)
Beta	−4.3867*** (0.5639)	−4.5175*** (0.5616)	−4.6096*** (0.5715)	−4.4227*** (0.5764)	−4.6172*** (0.5737)	−4.4328*** (0.5788)	−4.6039*** (0.5931)	−4.4218*** (0.5782)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	8188	7536	7392	6695	7392	6695	7392	6695
R^2	0.2981	0.2892	0.2982	0.2788	0.2977	0.2785	0.2978	0.2785

Note: The table contains panel regressions for ROA $t+1$ and ROA $t+2$ on the topic weight measure of innovation. The columns present the topic count of the base-LDA-model for *Text innovation* and the logged patent count, representing the variables of interest. Other controls include firm-year observations for *R&D/Sales*, log (*Sales*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

measure. We start by analysing the software industry where companies very rarely patent their innovations. Furthermore, the subgroup of non-patenting firms in our sample is qualitatively scrutinised to provide additional insights. Additionally, the impact of alternative sources of innovation text on the final measure is empirically tested. The topic distribution method is chosen as the subject of further analysis. While the topic weight method holds potential for further improvement and experimentation, considering the specific context of the current study, the topic distribution method emerges as a more reliable innovation metric with 10-Ks.

5.1 | Non-patenting firms

Descriptive statistics of the topic distribution measure of innovation for the software industry subsample (SIC code 7372) are presented in Table 11. At the bottom of the table, we report and test the difference in means compared to the full sample. The difference in means is statistically significant at the 1% level in every model. The software industry is generally seen as innovative despite having low patenting rates. Thus, this finding supports the assumption that the topic distribution innovation measure can measure innovativeness that is not only patent-based, and it assigns significantly better innovation scores for software industry firms.

35		40		60		80		log (*Patents*)	
t+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2
13.999*** (2.1807)	13.446*** (2.2046)	−25.042 (19.967)	−27.173 (21.726)	0.4770 (2.4293)	0.0450 (2.4018)	7.6565*** (2.3205)	6.4734*** (2.3689)		
								−2.8468*** (0.2334)	−2.4316*** (0.2398)
−0.0046*** (0.0008)	−0.0050*** (0.0008)	−0.0049*** (0.0008)	−0.0055*** (0.0008)	−0.0048*** (0.0008)	−0.0052*** (0.0008)	−0.0048*** (0.0008)	−0.0051*** (0.0008)	−0.0038*** (0.0008)	−0.0043*** (0.0008)
5.1149*** (0.2423)	4.5839*** (0.2550)	5.0342*** (0.2236)	4.5797*** (0.2396)	5.0256*** (0.2398)	4.4908*** (0.2518)	5.0295*** (0.2402)	4.4963*** (0.2523)	6.1741*** (0.2982)	5.4459*** (0.3125)
1.6107*** (0.4888)	1.5755*** (0.4917)	1.8817*** (0.4570)	1.6625*** (0.4604)	1.7851*** (0.4907)	1.7294*** (0.4948)	1.6828**** (0.4901)	1.6478*** (0.4938)	2.5418*** (0.5001)	2.3859*** (0.5021)
−0.0537** (0.0210)	−0.0326 (0.0238)	−0.0593*** (0.0199)	−0.0382* (0.0227)	−0.0627*** (0.0210)	−0.0400* (0.0238)	−0.0600*** (0.0210)	−0.0382 (0.0240)	−0.0651*** (0.0207)	−0.0398* (0.0232)
−4.7246*** (0.5740)	−4.5443*** (0.5792)	−4.4220*** (0.5686)	−4.4405*** (0.5687)	−4.5707**** (0.5682)	−4.3986*** (0.5755)	−4.6228*** (0.5734)	−4.4399*** (0.5785)	−4.5904*** (0.5603)	−4.3802*** (0.5657)
X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
7392	6695	8062	7375	7392	6695	7392	6695	7404	6718
0.3026	0.2835	0.2992	0.2870	0.2967	0.2778	0.2978	0.2786	0.3159	0.2926

5.2 | Qualitative inspection

We proceed by examining more closely some of the 10-K filings of non-patenting firms with the highest innovation scores. Table 12 provides example quotes from the annual reports. The excerpts were chosen based on their high innovation score ranking and evaluated using a topic model with 40 topics. As examples, we choose three companies with no patents filed during the year. The excerpts are examples of occurrences in which the companies talk about their business, innovative activities, R&D tasks or intellectual property.

The selected firms are Simulations Plus, Inc., Exponent, Inc. and Copart, Inc. As can be seen from the 2011 annual report of Simulations Plus, Inc., it operates in the highly competitive industry of pharmaceuticals, where the key to success is a significant investment in R&D (excerpt 1). Even though Simulations Plus, Inc. states in excerpt (3) that they hold two patents, their intellectual property is still related primarily to computer programs. Furthermore, the company's specific area of expertise is pharmaceutical research. Thus, the text-based metric recognises Simulations Plus, Inc. as a highly innovative firm, although it does not patent. Similar conclusions can be drawn from the second example of a non-patenting firm, Exponent Inc. Its entire team consists of researchers, and the key to its success is to provide innovative, cutting-edge solutions to its customers. The company aims to solve 'complicated issues' and uses different forms of 'analysis' in its solutions. To succeed in its field, the company needs to be more innovative than its competitors. Again, this innovativeness is recognised by the text-based metric, although Exponent Inc. is not patenting.

TABLE 10 The performance of the topic weight measure of innovation on Tobin's Q in the time periods $t+1$ and $t+2$.

	Dependent variable							
	Q							
No. of topics	**15**		**20**		**25**		**30**	
Time period	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$	$t+1$	$t+2$
Explanatory variable								
Text innovation (topic weight)	−0.1421** (0.0560)	−0.1234** (0.0598)	0.0019 (0.0297)	0.0281 (0.0303)	−0.0175 (0.0311)	0.0040 (0.0320)	−0.0389 (0.0313)	−0.0172 (0.0320)
log (*Patents*)								
R&D/Sales	0.000007 (0.00001)	0.00001 (0.00002)	0.000001 (0.00001)	0.000005 (0.00001)	0.000001 (0.00001)	0.000005 (0.00001)	0.0000004 (0.00001)	0.000005 (0.00001)
log (*Sales*)	0.0008 (0.0040)	0.0036 (0.0042)	−0.0039 (0.0029)	−0.0033 (0.0029)	−0.0039 (0.0029)	−0.0033 (0.0029)	−0.0040 (0.0029)	−0.0033 (0.0029)
log (*Age*)	−0.0823*** (0.0116)	−0.0774*** (0.0121)	−0.0584**** (0.0084)	−0.0561*** (0.0087)	−0.0582*** (0.0084)	−0.0557*** (0.0086)	−0.0578*** (0.0084)	−0.0554*** (0.0087)
Total debt/Total capital	0.0002 (0.0003)	0.0003 (0.0003)	0.00006 (0.0002)	0.0005* (0.0003)	0.00005 (0.0002)	0.0005* (0.0003)	0.00003 (0.0002)	0.0005* (0.0003)
Beta	−0.0287** (0.0117)	−0.0349*** (0.0124)	−0.0127 (0.0083)	−0.0175** (0.0086)	−0.0125 (0.0084)	−0.0174** (0.0086)	−0.0125 (0.0084)	−0.0173** (0.0086)
SIC4 FE	X	X	X	X	X	X	X	X
Year FE	X	X	X	X	X	X	X	X
Observations	8188	7438	7392	6695	7392	6695	7392	6695
R^2	0.0070	0.0065	0.0077	0.0084	0.0078	0.0082	0.0078	0.0083

Note: The table contains panel regressions for Tobin's Q $t+1$ and Tobin's Q $t+2$ on the topic weight measure of innovation. The columns present the topic count of the base-LDA-model for *Text innovation* and the logged patent count, representing the variables of interest. Other controls include firm-year observations for *R&D/Sales*, log (*Sales*), log (*Age*), *Total debt/total capital* and *Beta*. Industry and time fixed effects are included in each model and standard errors are robust. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

TABLE 11 Descriptive statistics for the topic distribution measure of innovation on the software industry subsample.

No. of topics	**15**	**20**	**25**	**30**	**35**	**40**	**60**	**80**
Count	922	922	922	922	922	922	922	922
Mean	2.6608	2.8397	3.3362	3.4475	4.2070	4.4641	3.9783	5.1587
Standard deviation	1.2880	1.3303	1.2835	1.2848	1.1443	1.2339	1.4977	1.0286
Median	2.5324	2.4686	3.0630	3.1724	4.2600	4.3363	3.4338	5.0897
t-test								
Full sample mean	4.9393	5.3595	5.7523	5.5409	5.7609	6.6926	6.6274	6.6231
Difference in means	2.2785***	2.5199***	2.1561***	2.0934***	1.5539***	2.2285***	2.6491***	1.4644***

Note: The table contains descriptive statistics for the topic distribution measure of innovation on the software industry subsample. In the last two rows we report the full sample mean and the difference in means between the full sample and the software industry subsample, and conduct a *t*-test for the difference in means. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

Finally, Copart Inc. is an online auction platform provider, with software developed in-house. Copart Inc. discusses development activities in various locations in their 2018 annual report (excerpts 2 and 3). They state that they see their platform as a competitive advantage

35		40		60		80		log (*Patents*)	
t+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2	*t*+1	*t*+2
−0.0622* (0.0359)	−0.0380 (0.0370)	−1.6816*** (0.3775)	−1.6249*** (0.5020)	0.2650*** (0.0482)	0.3273*** (0.0497)	−0.0508 (0.0423)	−0.0149 (0.0439)		
								0.0397*** (0.0040)	0.0420*** (0.0042)
0.0000001 (0.00001)	0.000004 (0.00001)	0.000006 (0.00001)	0.000003 (0.00001)	0.000002 (0.00001)	0.000006 (0.00001)	0.000001 (0.00001)	0.000005 (0.00001)	−0.000002 (0.00001)	0.00001 (0.00001)
−0.0043 (0.0030)	−0.0036 (0.0030)	−0.0007 (0.0028)	−0.0011 (0.0029)	−0.0049* (0.0030)	−0.0045 (0.0030)	−0.0039 (0.0029)	−0.0033 (0.0029)	−0.0255*** (0.0037)	−0.0256*** (0.0037)
−0.0576*** (0.0084)	−0.0553*** (0.0086)	−0.0579*** (0.0080)	−0.0543*** (0.0083)	−0.0566*** (0.0084)	−0.0536*** (0.0086)	−0.0577*** (0.0084)	−0.0555*** (0.0086)	−0.0690*** (0.0086)	−0.0674*** (0.0088)
0.00003 (0.0002)	0.0005* (0.0003)	0.00001 (0.0002)	0.0005** (0.0002)	0.00001 (0.0002)	0.0005** (0.0002)	0.00005 (0.0002)	0.0005* (0.0003)	0.0002 (0.0002)	0.0006** (0.0002)
−0.0120 (0.0084)	−0.0170** (0.0086)	−0.0197** (0.0081)	−0.0254*** (0.0084)	−0.0111 (0.0084)	−0.0157* (0.0087)	−0.0124 (0.0084)	−0.0173** (0.0086)	−0.0100 (0.0085)	−0.0132 (0.0088)
X	X	X	X	X	X	X	X	X	X
X	X	X	X	X	X	X	X	X	X
7392	6695	8062	7375	7392	6695	7392	6695	7401	6713
0.0082	0.0084	0.0092	0.0096	0.0136	0.0173	0.0080	0.0083	0.0263	0.284

and mention several competitive benefits from their in-house development activities, even though the activities have not resulted in filed patents.

The overall conclusion from the text excerpts is that the text-based metric can recognise the same dimension of innovation that can be estimated using patent counts and citations. However, more importantly, the examples of two non-patenting firms demonstrate that the metric can identify dimensions of innovation that cannot be measured using patent counts. All examples of non-patenting firms focus on research and development, and the text metric recognises this innovativeness. The text-based metric provides an alternative measure for innovation where all companies are on the same line, and innovation is measured more broadly and includes forms of innovation other than patents.

5.3 | Alternative text sources

We used the same innovation textbook as Bellstam et al. (2020) for all the analyses in the previous sections. In this section, we analyse the possible variation in the outcomes depending on the text source by generating alternative metrics from different innovation texts (Table A2). The following examples are based on the 40-topic model since the topic distribution method was most consistent with this model. We study the relationship between the measure and other innovation proxies by specifying the following equation:

TABLE 12 Excerpts from the 10-K filings of innovative non-patenting companies.

Company	Text Extract
Simulations Plus, Inc. (2011)	(1) Simulations Plus, Inc., which was incorporated in California in 1996 […] develops and produces software for use in pharmaceutical research and for education, as well as providing contract research services to the pharmaceutical industry. (2) We believe that our ability to grow and remain competitive in our markets is strongly dependent on significant investment into research and development ('R&D'). R&D activities include both enhancement of existing products and development of new products. […] R&D expenditures were approximately $1,846,000 during fiscal year 2011, of which $911,000 was capitalized. R&D expenditures during fiscal year 2010 were approximately $1,857,000, of which $887,000 was capitalized. (3) We own two patents that were acquired as part of our acquisition of certain assets of Bioreason, Inc. We primarily protect our intellectual property through copyrights and trade secrecy. Our intellectual property consists primarily of source code for computer programs and data files for various applications of those programs in both the pharmaceutical software and the disability products businesses. In the disability products business, electronic device schematics, mechanical drawings, and design details are also intellectual property. The expertise of our technical staff is a considerable asset closely related to intellectual property, and attracting and retaining highly qualified scientists and engineers is essential to our business.
Exponent, Inc. (2018)	(1) Exponent, Inc. […] is a science and engineering consulting firm that provides solutions to complex problems. Our multidisciplinary team of scientists, engineers, business and regulatory consultants brings together more than 90 different technical disciplines to solve complicated issues facing industry and government today. Our services include analysis of product development, product recall, regulatory compliance, and the discovery of potential problems related to products, people, property and impending litigation.
Copart, Inc. (2018)	(1) We are a leading provider of online auctions and vehicle remarketing services with operations in […]. Our goals are to generate sustainable profits for our stockholders, while also producing environmental and social benefits for the world, by promoting vehicle restoration, repair, and recycling; parts refurbishment and re-use; and facilitating the recovery and resilience of communities affected by severe climate events. (2) In addition, we have developed a database containing over 300 fields of real-time and historical information accessible by our sellers allowing for their generation of custom ad hoc reports and customer specific analysis. […] We have developed a computer system which provides a direct link to the DMV computer systems of multiple states, allowing us to expedite the processing of vehicle title paperwork. (3) We believe the introduction of our virtual auction platform increased the pool of available buyers for each sale, which resulted in added competition and an increase in the amount buyers are willing to pay for vehicles. We also believe that it improved the efficiency of our operations by eliminating the expense and capital requirements associated with live auctions.

$$Text\_inn_{it} = \alpha + \beta_1 \log\left(Patents_{it} + 1\right) + \beta_2 \log\left(Citations_{it} + 1\right) + \beta_3 R\&D/Sales_{it} + \mu_j + \gamma_t + \varepsilon_{it}. \quad (5)$$

The results are shown in Table 13 The different columns show the results for each text, numbered 1–5. As the results demonstrate, the innovation measure varies slightly depending on the innovation text source, and the method appears to be sensitive to the text source.

TABLE 13 Topic distribution measure of innovation with alternative innovation books (40-topic LDA) and different innovation proxies.

	Dependent variable				
	Alternative innovation book text innovation **(topic distribution, 40-topic)**				
Alternative text no.	**1**	**2**	**3**	**4**	**5**
log (*Patents*)	−0.0030 (0.0172)	−0.1173*** (0.0177)	−0.0247 (0.0200)	0.0851*** (0.0207)	0.1061*** (0.0290)
log (*Citations*)	−0.0191 (0.0131)	−0.0125 (0.0136)	−0.0420*** (0.0153)	−0.0665*** (0.0158)	−0.1173*** (0.0221)
R&D/sales	0.0001*** (1.59e-05)	0.0002*** (1.642e-05)	0.0002*** (1.849e-05)	0.0001*** (1.913e-05)	0.0001*** (2.683e-05)
SIC4 FE	X	X	X	X	X
Year FE	X	X	X	X	X
Observations	12,725	12,725	12,725	12,725	12,725
R^2	0.0071	0.0346	0.0148	0.0061	0.0041

Note: The table contains panel regressions for the topic distribution measure of innovation based on the 40-topic LDA model from alternative innovation textbooks. The dependent variable is the text-based innovation measure. The independent variables are innovation proxy variables. Industry and time fixed effects are included, as indicated. *, ** and *** indicate statistical significance at 10%, 5% and 1% levels, respectively.

Thus, these results give more evidence to the fact that great care is needed when incorporating text-based measures for abstract concepts like innovation. It is not just the LDA architecture but also the chosen reference texts that highly influence the performance and reliability of the metric.

6 | DISCUSSION AND CONCLUSIONS

In this study, we analysed text-based methods for measuring innovation from narrative disclosure in 10-K filings. Our study adds to the research on accounting and innovations (Bedford et al., 2021; Chenhall & Moers, 2015; He & Tian, 2018; Huang et al., 2021; Taipaleenmäki, 2014; Tang et al., 2021), extending the scope of research by specifically answering the recent calls in the accounting literature (Bellstam et al., 2020; Ranta et al., 2023) for better proxies of innovation by utilising ML methods to build a text-based proxy of innovation. Our study contributes to this literature by designing a measure that can estimate dimensions of innovation difficult for traditional proxies, like patent counts. We also analyse the robustness of text-based measures and identify an architecture-dependent sensitivity of these approaches for reliably measuring innovation. Our study thus makes methodological contributions to research on ML applications in accounting (Belloque et al., 2021; Cai et al., 2019; Clarkson et al., 2020; Zengul et al., 2021; Zhu et al., 2017).

The results demonstrate that the topic distribution method is a robust measure of innovation. It can identify patent-based innovation and is associated with the future financial performance of a company. However, while this design exhibited a degree of resilience to specific model adjustments (such as the number of topics), it still displayed some sensitivity to the selected reference innovation text. Conversely, our results indicate that the topic weight method lacked reliability when applied to 10-K filings. This method proved liable to alterations in LDA parameters, resulting in notable variations tied to the number of topics chosen for the LDA model. Moreover, it emerged as a relatively unreliable predictor of future performance. The influence of innovation on firm valuation and profitability is not straightforward, and previous research results on the topic have been mixed (Bowen et al., 2010; Feng, 2005; Hirshleifer et al., 2013; Jiménez-Jiménez & Sanz-Valle, 2011). The investigations of the present study regarding the topic weight method yielded similar, mixed results. Thus, although Bellstam et al. (2020) demonstrated a relatively good performance of the architecture with analyst reports, it appears that the structure of annual reports is such that they are not suitable for this method.

Since financial reports include statutory information, they necessarily include text that is unrelated to innovation. Thus, the legally required information could bias the results for the topic weight model. Furthermore, the required amount of disclosed information varies by company size, meaning that smaller companies can use a greater percentage of their statements to describe their products and R&D tasks. Only focusing on the share of the 'innovation topic' could lead to bias in which small companies with short 10-Ks could seem more innovative than large companies, which deteriorates the performance of the topic weight method. Overall, the results underline the fact that efficient text-based measures can be designed that use annual reports as an information source, but great care is needed when designing them. The constructed measure needs to be tested thoroughly to validate it before using it in practical applications.

Our results provide useful new information for future research seeking alternative data sources to measure innovation. Our findings are also valuable in terms of understanding the nature of innovation disclosure in 10-K filings. Using financial report narratives to measure innovativeness is a new method of innovation measurement, and the ability to predict future patents and citations per patent for large data sets at one time is also relevant for practice. In

addition, innovation measurement for firms without patents or R&D expenses is traditionally difficult but potentially possible using text-based methods.

The outcomes of our investigation also suggest that embracing open and transparent disclosure practices yields advantages for companies, investors and various stakeholders. Through the inclusion of strategic details and innovative initiatives within their 10-K filings, companies can effectively communicate valuable information to their stakeholders. This study substantiates the significance of these reports in delivering meaningful insights to stakeholders concerning the innovative facets embedded within corporate strategies and operations.

A possible limitation of this study may be impression management in firms' financial statements. If firms practice impression management and aim to seem more innovative in their financial reports than in reality, our innovation measurement could be positively biased. Also, vice versa, if innovative firms keep their innovations as trade secrets and do not disclose them, it is unclear whether our method would detect innovation correctly. The results with the topic distribution method regarding the future financial performance of a company are, thus, promising, as the measure was shown to be associated consistently with future performance. However, one possible source for the topic weight method's unreliability could be the impression management practices of a firm, if the method identifies typical 'hype' talk from the annual reports.

There are several avenues for future research. The innovation topic selection process for the topic weight method could be further developed to avoid relying entirely on the innovation textbook. In addition, for both methods, it would be useful to analyse the innovation text selection process further to define what kind of text source works best for measuring innovation using these methods. Finally, further research could investigate how much impression management or 'window dressing' affects the results and drives innovation-related disclosure in 10-K filings.

FUNDING INFORMATION

Support by the Evald & Hilda Nissi Foundation and OP Group Research Foundation, Grant/Award number: 20200132.

DATA AVAILABILITY STATEMENT

Data available on request from the authors.

ORCID

Essi Nousiainen https://orcid.org/0000-0002-3203-2723
Mika Ylinen https://orcid.org/0000-0003-3441-2129

REFERENCES

Acharya, V. & Xu, Z. (2017) Financial dependence and innovation: the case of public versus private firms. *Journal of Financial Economics*, 124(2), 223–243. Available from: https://doi.org/10.1016/J.JFINECO.2016.02.010

Ahmed, S., Ranta, M., Vähämaa, E. & Vähämaa, S. (2023) Facial attractiveness and CEO compensation: evidence from the banking industry. *Journal of Economics and Business*, 123, 106095. Available from: https://doi.org/10.1016/j.jeconbus.2022.106095

Antons, D., Grünwald, E., Cichy, P. & Salge, T.O. (2020) The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *R&D Management*, 50(3), 329–351. Available from: https://doi.org/10.1111/RADM.12408

Bao, Y., Ke, B., Li, B., Yu, Y.J. & Zhang, J. (2020) Detecting accounting fraud in publicly traded US firms using a machine learning approach. *Journal of Accounting Research*, 58(1), 199–235. Available from: https://doi.org/10.1111/1475-679X.12292

Basu, S., Ma, X. & Briscoe-Tran, H. (2022) Measuring multidimensional investment opportunity sets with 10-K text. *The Accounting Review*, 97(1), 51–73. Available from: https://doi.org/10.2308/TAR-2019-0110

Bedford, A., Ma, L., Ma, N. & Vojvoda, K. (2021) Patenting activity or innovative originality? *Accounting and Finance*, 61, 4191–4207. Available from: https://doi.org/10.1111/acfi.12730

Bedford, D.S., Bisbe, J. & Sweeney, B. (2019) Performance measurement systems as generators of cognitive conflict in ambidextrous firms. *Accounting, Organizations and Society*, 72, 21–37. Available from: https://doi.org/10.1016/j.aos.2018.05.010

Bei, C., Liu, S., Liao, Y., Tian, G. & Tian, Z. (2021) Predicting new cases of COVID-19 and the application to population sustainability analysis. *Accounting and Finance*, 61(3), 4859–4884. Available from: https://doi.org/10.1111/acfi.12785

Belloque, G., Linnenluecke, M.K., Marrone, M., Singh, A.K. & Xue, R. (2021) 55 years of *Abacus*: evolution of research streams and future research directions. *Abacus*, 57, 593–618. Available from: https://doi.org/10.1111/abac.12232

Bellstam, G., Bhagat, S. & Cookson, J.A. (2020) A text-based analysis of corporate innovation. *Management Science*, 67(7), 4004–4031. Available from: https://doi.org/10.1287/MNSC.2020.3682

Bertomeu, J., Cheynel, E., Floyd, E. & Pan, W. (2021) Using machine learning to detect misstatements. *Review of Accounting Studies*, 26(2), 468–519. Available from: https://doi.org/10.1007/s11142-020-09563-8

Bisbe, J. & Malagueño, R. (2009) The choice of interactive control systems under different innovation management modes. *The European Accounting Review*, 18, 371–405. Available from: https://doi.org/10.1080/09638180902863803

Blei, D.M., Ng, A.Y., Jordan, M.I. & Lafferty, J. (2003) Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(4/5), 993–1022.

Bowen, F.E., Rostami, M. & Steel, P. (2010) Timing is everything: a meta-analysis of the relationships between organizational performance and innovation. *Journal of Business Research*, 63(11), 1179–1185. Available from: https://doi.org/10.1016/J.JBUSRES.2009.10.014

Brown, N.C., Crowley, R.M. & Elliott, W.B. (2020) What are you saying? Using *topic* to detect financial misreporting. *Journal of Accounting Research*, 58(1), 237–291. Available from: https://doi.org/10.1111/1475-679X.12294

Buehlmaier, M.M.M. & Whited, T.M. (2018) Are financial constraints priced? Evidence from textual analysis. *The Review of Financial Studies*, 31(7), 2693–2728. Available from: https://doi.org/10.1093/RFS/HHY007

Cai, C.W., Linnenluecke, M.K., Marrone, M. & Singh, A.K. (2019) Machine learning and expert judgement: analyzing emerging topics in accounting and finance research in the Asia–Pacific. *Abacus*, 55, 709–733. Available from: https://doi.org/10.1111/abac.12179

Cai, X., Pan, H., Gao, C., Wang, C. & Lu, L. (2021) Top executive tournament incentives and corporate innovation output. *Accounting and Finance*, 61, 5893–5924. Available from: https://doi.org/10.1111/acfi.12850

Chang, X., Fu, K., Low, A. & Zhang, W. (2015) Non-executive employee stock options and corporate innovation. *Journal of Financial Economics*, 115, 168–188. Available from: https://doi.org/10.1016/j.jfineco.2014.09.002

Chenhall, R.H. & Moers, F. (2015) The role of innovation in the evolution of management accounting and its integration into management control. *Accounting, Organizations and Society*, 47, 1–13. Available from: https://doi.org/10.1016/j.aos.2015.10.002

Ciftci, M. & Zhou, N. (2016) Capitalizing R&D expenses versus disclosing intangible information. *Review of Quantitative Finance and Accounting*, 46(3), 661–689. Available from: https://doi.org/10.1007/S11156-014-0482-0

Clarkson, P.M., Ponn, J., Richardson, G.D., Rudzicz, F., Tsang, A. & Wang, J. (2020) A textual analysis of US corporate social responsibility reports. *Abacus*, 56, 3–34. Available from: https://doi.org/10.1111/abac.12182

Cooper, M., Knott, A. & Yang, W. (2022) RQ innovative efficiency and firm value. *Journal of Financial and Quantitative Analysis*, 57(5), 1649–1694. Available from: https://doi.org/10.1017/S0022109021000417

Ding, K., Lev, B., Peng, X., Sun, T. & Vasarhelyi, M.A. (2020) Machine learning improves accounting estimates: evidence from insurance payments. *Review of Accounting Studies*, 25(3), 1098–1134. Available from: https://doi.org/10.1007/s11142-020-09546-9

Donovan, J., Jennings, J., Koharki, K. & Lee, J. (2021) Measuring credit risk using qualitative disclosure. *Review of Accounting Studies*, 26(2), 815–863. Available from: https://doi.org/10.1007/S11142-020-09575-4

Dyer, T., Lang, M. & Stice-Lawrence, L. (2017) The evolution of 10-K textual disclosure: evidence from latent Dirichlet allocation. *Journal of Accounting and Economics*, 64(2–3), 221–245. Available from: https://doi.org/10.1016/j.jacceco.2017.07.002

Dziallas, M. & Blind, K. (2019) Innovation indicators throughout the innovation process: an extensive literature analysis. *Technovation*, 80–81, 3–29. Available from: https://doi.org/10.1016/j.technovation.2018.05.005

Feng, G.U. (2005) Innovation, future earnings, and market efficiency. *Journal of Accounting, Auditing and Finance*, 20(4), 385–418. Available from: https://doi.org/10.1177/0148558x0502000405

Frankel, R., Jennings, J. & Lee, J. (2016) Using unstructured and qualitative disclosures to explain accruals. *Journal of Accounting and Economics*, 62(2–3), 209–227. Available from: https://doi.org/10.1016/J.JACCECO.2016.07.003

Garanina, T., Ranta, M. & Dumay, J. (2021) Blockchain in accounting research: current trends and emerging topics. *Accounting, Auditing & Accountability Journal*, 35(7), 1507–1533. Available from: https://doi.org/10.1108/AAAJ-10-2020-4991

Glaeser, S.A. & Landsman, W.R. (2021) Deterrent disclosure. *The Accounting Review*, 96(5), 291–315. Available from: https://doi.org/10.2308/TAR-2019-1050

Grabner, I., Posch, A. & Wabnegg, M. (2018) Materializing innovation capability: a management control perspective. *Journal of Management Accounting Research*, 30, 163–185. Available from: https://doi.org/10.2308/jmar-52062

Guo, B., Paraskevopoulou, E. & Sánchez, L.S. (2019) Disentangling the role of management control systems for product and process innovation in different contexts. *The European Accounting Review*, 28(4), 681–712. Available from: https://doi.org/10.1080/09638180.2018.1528168

Hall, B.H., Helmers, C., Rogers, M. & Sena, V. (2013) The importance (or not) of patents to UK firms. *Oxford Economic Papers*, 65(3), 603–629. Available from: https://doi.org/10.1093/oep/gpt012

He, J. & Tian, X. (2018) Finance and corporate innovation: a survey. *Asia-Pacific Journal of Financial Studies*, 47, 165–212. Available from: https://doi.org/10.1111/ajfs.12208

Helling, A.R., Maury, B. & Liljeblom, E. (2020) Exit as governance: do blockholders affect corporate innovation in large US firms? *Accounting and Finance*, 60, 1703–1725. Available from: https://doi.org/10.1111/acfi.12509

Henri, J.F. & Wouters, M. (2020) Interdependence of management control practices for product innovation: the influence of environmental unpredictability. *Accounting, Organizations and Society*, 86, 101073. Available from: https://doi.org/10.1016/J.AOS.2019.101073

Hirshleifer, D., Hsu, P.H. & Li, D. (2013) Innovative efficiency and stock returns. *Journal of Financial Economics*, 107(3), 632–654. Available from: https://doi.org/10.1016/j.jfineco.2012.09.011

Hoberg, G. & Maksimovic, V. (2015) Redefining financial constraints: a text-based analysis. *The Review of Financial Studies*, 28(5), 1312–1352. Available from: https://doi.org/10.1093/RFS/HHU089

Holmstrom, B. (1989) Agency costs and innovation. *Journal of Economic Behavior & Organization*, 12(3), 305–327. Available from: https://doi.org/10.1016/0167-2681(89)90025-5

Huang, H.J., Habib, A., Sun, S.L., Liu, Y. & Guo, H. (2021) Financial reporting and corporate innovation: a review of the international literature. *Accounting and Finance*, 61, 5439–5499. Available from: https://doi.org/10.1111/acfi.12764

Jiménez-Jiménez, D. & Sanz-Valle, R. (2011) Innovation, organizational learning, and performance. *Journal of Business Research*, 64(4), 408–417. Available from: https://doi.org/10.1016/J.JBUSRES.2010.09.010

Jones, S. & Alam, N. (2019) A machine learning analysis of citation impact among selected Pacific Basin journals. *Accounting and Finance*, 59(4), 2509–2552. Available from: https://doi.org/10.1111/acfi.12584

Kim, C.(F.), Wang, K. & Zhang, L. (2019) Readability of 10-K reports and stock price crash risk. *Contemporary Accounting Research*, 36(2), 1184–1216. Available from: https://doi.org/10.1111/1911-3846.12452

Kogan, L., Papanikolaou, D., Seru, A. & Stoffman, N. (2017) Technological innovation, resource allocation, and growth. *Quarterly Journal of Economics*, 132(2), 665–712. Available from: https://doi.org/10.1093/qje/qjw040

Lehavy, R., Li, F. & Merkley, K. (2011) The effect of annual report readability on analyst following and the properties of their earnings forecasts. *The Accounting Review*, 86(3), 1087–1115.

Lewis, C. & Young, S. (2019) Fad or future? Automated analysis of financial text and its implications for corporate reporting. *Accounting and Business Research*, 49(5), 587–615. Available from: https://doi.org/10.1080/00014788.2019.1611730

Liang, Y. (2022) The effect of capital and labour distortion on innovation. *Accounting and Finance*, 63, 1709–1737. Available from: https://doi.org/10.1111/acfi.12924

Lu, Q. & Chesbrough, H. (2022) Measuring open innovation practices through topic modelling: revisiting their impact on firm financial performance. *Technovation*, 114, 102434. Available from: https://doi.org/10.1016/J.TECHNOVATION.2021.102434

Moulang, C. (2015) Performance measurement system use in generating psychological empowerment and individual creativity. *Accounting and Finance*, 55, 519–544. Available from: https://doi.org/10.1111/acfi.12059

Mukherjee, A., Singh, M. & Žaldokas, A. (2017) Do corporate taxes hinder innovation? *Journal of Financial Economics*, 124(1), 195–221. Available from: https://doi.org/10.1016/j.jfineco.2017.01.004

Müller-Stewens, B., Widener, S.K., Möller, K. & Steinmann, J.-C. (2020) The role of diagnostic and interactive control uses in innovation. *Accounting, Organizations and Society*, 80, 101078. Available from: https://doi.org/10.1016/j.aos.2019.101078

Nuhu, N.A., Baird, K. & Su, S. (2022) The impact of interactive and diagnostic levers of eco-control on eco-innovation: the mediating role of employee environmental citizenship behaviour. *Accounting and Finance*, 63, 2245–2271. Available from: https://doi.org/10.1111/acfi.12967

Plečnik, J.M., Yang, L.L. & Zhang, J.H. (2022) Corporate innovation and future earnings: does early patent disclosure matter? *Accounting and Finance*, 62, 2011–2056. Available from: https://doi.org/10.1111/acfi.12851

Ranta, M. & Ylinen, M. (2023a) Employee benefits and company performance: evidence from a high-dimensional machine learning model. *Management Accounting Research*, 100876. Available from: https://doi.org/10.1016/j.mar.2023.100876

Ranta, M. & Ylinen, M. (2023b) Board gender diversity and workplace diversity: a machine learning approach. *Corporate Governance*, 23, 995–1018. Available from: https://doi.org/10.1108/CG-01-2022-0048

Ranta, M., Ylinen, M. & Järvenpää, M. (2023) Machine learning in management accounting research: literature review and pathways for the future. *The European Accounting Review*, 32, 607–636. Available from: https://doi.org/10.1080/09638180.2022.2137221

Saidi, F. & Žaldokas, A. (2021) How does firms' innovation disclosure affect their banking relationships? *Management Science*, 67(2), 742–768. Available from: https://doi.org/10.1287/mnsc.2019.3498

Speckbacher, G. & Wabnegg, M. (2020) Incentivizing innovation: the role of knowledge exchange and distal search behavior. *Accounting, Organizations and Society*, 86, 101142. Available from: https://doi.org/10.1016/J.AOS.2020.101142

Taipaleenmäki, J. (2014) Absence and variant modes of presence of management accounting in new product development – theoretical refinement and some empirical evidence. *The European Accounting Review*, 23(2), 291–334.

Tang, X., Shi, J., Han, J., Shu, A. & Xiao, F. (2021) Culturally diverse board and corporate innovation. *Accounting and Finance*, 61, 5655–5679. Available from: https://doi.org/10.1111/acfi.12772

Tidd, J., Bessant, J. & Pavitt, K. (2005) *Managing innovation: integrating technological, market and organizational change*. New York: John Wiley & Sons.

Ylinen, M. & Gullkvist, B. (2014) The effects of organic and mechanistic control in exploratory and exploitative innovation. *Management Accounting Research*, 25(1), 93–112.

Ylinen, M. & Ranta, M. (2023) Employer ratings in social media and firm performance: evidence from an explainable machine learning approach. *Accounting and Finance*, 1–30. Available from: https://doi.org/10.1111/acfi.13146

Zengul, F.D., Oner, N., Byrd, J.D. & Savage, A. (2021) Revealing research themes and trends in 30 top-ranking accounting journals: a text-mining approach. *Abacus*, 57, 468–501. Available from: https://doi.org/10.1111/abac.12214

Zhang, Z., Wu, H., Ying, S.X. & You, J. (2023) Corporate innovation and disclosure strategy. *Abacus*, 59, 76–133. Available from: https://doi.org/10.1111/abac.12248

Zhou, L.J. & Sadeghi, M. (2021) The long-run role of innovation in the IPO market: inhibition or promotion? *Accounting and Finance*, 61, 3735–3779. Available from: https://doi.org/10.1111/acfi.12799

Zhu, Y., Wu, Z., Zhang, H. & Yu, J. (2017) Media sentiment, institutional investors and probability of stock price crash: evidence from Chinese stock markets. *Accounting and Finance*, 57(5), 1635–1670. Available from: https://doi.org/10.1111/acfi.12355

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Nousiainen, E., Ranta, M., Ylinen, M. & Järvenpää, M. (2024) Using machine learning and 10-K filings to measure innovation. *Accounting & Finance*, 00, 1–29. Available from: https://doi.org/10.1111/acfi.13245

APPENDIX

TABLE A1 Variable definitions.

Variable	Definition
log (*Patents*)	The natural logarithm of the patent count of firm *i* in year *t*
log (*Citations*)	The natural logarithm of the patent citation count of firm *i* in year *t*
log (*Sales*)	The natural logarithm of the net sales of firm *i* in year *t*
log (*Assets*)	The natural logarithm of the total assets of firm *i* in year *t*
log (*Age*)	The natural logarithm of the age of firm *i* in year *t*
R&D/*Sales*	Research and development expenses of firm *i* in year *t* divided by sales
Total debt/*Total capital*	Total debt of firm *i* in year *t* divided by total capital
ROA	Return on assets of firm *i* in year *t*
log (*Q*)	Natural logarithm of the Tobin's Q of firm *i* in year *t*. Tobin's Q calculated as (total assets + market capital − equity)/total assets
Beta	Beta of firm *i* in year *t*

TABLE A2 Alternative innovation texts.

Alternative innovation text no.	Book	Chapters	Pages
1	Korres, G.M. (2012). *Handbook of innovation economics*. Nova Science Publishers.	1	1–43
2	Atkinson, R.D. & Ezell, S.J. (2012). *Innovation economics: the race for global advantage*. Yale University Press.	6	162–189
3	Link, A.N. & Siegel, D.S. (2007). *Innovation, entrepreneurship, and technological change*. Oxford University Press.	1–3	1–39
4	Mäder, A., Kunz, C., Ninck, A., Hurni, D. & Tokarski, K.O. (2015). Innovation management: In: Machado, C. & Paulo Davim, J. (Eds.) *Research and industry*. De Gruyter.	1	1–37
5	Talukder, M. (2014). *Managing innovation adoption: from innovation to implementation*. Taylor & Francis.	1	1–6