Essi Nousiainen

Essays on Corporate Textual Disclosure

ACTA WASAENSIA 552

University of Vaasa
VAASAN YLIOPISTO

ISBN 978-952-395-186-0 (print)
978-952-395-187-7 (online)

ISSN 0355-2667 (Acta Wasaensia 552, print)
2323-9123 (Acta Wasaensia 552, online)

URN https://urn.fi/URN:ISBN:978-952-395-187-7

PunaMusta Oy, Joensuu, 2025.

ACADEMIC DISSERTATION

To be presented, with the permission of the Board of the School of Accounting and Finance of the University of Vaasa, for public examination on the 4th of April, 2025, at noon.

Compilation dissertation, School of Accounting and Finance, Accounting.

Author	Essi Nousiainen https://orcid.org/0000-0002-3203-2723
Supervisor(s)	Associate Professor Mikko Ranta University of Vaasa. School of Accounting and Finance, Accounting. Professor Marko Järvenpää University of Vaasa. School of Accounting and Finance, Accounting. Dr. Mika Ylinen University of Vaasa. School of Accounting and Finance, Accounting.
Custos	Associate Professor Mikko Ranta University of Vaasa. School of Accounting and Finance, Accounting.
Reviewers	Professor Seppo Ikäheimo Aalto University, School of Business, Accounting. Professor Hannu Ojala University of Eastern Finland, Faculty of Social Sciences and Business Studies, Business School.
Opponent	Professor Seppo Ikäheimo Aalto University, School of Business, Accounting.

Tiivistelmä

Tämä on kokoomaväitöskirja, jonka tutkimuskohde on laskentatoimen tekstimuotoinen raportointi innovaatioiden, vastuullisuuden, sekä lohkoketjuteknologian aihepiireistä. Väitöskirja hyödyntää tekoälyavusteisia tekstianalyysimenetelmiä, kuten Latent Dirichlet Allocation (LDA) sekä tunneanalyysi, yhdysvaltalaisten yritysten laskentatoimen raporteissa (10-K, S-1) esiintyvän vapaaehtoisen raportoinnin tutkimiseen. Tulokset painottavat ei-taloudellisen raportoinnin tärkeyttä ja roolia myös laajemmassa kontekstissa. Tutkimus soveltaa legitimiteetti- ja signalointiteorioita selittääkseen yritysten intoa vapaaehtoiseen raportointiin huolimatta raportointikustannuksista ja lakisääteisten vaatimusten puutteesta.

Essee I tutkii innovaatioraportointia vaihtoehtoisena mittarina perinteisille patentteihin perustuville innovaatiomittauksille. Tutkimuksessa esitetään uusi tekstipohjainen lähestymistapa, joka tunnistaa yritysten innovatiivisuuden, eikä vaadi patenttiaineistoa. Essee II taas käsittelee vastuullisuusraportointia ja sen löydökset paljastavat, että ostajaa etsivät yritykset lisäävät vastuullisuusraportointiaan ilman vastaavaa parannusta vastuullisuussuoritteissa. Tutkimus korostaa vastuullisuusraportointia strategisena työkaluna yritysten legitimoinnissa myös silloin, kun suoritteet eivät kohtaa kerrotun kanssa. Essee III tutkii, miten lohkoketjuteknologia ilmenee listalleottoesitteissä. Tutkimuksen löydöksissä havaitaan siirtyminen varovaisempaan sävyyn raportoinnissa, erityisesti koskien kryptovaluutta-aihepiirejä. Tulokset heijastelevat yritysten haasteita liittyen markkinatilanteisiin ja sääntelyyn. Raportointi lohkoketjusovelluksista oli kuitenkin sävyltään aikaisempaa varmempaa suhteessa kryptovaluutta-aiheisiin liittyvinä löydöksinä.

Kokonaisuutena väitöskirjan lisäarvo on ei-taloudellisen raportoinnin tutkimuksessa laskentatoimen tieteenalalla. Väitöskirja luo uutta tietoa raportoinnin seurauksista yrityksille, yhteiskunnalle, sijoittajille sekä sidosryhmille. Lisäksi väitöskirja havainnollistaa läpinäkyvyyden tärkeyttä yritysten raportoinnissa sekä koneoppimis- ja tekoälyavusteisten tekstianalyysimenetelmien potentiaalia innovaatio-, vastuullisuus- ja teknologia-aiheiden yksityiskohtaisessa tutkimisessa. Sijoittajat ja sidosryhmät voivat hyötyä tutkimuksen tuomasta ymmärryksestä siirtymävaiheessa olevien yritysten tekstiraportointia kohtaan.

Asiasanat: Laskentatoimen raportti, 10-K, S-1, koneoppiminen, tekstianalyysi, vapaaehtoinen raportointi

Abstract

This is a compilation dissertation exploring textual accounting disclosures with a focus on innovation, ESG (environmental, social and governance) and blockchain technology disclosures. Using computational methods such as Latent Dirichlet Allocation (LDA) and sentiment analysis, the research analyzes voluntary disclosures in U.S. public companies' 10-K and S-1 filings. The study highlights the importance of non-financial disclosures in corporate reporting and their role in a broader context. The dissertation draws from legitimacy and signaling theories to explain why companies engage in voluntary disclosure despite its cost and the lack of statutory requirements.

In Essay I, innovation disclosure is explored as an alternative to traditional patent-based methods for measuring a company's innovative activities. The research proposes and validates a novel text-based approach that identifies innovation in firms and without relying on patents. Essay II investigates ESG disclosures, revealing that companies seeking acquisitions often increase their ESG-related disclosures without a corresponding improvement in their actual ESG performance. This highlights the strategic use of ESG disclosures as a tool to enhance market perceptions and legitimacy, even when the underlying performance does not align with the narrative. Blockchain disclosures are analyzed in Essay III to understand how emerging companies present their engagement with this technology during the IPO process. The research reveals a shift towards more cautious and uncertain language, especially in relation to cryptocurrencies, reflecting market and regulatory challenges. However, confidence in blockchain technology solutions has increased over time.

Overall, this dissertation contributes to the accounting literature on non-financial disclosure by providing new knowledge on textual disclosure and its implications for society, investors, stakeholders and the companies themselves. Furthermore, the research highlights the importance of transparency in voluntary disclosures and the potential of machine learning and natural language processing techniques to provide more nuanced assessments of innovation, ESG, and technology topics. Understanding the textual disclosure of companies in transitional phases can be insightful to investors and stakeholders.

Keywords: Accounting disclosure, 10-K, S-1, machine learning, textual analysis, voluntary disclosure

ESIPUHE

Aloitin väitöskirjaprojektini vuonna 2021 heti maisterin tutkinnon valmistuttua. Työskentelin silloin tutkimusavustajana ja olin jo haaveillut väitöskirjan teosta, kun lopulta Mikko Rannan ja Heidi Kuusniemen kannustuksesta uskalsin lähteä tähän projektiin. Matkan varrella on ollut haasteita, ja joskus olen miettinyt, että helpommallakin olisi voinut päästä, mutta tämä työ on antanut ja opettanut valtavasti, enkä vaihtaisi kokemusta mihinkään. Vaikka väitöskirjatyö on usein yksinäistä, on matkan varrella ollut paljon ihmisiä, joiden tuki on ollut korvaamatonta.

Aloin toden teolla perehtyä tekoälyavusteiseen tekstianalyysiin juurikin aiemmin mainitun tutkimusavustajan pestin myötä. Kiinnostuin aiheesta sen verran, että siitä tuli luonteva valinta väitöskirjan aiheeksi laskentatoimen raportoinnin kontekstissa. Aihe säilyi samana koko prosessin ajan, vaikkakin hiukan tarkentui matkan varrella. Väitöskirjan teon aikana olen useaan otteeseen kyseenalaistanut omaa kypsyyttäni ja osaamistani, mutta näin jälkikäteen katsottuna juuri nämä pohdinnat ovat olleet olennainen osa lopullista teosta. Sikäli kun väitöskirjan teko on tutkimustyön opettelua, niin vähintään yhtä paljon olen oppinut itsestäni.

En voi tarpeeksi kiittää ensimmäistä ohjaajaani, apulaisprofessori Mikko Rantaa. Kannustit minua aloittamaan väitöskirjan ja uskoit kykyihini. Olen oppinut sinulta valtavasti tutkimustyöstä ja olet antanut arvokkaita neuvoja ja tukea koko prosessin ajan. Lisäksi olet ollut loputtoman kärsivällinen virheistä ja vaikeuksista huolimatta. Kiitän myös muita ohjaajiani, professori Marko Järvenpäätä ja Mika Ylistä arvokkaasta yhteistyöstä, tuestanne, ja kaikesta, mitä olen teiltä oppinut erityisesti laskentatoimen tutkimusperinteestä ja teorioista. Lisäksi lämmin kiitos yhteystyöstä väitöskirjan toisessa esseessä apulaisprofessori Tatiana Kingille. On ollut suuri kunnia kirjoittaa artikkeleita teidän kanssanne. Kiitos myös kandi- ja graduohjaajalleni Jaana Rahkolle, jonka ohjauksessa otin ensiaskeleni tutkimustyön parissa.

Erityiskiitos myös professori Heidi Kuusniemelle, joka oli ensimmäisten joukossa kannustamassa minua tähän haasteeseen. Tukesi on ollut korvaamatonta. Kiitos, että otit minut mukaan Digital Economyyn, joka on ollut merkittävä tekijä työni ja väitöskirjani lopputuloksen kannalta. On ollut hienoa olla osa Digital Economy -tutkimusalustaa ja osallistua moniin tutkimushankkeisiin. Suuret kiitokset kaikille, joiden kanssa olen saanut tehdä yhteistyötä näissä hankkeissa – olen oppinut teiltä valtavasti.

Kiitän väitöskirjani esitarkastajaa, professori Hannu Ojalaa, sekä esitarkastajaa ja vastaväittäjää, Seppo Ikäheimoa, arvokkaasta palautteestanne, joka auttoi

kehittämään työtäni. Lisäksi olen kiitollinen työni rahoittajille, Evald ja Hilda Nissin säätiölle ja Liikesivistysrahastolle henkilökohtaisista apurahoista, joiden turvin sain väitöskirjani saatettua loppuun. Lisäksi kiitän kaikkia Vaasan yliopiston laskentatoimen yksikön jäseniä, joihin olen tutustunut väitöskirjaprojektin aikana.

Muistelen lämmöllä väitöskirjan kirjoittamisvuosissani erityisesti sitä, että olen saanut matkan varrelta hienoja ystäviä akateemisesta maailmasta. Kiitos Cem Özcan, Piia Korri ja Ly Pham vertaistuesta ja ystävyydestä. Tunnen itseni onnekkaaksi, koska minulla on niin monta hyvää ja pitkäaikaista ystävää myös työn ulkopuolella, jotka auttavat pitämään elämän tasapainossa. Erityisesti vaikeina hetkinä ystävien tuki on ollut korvaamatonta ja auttanut jaksamaan eteenpäin. Suurimmat kiitokset, Vilma, Eerika, Anni, Tiia, Clarissa ja kaikki muutkin ystäväni.

Perhe on ollut minulle aina tärkeä tukipilari. Olen kiitollinen vanhemmilleni Arjalle ja Tapanille kannustuksesta ja kärsivällisyydestä kaikkeen, mitä olen päättänyt tehdä elämässäni. Kiitän myös muita perheenjäseniäni ja kumppanini perhettä läheisyydestä ja tuesta. Kiitos erityisesti Anitalle ja Petrille sekä isovanhemmilleni Valmalle ja Jaakolle, joita pääsin Vaasan-vuosinani liian harvoin tapaamaan Pohjois-Karjalaan.

Suurimman kiitoksen ansaitsee kumppanini Otto, joka on ollut vierelläni koko väitöskirjaprosessin ajan. Olet jaksanut kannustaa ja tukea minua niin hyvinä kuin vaikeinakin hetkinä, ja ilman sinua tämä työ ei todellakaan olisi valmistunut. Toivon voivani olla sinulle yhtä suuri tuki.

Helsingissä 27.1.2025

Essi Nousiainen

Contents

Tables

Abbreviations

ML	Machine Learning
LDA	Latent Dirichlet allocation
ESG	Environmental, social and governance
M&A	Mergers and acquisitions
NLP	Natural language processing
CSR	Corporate social responsibility
MD&A	Manager's Discussion and Analysis

Essays

[1] Nousiainen, E., Ranta, M., Ylinen, M., & Järvenpää, M. (2024). Using Machine Learning and 10-K Filings to Measure Innovation. *Accounting & Finance, 64*(4), 3211-3239. https://doi.org/10.1111/ACFI.13245 CC BY-NC-ND. Reprinted with the kind permission of John Wiley & Sons.

[2] King, T; Nousiainen, E & Ranta, M. (2024). ESG Disclosure and ESG Performance of Seeking Buyer Companies [Revised and resubmitted]. An earlier version of this paper was presented at the 1st Botnia Accounting & Auditing Seminar, May 2022.

[3] Nousiainen, E. & Ranta, M. (2024). A Survey of Blockchain Disclosure by Firms in the Initial Public Offering (IPO) Process. [Unpublished manuscript]. An earlier version of this paper was presented at the National Accouting Tutorial, March 2024 and the 27th Nordic Academy of Management (NFF) Conference, August 2024.

Author’s Contribution

Essay 1. First author. Defining the research setting, collecting and managing research data, data analysis, research methods, verifying and analyzing the findings, writing of the first draft, editing the manuscript at various stages.

Essay 2. Co-author. Defining the research setting, collecting and managing research data, data analysis, research methods, verifying and analyzing the findings, writing of the first draft, editing the manuscript at various stages.

Essay 3. First author. Defining the research setting, collecting and managing research data, data analysis, research methods, verifying and analyzing the findings, visualizing the results, writing of the first draft, editing the manuscript at various stages.

1 INTRODUCTION

ESG, innovation, and technology disclosures are somewhat different types of disclosures, but are driven by a common goal of enhancing transparency and reducing information asymmetry. Innovation is an important driver for companies' long-term growth (Bellstam et al., 2020; Chang et al., 2015; Holmstrom, 1989), hence it has been researched in accounting literature on multiple occasions and remains an important research target (e.g. Chenhall et al., 2011; Huang et al., 2021; Tian & Wang, 2014). Additionally, environmental, social, and governance (ESG) issues have become important factors in investment decisions (Christensen et al., 2022), and adopting ESG practices has proven benefits for companies (e.g. Dhaliwal et al., 2011; Eliwa et al., 2021; Gregory et al., 2014; Li et al., 2018). Finally, blockchain technology is a potentially disruptive phenomenon that is entering traditional financial markets. Blockchain technology poses both challenges and opportunities to companies (Al Shamsi et al., 2023; Foley et al., 2019; Lee et al., 2024; Casey & Vigna, 2018), which creates an important ground for research into this technology whose impacts on society and individual companies remain underexplored.

This dissertation explores the role of textual accounting disclosures in 10-K and S-1 filings, with a focus on voluntary disclosure practices by public companies in the U.S. These filings, required by the SEC, include extensive narrative sections that provide detailed insights into a company's business and financial condition (Investor.gov, n.d.-a, n.d.-b; US Securities & Exchange Commission, 1934). Companies often use voluntary disclosures within these filings to convey additional important information beyond mandatory requirements, such as management's discussion and analysis (MD&A) and other narrative elements (Loughran & McDonald, 2016). A body of prior research has investigated voluntary disclosures, demonstrating that the textual sections in accounting reports provide valuable insights into the reporting company, and emphasizing that accounting information users should extend their attention beyond numeric financial statements. For instance, prior research has utilized these voluntary disclosures to predict future investments, detect financial misreporting, and measure financial constraints (Basu et al., 2022; Brown et al., 2020; Buehlmaier & Whited, 2018). Voluntary disclosures are also seen as a tool for enhancing transparency and building trust with stakeholders, which can positively influence market perceptions and investor decisions. Studies have shown that firms use voluntary disclosures, such as environmental, social and governance (ESG) reports, to respond to stakeholder pressures and regulatory expectations, aiming to improve their legitimacy and reputation (Deegan et al., 2002; Patten, 1992). Furthermore, it has been shown that companies disclose technology-related matters to send signals to the stakeholders regarding profitability (Liu et al., 2024), and that investors also react to technology disclosures (Nishant et al., 2017).

Following the importance of studying textual disclosure discussed above, the secondary aim of this dissertation is to apply computational textual analysis methods to address research problems related to accounting disclosure. Textual analysis as a research method has an opportunity to solve many research problems in the accounting field, particularly as accounting reports continue to grow in length and the volume of textual disclosure expands (Dyer et al., 2017). This trend creates new opportunities for researchers to examine the increasing numbers of disclosures. The benefit of textual analysis is that larger datasets can be gauged for underlying patterns, which would not be possible in a similar manner in manual, qualitative analysis. According to Loughran and McDonald (2016), textual analysis is an emerging area in accounting research. The aim of textual analysis is typically to computationally extract meaning from text, with the main categories being word lists, sentiment analysis, topic modeling, or measures of document similarity. The textual analysis applications used in the essays of this dissertation encompass most of these categories, with the main emphasis on topic modeling. Nevertheless, textual analysis is less precise compared to common econometric models, and there are many tripwires in analyzing company disclosures (Loughran & McDonald, 2016), which is why careful research design and a thorough understanding of the studied topic is required.

There is previous literature on innovation measures alternative to patent counts, including studies by Bellstam et al. (2020), Kogan et al. (2017), Mukherjee et al. (2017) and Cooper et al. (2020). However, these attempts of measuring innovation are still patent-derived (Kogan et al., 2017), limited to a specific type of innovation (Mukherjee et al., 2017), or require costly data (Bellstam et al., 2020). Recent research has called for alternative ways of measuring innovation (Lu & Chesbrough, 2022; Ranta et al., 2023), since the traditionally used patenting data can be difficult to acquire and not all companies file patents (Hall et al., 2013), in addition, companies may be in a situation where filing a patent is not in their best interest (Ciftci & Zhou, 2016; Saidi & Žaldokas, 2021). Nevertheless, innovation proxies are used in a variety of research papers and might also serve interest towards investors and company stakeholders, and thus, a proxy that would encompass all companies and overcome the shortfalls of previous attempts is necessary. The first essay in this dissertation answers these calls in previous literature by introducing a novel measure of innovation from easily accessible textual data, namely 10-K filings.

In addition to innovation, other elements can also be quantitatively measured from disclosure text. Essay II investigates ESG disclosure and performance of ‘seeking buyer' companies, and demonstrates a new way to measure ESG disclosure from 10-K filings with a word embedding model. Previous research has studied the earnings management behavior of seeking buyer companies (Anagnostopoulou & Tsekrekos, 2015), in addition, previous work on the role of ESG on the outcomes of M&A (Mergers and acquisitions) deals includes e.g. Aktas et al. (2011a), Bose et al. (2021), Deng et al. (2013), and Fairhurst and Greene (2022). However, there is no previous research on the ESG orientations of seeking buyer companies, i.e. a setting where the deal is specifically buyer-initiated. As the importance of ESG orientations in an M&A setting has been established in previous research, the second essay investigates its role in a situation where the target company has the opportunity to prepare for the upcoming acquisition.

Natural language processing (NLP) methods are useful for quantifying textual information, for example, when using textual measures in statistical models. The algorithms can, however, also be utilized for descriptive analysis to uncover the contents and trends in the disclosure. Essay III focuses on blockchain disclosure of companies in the IPO process. Prior studies on blockchain disclosure include Cheng et al. (2019) who studies speculative blockchain disclosure, Stratopoulos et al. (2022) who use blockchain disclosure to study the adoption rate of blockchain technologies, and Yen and Wang (2021) who study the value relevance of blockchain disclosures. Considering that investors react negatively to regulation in the cryptocurrency market (Chokor & Alfieri, 2021), and that Bitcoin is extensively used for criminal purposes (Foley et al., 2019) creating public controversy towards cryptocurrencies, it is valuable to understand how companies associated with cryptocurrencies and blockchain technologies view them. While previous research addresses blockchain disclosure in various contexts, it does not answer how blockchain applications are seen from the point of view of emerging companies and reflected in their business models and strategies.

To conclude, this dissertation explores the role of textual accounting disclosures, particularly in the context of ESG, innovation, and technology, emphasizing the use of computational textual analysis to enhance transparency and reduce information asymmetry. It consists of three essays. The first develops a novel text-based measure of innovation from 10-K filings, addressing the limitations of traditional patent-based methods. The second essay investigates the ESG practices of companies seeking acquisition, introducing a new method for measuring ESG disclosures through word embedding models. The third essay uses text mining to analyze blockchain disclosure trends in S-1 filings during the IPO process, providing insights into the evolving landscape of blockchain reporting. Overall, the dissertation provides novel insights into accounting literature utilizing computational textual analysis methods.

The structure of this chapter is as follows: Section 2 presents the contribution of the dissertation as a whole and the individual essays. Section 3 covers the background and theoretical foundations of this dissertation. Section 4 introduces the methodological considerations and data. Finally, Section 5 summarizes the essays that constitute this dissertation.

2 CONTRIBUTION

As a whole, this dissertation contributes to the accounting literature examining voluntary disclosure by providing new knowledge regarding the analysis of voluntary textual disclosure and its implications for society, investors and stakeholders. The work also advances accounting research by exploring applications of machine learning and natural language processing techniques, which are emerging in accounting research, but still remain less explored. Additionally, the contribution includes new methods to quantify textual information in accounting reports. More broadly, the essays introduce novel methods for measuring and quantifying different aspects within the text in financial reports.

The research argues for transparent and standardized disclosure. Previous research has questioned the credibility of voluntary disclosure (Michael & Dixon, 2019). This dissertation also argues that voluntary disclosure is often inconclusive and up for interpretation, and it may be hard to distinguish between symbolic and substantive disclosure. The findings on ESG and blockchain disclosures offer investors and stakeholders a nuanced understanding of corporate narratives, which can help them evaluate the substance of disclosures, allowing for more informed decision-making, particularly in the context of the studies in this dissertation. The methods in this dissertation used to analyze textual disclosures aid investors and stakeholders in assessing company strategy and performance. From a corporate communication point of view, firms can build trust with stakeholders and decrease information asymmetry by adopting more transparent reporting. Companies may use strategic disclosures of ESG, innovation or technology to enhance their reputation. Practitioners could also benefit from the methodological contributions of this dissertation, for example, the methods could be applied by managers to assess their competitors and potential customers. Also, the orientation of a given customer or competitor towards ESG, innovation, or blockchain could be of interest and help companies benchmark common practices or address challenges.

The first essay presents a method of innovation measurement drawn from 10-K filings, differing from previous attempts to construct an innovation measure alternative to patent counts. Therefore, this essay extends the research stream on alternative innovation measurements (Bellstam et al., 2020; Cooper et al., 2020; Kogan et al., 2017; Mukherjee et al., 2017). 10-K filing text has not been previously used for innovation measurement, and we demonstrate how the previously developed method of Bellstam et al. (2020) is not suitable for 10-K text data, even if effective on analyst reports, and present an improvement, simultaneously answering the call by Bellstam et al. to test their method on other text sources. Furthermore, the first essay offers a methodological contribution by developing a new procedure for utilizing Latent Dirichlet Allocation (LDA) in creating a measure of innovation that could be potentially applied to other disclosure topics as well. The outcomes of Essay I 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. In addition, practitioners could benefit from the innovation measurement method in assessing and benchmarking the practices and innovativeness of competitors and customers.

The second essay provides several valuable contributions to the accounting literature and M&A research. First, it builds on and extends existing research on seeking buyers (Anagnostopoulou & Tsekrekos, 2015) by examining ESG activities specifically in those firms, revealing that they engage in enhanced ESG disclosure but do not exhibit higher ESG performance compared to peers. Second, Essay II introduces a novel method for evaluating ESG information using a word embedding model following Li et al. (2021), allowing for a more nuanced analysis of ESG disclosures in 10-K forms. This approach answers calls for text-based measures made in prior literature (Ranta et al., 2023), and challenges traditional third-party ESG metrics by capturing a broad range of ESG-related language. Lastly, it provides evidence that enhanced ESG disclosure does not correlate with a higher likelihood of acquisition, emphasizing that financial performance indicators like leverage and sales growth are more critical to acquirers. The findings of Essay II suggest the need for standardized ESG reporting rules to ensure clarity and relevance in assessing ESG outcomes. The findings indicate that acquirers should investigate the ESG performance of target companies to verify that the disclosure reflects an actual adoption of ESG practices. Furthermore, the findings strengthen the view that ESG disclosure requires standardization to ensure the reliability of the information.

The third essay makes important contributions to the literature on blockchain and accounting disclosure (Cheng et al., 2019; Stratopoulos et al., 2022; Yen & Wang, 2021). First, it advances our understanding by analyzing the content of blockchain-related disclosures in emerging companies, revealing that these disclosures range from speculative buzzwords to established business models and products. Second, it examines the linguistic features of these disclosures over time, providing insights into emerging trends in blockchain disclosure sentiment, in addition to creating a methodology that could be applied to other disclosure themes. The study finds that the use of uncertain and weak modal words has increased in the course of time for the most of the topics, whereas strong modal words have decreased, which could indicate an overall increase in uncertainty. The findings imply that public controversy and a variety of risks and challenges are affecting the disclosure tone, and the study extends the conversations on blockchain controversies in existing literature (Chokor

& Alfieri, 2021; Foley et al., 2019). However, the findings for blockchain technology solutions indicate a more confident tone, which creates a distinct difference between blockchain solutions and cryptocurrencies. Lastly, the study sheds light on the implementation and adoption of blockchain technologies in business, supporting the findings of Cheng et al. (2019) that the existence of disclosure may not always reflect the actual integration of blockchain in business operations. Practical implications suggest that IPO investors should also scrutinize the disclosures and critically evaluate any references to blockchain technologies. Also, companies are seemingly adopting a more cautious approach toward cryptocurrencies, possibly due to market conditions or concerns about rising regulation.

To summarize, this dissertation contributes to the accounting literature on voluntary disclosure by introducing innovative methods and insights across three key areas. The first essay presents a novel measure of innovation derived from 10-K filings, advancing beyond traditional patent-based metrics and offering a new procedure using Latent Dirichlet Allocation. The second essay examines ESG activities in M&A settings, revealing that enhanced ESG disclosure in firms seeking acquisition does not correlate with higher ESG performance or increased acquisition likelihood, underscoring the need for standardized ESG reporting. The third essay explores blockchain disclosure, identifying trends in linguistic features that reflect increasing uncertainty in most topics, with the exception of blockchain solutions. Together, these essays enhance our understanding of voluntary disclosure practices and their implications for businesses, investors, and stakeholders, underscoring the importance of transparency.

3 BACKGROUND

This section presents the previous literature informing this dissertation. First, I introduce the concept of textual accounting disclosure and its treatment in prior accounting literature. Next, I discuss key theories related to voluntary disclosure, framing them as tools to better understand the phenomenon rather than as a unified theoretical lens for the dissertation. These theories, while not uniformly applied across all essays in this work, serve to contextualize the dissertation's examination of voluntary disclosure. Given the dissertation's focus on voluntary disclosure in the contexts of innovation, ESG, and blockchain or technology disclosures, I concentrate on these themes when discussing their relevance to the identified theories.

3.1 Textual accounting disclosure

This dissertation studies textual disclosures reported specifically in form 10-K and S-1 filings. Publicly reporting companies in the United States are required to submit an annual report in Form 10-K (Investor.gov, n.d.-a). The filing gives an overview of the company's business and contains financial statements. The form includes 15 items, some of which are not required by all firms (US Securities & Exchange Commission, 1934). The two items containing the most written narrative are Items 1. Business and 7. Management's Discussion and Analysis of Financial Condition and Results of Operations (MD&A), but important extensive disclosures can also appear in other parts of the filing, such as footnotes (Loughran & McDonald, 2016). Even though the required contents of the disclosures are strictly regulated, there is plenty of space for the reporting companies to include other important topics. Similar requirements apply for S-1 filings, except that the S-1 filing is the registration statement form filed when a company intends to go public, often containing the prospectus, which describes the offering and the company to potential investors (Investor.gov, n.d.-b).

Frazier et al. (1984) is one of the earliest studies on analyzing narrative accounting disclosure, where the authors suggest a methodology using a computer program to classify and statistically analyze the narrative sections of financial statements. Jones and Shoemaker (1994) and Lewis et al. (1986) also pioneer the analysis of accounting narratives, focusing on readability, which is nowadays considered a relatively simple form of analysis. Earlier works on accounting narratives in general are focused on small-sample manual analysis, with natural language processing and computational linguistics as relatively new phenomena (Li, 2010a). However, some examples of earlier studies using NLP or computational linguistics on accounting narratives exist, and include Feldman et al. (2010), Lehavy et al. (2011), Li (2008) and Li (2010b). For a comprehensive review of early literature on textual analysis of accounting disclosures, refer to Li (2010a).

Recent work on accounting narrative sections includes Basu et al. (2022) predicting future investments using identified keywords from 10-K text, Brown et al. (2020) using topic modeling to detect financial misreporting, and Buehlmaier and Whited (2018) who measure financial constraints from annual report text and study their impact on stock returns. Donovan et al. (2021) measure credit risk from the Management's discussion and analysis section in 10-K filings, and similarly, Frankel et al. (2016) also study MD&A sections to estimate accruals, whereas Dyer et al. (2017) study the evolution of 10-K textual disclosure trends with a topic model. Hoberg and Maksimovic (2015) study financial constraints from 10-K text, and Kim et al. (2019) use 10-K readability to predict stock price crash risk. S-1 filing disclosure has also been empirically investigated in previous works, including Huang et al. (2019), Loughran and McDonald (2013), Ma et al. (2021), Wasiuzzaman et al. (2018), and Chen et al. (2023). To sum up, quantitative disclosure research is commonly done by constructing a proxy for something from the text, and relating it to some other firm characteristic utilizing statistical modeling.

3.2 Legitimacy theory

One of the theories used in accounting research for explaining voluntary disclosure is legitimacy theory. Seeking legitimacy can explain companies engaging in voluntary disclosure. CSR disclosure is especially discussed in this context. Organizations require the support of the actors of the society around them to operate, and take different actions to ensure their legitimacy towards the society. Suchman (1995) defines legitimacy as "a generalized perception or assumption that the actions of an entity are desirable, proper, or appropriate within some socially constructed system of norms, values, beliefs, and definitions." To ensure their legitimacy, organizations should conform to the strategies and practices accepted by the general public and the regulators (Deephouse, 1996).

Legitimacy can be viewed as something controlled by the organization, or something that comes to the organization from the outside. Suchman (1995) finds two different streams of legitimacy research; those viewing legitimacy through a strategic approach as an organizational resource that firms exploit for their strategic goals, whereas those viewing legitimacy through an institutional approach see that the surrounding institutions shape the organization and vice versa, with less emphasis on managerial control.

The scientific literature recognizes different legitimization strategies, an organization can legitimate itself through three actions: by adapting its output, goals and means of operation, by trying to alter the definition of social legitimacy through communication to conform to its current practices, or by trying to become identified through communication with symbols, values or institutions that are seen as socially legitimate (Dowling & Pfeffer, 1975). Deephouse (1996) and DiMaggio and Powell (1983) also argue that organizations conforming to the same practices as others are seen as more legitimate by societal actors.

However, as Deephouse (1996) points out, there are power structures that may be biasing the legitimacy of the organization, as older companies are more likely to be higher in the up in the power hierarchy and "*be enforced by powerful social actors.*" And as Tilling and Tilt (2010) argue, organizations may be using voluntary reporting as a legitimacy tool to mask actions that are not aligned with the reported values. Also Palazzo and Scherer (2006) recognize the power hierarchies within the definition of organizational legitimacy, and propose a discursive approach.

When it comes to the downsides of organizations pursuing legitimacy, there are two different practices: substantive and symbolic (Ashforth & Gibbs, 1990). Substantive management consists of material changes in the organization's operations and goals that are inclined with society. Symbolic management, on the other hand, is when the organization is merely portraying their actions in a way that appears to be consistent with the societal values. The motivations behind legitimating through ESG disclosure have been discussed in a multitude of studies. Neu et al. (1998) question the role of environmental disclosures in annual reports, and whether their aim is to obfuscate the negative aspects. Lewis and Unerman (1999) discuss the role of ethical relativism in the different contents of CSR disclosure, as in what is seen as good or bad in a society is time and culture dependent. Overall, the motivations and practices of ESG and CSR disclosure from a legitimacy point of view have been widely discussed in accounting literature (i.e. Ball & Craig, 2010; Burritt & Schaltegger, 2010; Cho et al., 2012, 2015; Kolk et al., 2008).

Empirical evidence also supports the argument that firms are using ESG disclosures to gain legitimacy by responding to stakeholder pressures and regulatory requirements. For example, companies in the oil and gas industry often enhance their legitimacy through comprehensive environmental reporting following major incidents or unfavorable press (Deegan et al., 2002; Patten, 1992). There is also evidence that companies in their respective industries react to major social incidents by increasing social disclosure to legitimize their existence (Deegan et al., 2000). Furthermore, companies will also respond to environmental concerns raised by the media in their disclosures of environmental information (Brown & Deegan, 1998).

Hence, companies may be using ESG and CSR reporting for social legitimacy. Depending on the view, the companies may also be engaging in strategic legitimacy, and using CSR reporting to establish their legitimacy within the society, or from a different point of view, social norms towards CSR may be shaping their CSR dispositions.

3.3 Signaling theory

Signaling theory is a theory that is closely linked to legitimacy theory. Companies may try gain legitimacy by sending signals about their unobservable qualities to investors (Connelly et al., 2010), however, signaling also serves other purposes for companies, examined in this section. Akerlof (1970) laid the groundwork for signaling theory by discussing information asymmetry between the organization's insiders and outsiders through car sales. Signaling means that when two parties have different access to information, in a setting of manager and an investor, the manager knows more about the firm than the investor, and due to the uncertainty, the investor will offer a lower price, but the manager can voluntarily give up more information to increase the price offer (Connelly et al., 2010). More formally, one party must choose how to send the information (signal) to the other party, who will then choose how to interpret it. The early work of Spence (1973) describes signaling in the job market, where job applicants acquire degrees to signal their fit for a job and reduce information asymmetry. Effective signaling should be profitable only for the good candidate and challenging to mimic, with signaling costs negatively correlated with productivity.

In accounting research, signaling theory is connected to voluntary disclosure, and companies may be using ESG, technology, and innovation disclosures alike as signaling mechanisms. For example, they may disclose extensive ESG metrics or technology advancements, which require resources, transparency, and accountability. These actions serve as signals that differentiate them from bad companies, which might struggle to meet the same standards or bear similar costs. The effectiveness of a signal depends on its quality (e.g., accuracy, reliability), frequency, and interpretability, as noted by Connelly et al. (2010). For instance, in the work of Ross (1977), the disclosure of managerial incentives acted as a credible signal of a firm's financial stability. Similarly, Verrecchia (1983) emphasized how managers can use discretion to withhold or disclose information, thereby signaling the firm's risk profile. Zhang and Wiersema (2009) conclude that CEO's use financial statements to signal firms' unobservable (private) qualities to investors.

Other works on disclosure and signaling include Gelb (2000) who studied dividends and accounting disclosures as signaling mechanisms, finding that firms in industries with lower entry barriers see dividends and stock repurchases as less costly signaling mechanisms compared to accounting disclosures. Datt et al. (2019) found that US organizations signal better carbon performance with more voluntary carbon disclosure. Similarly, Sun et al. (2024) find that firms with a high ESG performance signal by having a more optimistic tone in ESG disclosure. However, they also find that ESG reporting principles constrain the manipulation of disclosure tones. It is worth noting that previous literature is not fully conclusive on whether environmental disclosure is a signal of better financial performance, a signal of better environmental performance, or something else.

Liu et al. (2024) explain technology disclosures with signaling theory and found that companies use digital supply chain announcements to signal profitability. Nishant et al. (2017) studied stock market reactions to information technology announcements from a signaling perspective. Thus, in addition to environmental disclosure, companies may use technology and innovation disclosure as a signaling tool to investors.

However, disclosure is not always the best choice for all organizations under all circumstances. Ciftci and Zhou (2016) find that disclosing innovations and patents can risk leaking strategic information to competitors, being more beneficial for firms in industries with strong intellectual property protection. Saidi and Žaldokas (2021) discuss the strategic factors of disclosure in a patent context. Even though Spence (1973) states that a signal should only be profitable for the good candidate, disclosure as a signaling mechanism can be fertile ground for manipulation and obfuscation by managers. In this vein, Lambertsen (2024) highlights how managers may exploit investors' understanding limitations, biasing signal interpretation.

3.4 Stakeholder theory and institutional theory

Further theories explaining voluntary disclosure include stakeholder theory and institutional theory. I examine these theories as alternative or complementary theories to legitimacy and signaling theories. Stakeholder theory posits that organizations are accountable beyond their shareholders. Thompson (1967) introduced the idea that an organization must manage its external environment and interact with external groups. Freeman (1984) defines that: *“A stakeholder in an organization is (by definition) any group or individual who can affect or is affected by the achievement of the organization's objectives.”* Organizational success requires commitment towards the stakeholders (Moneva et al., 2007), with stakeholder disclosure being one method of engagement. For example, the GRI standards are sustainability reporting standards focused on stakeholder needs (GRI, 2021). Prior works applying stakeholder theory on CSR disclosure dialogue include Burchell and Cook (2006) who study the impact of CSR dialogue processes, and Kaur and Lodhia (2018) who study stakeholder engagement practices throughout the sustainability reporting process.

Institutional theory explains how isomorphic processes push organizations to become increasingly similar to maintain legitimacy (DiMaggio & Powell, 1983). Conforming to institutional norms may explain CSR practice adoption (Jennings & Zandbergen, 1995). Lammers and Barbour (2006) suggest an institutional theory for organizational communication, part of which includes accounting disclosure, stating that organizational communication sustains institutions and aligns organizations with them. It is also noted that successful external communication often references institutional rules and norms.

4 METHODOLOGY

4.1 Data and methodology

The methodologies of the essays in this dissertation are itemized in Table 1. All of the essays use quantitative methodologies, machine learning, and statistical modeling, in addition to textual analysis. According to Lukka (2010), the economics-based research agenda with large archival datasets is the dominant paradigm in accounting research, which this thesis also falls into. This dissertation also mainly falls into the category of positivistic accounting research according to Baker's (2011) definition of *"research following the scientific model, in which several hypotheses were proposed, numerical data were collected, and statistical analyses were performed to test the hypotheses",* with Essay III falling more into the category of descriptive accounting research (Baker, 2011).

Table 1. Data sources and research design

	Essay I	Essay II	Essay III
Research type	Quantitative	Quantitative	Quantitative
Methods	LDA Statistical modeling Qualitative robustness check	Word embedding Statistical modeling	LDA Sentiment analysis Statistical modeling for time trends
Data sources	10-K filings Patent data (Kogan et al., 2017) LSEG Eikon company fundamentals	10-K filings Seeking buyer and completed deals data from LSEG Eikon Deal Screener ESG data from LSEG Eikon and Bloomberg LSEG Eikon company fundamentals	S-1 filings Sentiment word lists (Loughran et al., 2011)

The data source encompassing all of the essays in this dissertation is textual data in the form of company filings. The additional variables for statistical models are collected from the LSEG Eikon database, Bloomberg database, and Noah Stoffman's patent database (Kogan et al., 2017), and Loughran et al. (2011) sentiment word lists.

4.2 Data preprocessing

To reduce noise and the number of dimensions, textual data often requires preprocessing prior to being fed to the algorithm, depending on the ML algorithm. Pre-processing helps infer more meaningful results from the text mining model. The pre-processing steps used in the essays include e.g. 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. The dictionary is also often limited so that the most common and the rarest words are deleted, because if the dictionary contains words that appear only in a single document, the word is not informative of the trends of the whole corpus, and the same applies for words that appear in every document.

4.3 LDA for textual disclosure analysis

Latent Dirichlet Allocation (LDA) is a machine learning algorithm that I use in Essays I and III to analyze textual information in company filings. LDA is a common machine learning algorithm for analyzing accounting reports (Blei et al., 2003), and was the most practical choice for the analytical purposes of Essays I and III. LDA is designed to uncover latent thematic structures within textual datasets by assuming that documents are mixtures of topics, and that topics are mixtures of words. Through iterative processing, LDA probabilistically assigns words to topics, revealing the underlying topical composition of documents. This approach allows for the identification of established themes within the dataset, enhancing our understanding of the latent semantic structures that characterize the disclosures.

The application areas of LDA in accounting cover a wide range of purposes and topics, making it highly a versatile algorithm. For example, LDA can be used as a tool for literature review, as identified by Garanina et al. (2022) and Cai et al. (2019), constructing various measures from text (Bellstam et al., 2020), or studying the topical content of company disclosures (Dyer et al., 2017; Gao et al., 2020; Stratopoulos et al., 2022). Essay I utilizes LDA in a similar manner to Bellstam et al. (2020) by using LDA to construct a text-based measure on innovation, whereas Essay III utilizes a similar methodology to Gao et al. (2020), who use LDA to study cybersecurity risk disclosures, but Essay III applies the methods to blockchain disclosure. As is demonstrated by the very different use cases in the essays and prior literature, LDA enables a wide range of application areas.

4.4 Disclosure sentiment analysis

Sentiment analysis in accounting is often based on the word lists of Loughran and McDonald (2011), which are specific to the financial context. Loughran and McDonald create financial word lists for six sentiment types (negative, positive, uncertainty, litigious, strong modal, and weak modal) that are used in Essay III of this dissertation. Sentiment analysis is especially useful for understanding the tones of textual content, which is why it is used in Essay III for uncovering insights on blockchain disclosure. Previously, sentiment has been studied in the context of either positive or negative tone (Feldman et al., 2010; Li, 2008; Tetlock et al., 2008), but the "new" financial text-specific word lists significantly expand the possibilities of sentiment analysis. Particularly, while sentiment is usually calculated as a count or share of sentiment words in the full text unit, recent research advances include a more efficient ML-based sentiment analysis for positive and negative tone (Frankel et al., 2021).

Sentiment has been connected with various company-related features. Examples of research employing sentiment types beyond positive and negative include Li et al. (2022) who discovered that firms engaging in earnings management tend to use more positive and modal words, Ertugrul et al. (2017) who found a correlation between the use of uncertain or weak modal words and stricter loan terms, as well as a higher risk of future stock price crashes, and Patelli and Pedrini (2013) who found that optimism in CEO letters was associated with better past and future performance.

4.5 Word embedding models in accounting

Li et al. (2021) developed a methodology for measuring corporate culture using a word embedding model. Their approach can also be used for measuring other textual themes, such as ESG. Word embedding models presented in Mikolov et al. (2013a) and Mikolov et al. (2013b) can be utilized for creating dictionaries in an objective manner by representing words as continuous vectors in a high-dimensional space, capturing semantic relationships between words based on their context in large text corpora. These models, specifically Word2Vec, learn word representations by predicting a word given its surrounding context or by predicting the context given a word. The resulting word embeddings can efficiently capture semantic similarities, such as grouping together synonyms or words with related meanings. This approach is useful for finding and organizing semantically related words, enabling the development of more accurate and contextually aware dictionaries – in other words, it can be utilized for measuring a specific type of language in text, since it can find contextually similar words. Therefore, Essay II of this dissertation utilizes the word embedding models and methodology specified in Mikolov et al. (2013a), Mikolov et al. (2013b), and Li et al. (2021) to measure ESG disclosure.

5 SUMMARY OF THE ESSAYS

This section summarizes the three essays studying textual and other types of company disclosure from the point of view of voluntary innovation, ESG and blockchain disclosure, utilizing machine learning and text mining techniques.

5.1 Essay I: Using Machine Learning and 10-K Filings to Measure Innovation

This essay studies innovation disclosure in public U.S. companies' 10-K filings. In this paper we propose a new method of measuring innovation using 10-K filings and a machine learning algorithm called Latent Dirichlet Allocation (LDA). The 10-K filing includes information about the company's strategy and activities, and should presumably be informative regarding the company's innovative activities. The limitations of traditional innovation measurement proxies have been discussed in accounting literature – namely, not all companies file patents (Hall et al., 2013); the filed patents are context- and industry-specific (Guo et al., 2019); and filing a patent may not be in a company's best interest under some circumstances (Ciftci & Zhou, 2016; Saidi & Žaldokas, 2021). These findings in previous research call for alternative ways of measuring innovation in a reliable manner. Previous studies have successfully built text-based measures of innovation (Bellstam et al., 2020; Lu & Chesbrough, 2022), and have also used other novel methodologies for innovation measurement (Cooper et al., 2020; Kogan et al., 2017; Mukherjee et al., 2017).

We use unsupervised ML to construct the innovation measure, specifically Latent Dirichlet Allocation (LDA), due to its capacity to analyze text without predefined response variables. Multiple LDA models are trained with varied topic numbers to gauge method sensitivity. Two alternative innovation measurement methods are proposed based on LDA outputs: the topic weight method and the topic distribution method. The former identifies an "innovation topic" by comparing word distributions with a designated innovation textbook, while the latter assesses divergence between topic distributions of annual reports and the innovation textbook. These measures are validated against patent-based indicators and firm performance variables using panel regressions. For data and sample, US companies' annual reports are utilized as a textual innovation source, alongside accounting and patent data. The sample comprises SEC 10-K filings from 2008 to 2018, aiming for cross-industry representation to ensure metric applicability.

Panel regression models controlling for industry- and year-fixed effects and company characteristics with logged patent and citation variables as the response variables are employed to study the association of the text-based innovation variables and the patent variables. The analysis reveals that the topic distribution method consistently demonstrates a strong association with patenting, with statistically significant and positive coefficients across various models. However, the topic weight method exhibits less consistency, displaying both positive and negative coefficients, indicating sensitivity to initial specifications. Further analysis focuses on a subsample of patenting firms, with similar results observed. Additionally, the association between text-based innovation measures and firm performance is explored. The topic distribution method consistently outperforms the patent count proxy in predicting firm performance, while the topic weight method displays somewhat inconsistent results, particularly with Tobin's Q.

Additional robustness tests on the topic distribution method confirm that the topic distribution method is a reliable innovation metric, assigning notably better scores for software industry firms despite their low patent filings. In addition, through qualitative inspection of non-patenting firms' annual reports, we find that the text-based metric accurately identifies innovative activities that are not patent-based, highlighting its ability to measure innovation more broadly. Finally, we explore the impact of alternative sources of innovation text on the topic distribution measure and find that the innovation measure varies slightly depending on the text source, indicating the sensitivity of the method to the chosen reference texts.

The results of this study emphasize the importance of careful consideration when incorporating text-based measures for abstract concepts like innovation, as both the LDA configuration and the chosen reference texts influence the metric's performance and reliability. The essay suggests 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 to their stakeholders without compromising their trade secrets by filing patents. In addition, while extending beyond patent-based innovation, the measure is able to predict future patents and citations, which also has practical use cases.

5.2 Essay II: ESG Disclosure and ESG Performance of Seeking Buyer Companies

Involvement in environmental, social, and governance (ESG) activities has gained importance in investment decisions and company valuations (Christensen et al., 2022), with significant impacts on financial performance (Li et al., 2018; Nekhili et al., 2021), cost of capital (Dhaliwal et al., 2011), competitive advantage (Russo & Fouts, 1997), and stakeholder relationships (Tregidga & Laine, 2022). Research has primarily focused on the effects of ESG on mergers and acquisitions (M&A) showing benefits such as better announcement returns and higher takeover premiums (Aktas et al., 2011b; Bose et al., 2021; Deng et al., 2013; Fairhurst & Greene, 2022). However, there is limited exploration of ESG activities in target-initiated M&A deals. Previous research on target-initiated M&A deals is limited to the role of earnings management in the process (Anagnostopoulou & Tsekrekos, 2015). This second essay examines whether firms seeking buyers enhance their ESG activities to attract potential buyers, based on legitimacy and signaling theories.

In mergers and acquisitions, target firms seeking buyers may publicly announce their intentions for i.e. strategic reasons (Aktas et al., 2011b). These firms often engage in earnings management to appear more attractive, similar to companies preparing for IPOs, which can lead to negative long-term outcomes (Alhadab et al., 2015). ESG activities are used to build positive impressions (Neu et al., 1998), reduce cost of capital (Dhaliwal et al., 2011, 2012), and increase firm value (Buchanan et al., 2018; Cho et al., 2012), signaling compliance with societal expectations and reducing information asymmetry (Cui et al., 2018). However, excessive ESG disclosure without genuine performance can raise skepticism and deter potential buyers. According to a stream of previous research, stakeholders may see ESG activities also unfavorable and a waste of resources (see Masulis & Reza, 2015). The research hypotheses of the second essay consider whether seeking buyer firms are characterized by higher ESG disclosure and performance prior to the announcement date, and whether their ESG activities have a relationship with the likelihood of being acquired.

The second essay studies the ESG practices of seeking buyer companies with the help of three different ESG proxies. The first proxy is the Bloomberg ESG disclosure score, based on publicly available reports and sources, which measures the extent of disclosed ESG information. Recognizing the limitations of this metric, the study also uses a machine learning-based measure derived from textual analysis of 10-K forms to objectively assess ESG disclosure. Additionally, the LSEG ASSET4 rating is used to evaluate actual ESG performance, considering both qualitative and quantitative data. The sample consists of publicly listed US firms seeking buyers between 2000 and 2021, with data sourced from the LSEG Eikon M&A database and the SEC Edgar database, resulting in 246 deals from various industry sectors.

The findings indicate that seeking buyer companies are involved in a higher ESG disclosure measured by the ML ESG disclosure proxy. However, the findings indicate that their ESG performance does not differ compared to their peers. In addition, by examining an interaction term for ESG disclosure and performance, we find that a high ESG disclosure and performance are not likely to appear within the same companies, which supports the finding that seeking buyer companies are high disclosure – low performance. Regarding the effect of seeking buyer ESG practices on the likelihood of being acquired, we do not find a statistically significant relationship. Our additional analyses and robustness test confirm the positive relationship between ESG disclosure and the propensity of being a seeking buyer company.

Existing literature on the relationship between ESG activities and M&A outcomes is limited, focusing mainly on deals initiated by acquiring companies, examining factors like announcement returns, takeover premiums, and acquisition decisions. The second essay extends the literature by exploring ESG activities in firms seeking buyers, finding that these firms disclose more ESG information but do not differ in ESG performance compared to peers, suggesting this disclosure may be superficial. Enhanced ESG practices do not correlate with a higher likelihood of acquisition, indicating that acquiring firms prioritize financial characteristics like leverage and sales growth over ESG activities. The study contributes to understanding corporate disclosures by firms seeking buyers, and suggests practical implications for acquiring firms to verify the genuine integration of ESG strategies. It aligns with regulatory efforts to standardize ESG reporting and highlights the need for more comprehensive research across different markets and industries.

5.3 Essay III: A Survey of Blockchain Disclosure by Firms in the Initial Public Offering (IPO) Process

The subject of this Essay III is at the intersection of emerging technologies and traditional markets. The study focuses on newly listing companies involved with blockchain technologies. By analyzing S-1 filings of firms in the IPO process, the research employs advanced text mining techniques to scrutinize the content and sentiment of blockchain disclosures over time, offering a deeper understanding of underlying themes and patterns. Building on prior studies (Cheng et al., 2019; Stratopoulos et al., 2022), the findings reveal how emerging companies are integrating blockchain into their business models and how the disclosure strategies change over time, affected by global events and market circumstances.

This study concerns IPO disclosure. Previous work that this study builds on, has studied the impact of IPO disclosure on IPO returns, misleading information, and risk factors (Loughran & McDonald, 2013; Ma et al., 2021; Wasiuzzaman et al., 2018). Prior work on blockchain disclosure finds a shift from cryptocurrency-focused content to broader business applications as firms recognize the technology's potential to enhance transparency, security, and efficiency (Cheng et al., 2019; Stratopoulos et al., 2022). However, the motivation behind engaging in blockchain disclosure may be technological opportunism, which according to earlier findings, may increase profitability or help gain competitive advantage (Sarkees, 2011; Srinivasan et al., 2002).

The accounting literature suggests that the sentiment and tone in corporate disclosures, including those related to blockchain, carry significant informational value. For example, Tsang et al. (2023) found that risk disclosure sentiments impact investor responses, while González et al. (2019) linked uncertain sentiment in IPO prospectuses to IPO underpricing. In blockchain disclosures, the use of uncertain or strong modal words may signal a firm's uncertainty about blockchain technologies. Firms might manipulate the tone of their disclosures to shape perceptions, as seen in Li et al. (2022) where positive language was used to obscure earnings management. Ertugrul et al. (2017) also found that uncertain language correlated with stricter loan terms and higher risks of stock price crashes. Conversely, Patelli and Pedrini (2013) associated optimistic CEO letters with better performance. Thus, the tone of blockchain disclosures can reflect a firm's genuine optimism or a strategic intent to influence readers.

The data used in this study comprises of 712 S-1 filings containing at least one of our blockchain keywords. Paragraphs discussing blockchain topics are separated from the filings, resulting in 18,455 individual paragraphs. A 5-topic LDA model is then trained on the paragraph corpus and the Loughran et al. (2011) sentiment word counts are calculated for each paragraph. These are then combined with a regression equation to analyze the changes in each sentiment over time. Five blockchain disclosure topics emerge from the corpus. The topics are labeled Risk Factors, General Business, Cryptocurrency Mining, Cryptocurrency Trading & Investment, and Blockchain Technology solutions, and each paragraph is assigned one topic based on its importance in the document.

The analysis of blockchain-related S-1 disclosures reveals notable shifts in sentiment and language across different topics and industries between 2016 and 2022. In general, there has been an increase in uncertain and weak modal language, particularly in topics related to risk factors, general financial issues, mining, and trading/investment, reflecting growing cautiousness and uncertainty within companies with regard to blockchain. This trend contrasts with the blockchain solutions topic, which saw a decrease in negative and uncertain language, suggesting increased confidence in these technologies. Industry-specific analysis shows that the Finance, Insurance, and Real Estate sector, along with Services and Retail, experienced significant increases in uncertainty. These findings suggest that while blockchain technology, particularly cryptocurrency, initially generated enthusiasm, emerging challenges and regulatory pressures have possibly led to a more cautious and uncertain communication approach, especially in sectors somewhat involved with digital assets.

To conclude, the study reveals a significant increase in uncertain and cautious language across the majority of blockchain-related S-1 topics between 2016 and 2022, particularly in cryptocurrency disclosures, signaling growing apprehension rather than the expected decrease in uncertainty as the technology matures. This trend suggests that companies may be facing financial challenges, adjusting to regulatory pressures, or are expressing general disappointment in the applications. The only exception is the Blockchain Solutions topic, where language became more confident. The research, which also highlights how blockchain disclosure varies across industries, offers new insights into the trends and waves of enthusiasm surrounding blockchain technologies, and the strategic positioning of firms in the IPO process. Our findings suggest that heightened uncertainty in the cryptocurrency disclosure signals an increase in risk, and the investors should carefully scrutinize IPO disclosures, particularly regarding cryptocurrencies, as companies are seemingly adopting a more cautious approach reflecting market conditions.

References

Akerlof, G. A. (1970). The market for "lemons": Quality uncertainty and the market mechanism. *Quarterly Journal of Economics*, *84*(3), 488–500. https://doi.org/10.2307/1879431

Aktas, N., de Bodt, E., & Cousin, J. G. (2011a). Do financial markets care about SRI? Evidence from mergers and acquisitions. *Journal of Banking & Finance*, *35*(7), 1753–1761. https://doi.org/10.1016/J.JBANKFIN.2010.12.006

Aktas, N., de Bodt, E., & Cousin, J. G. (2011b). Do financial markets care about SRI? Evidence from mergers and acquisitions. *Journal of Banking & Finance*, *35*(7), 1753–1761. https://doi.org/10.1016/J.JBANKFIN.2010.12.006

Alhadab, M., Clacher, I., & Keasey, K. (2015). Real and accrual earnings management and IPO failure risk. *Accounting and Business Research*, *45*(1), 55–92. https://doi.org/10.1080/00014788.2014.969187

Al Shamsi, M., Smith, D., & Gleason, K. (2023). Space transition and the vulnerabilities of the NFT market to financial crime. *Journal of Financial Crime*, *30*(6), 1664–1673. https://doi.org/10.1108/JFC-09-2022-0218/FULL/PDF

Anagnostopoulou, S. C., & Tsekrekos, A. E. (2015). Earnings management in firms seeking to be acquired. *The British Accounting Review*, *47*(4), 351–375. https://doi.org/10.1016/J.BAR.2014.07.001

Ashforth, B. E., & Gibbs, B. W. (1990). The Double-Edge of Organizational Legitimation. *Organization Science*, *1*(2), 177–194. https://doi.org/10.1287/ORSC.1.2.177

Baker, C. R. (2011). A genealogical history of positivist and critical accounting research. *Accounting History*, *16*(2), 207–221. https://doi.org/10.1177/1032373210396335

Ball, A., & Craig, R. (2010). Using neo-institutionalism to advance social and environmental accounting. *Critical Perspectives on Accounting*, *21*(4), 283–293. https://doi.org/10.1016/J.CPA.2009.11.006

Basu, S., Ma, X., & Briscoe-Tran, H. (2022). Measuring Multidimensional Investment Opportunity Sets with 10-K Text. *ACCOUNTING REVIEW*, *97*(1), 51–73. https://doi.org/10.2308/TAR-2019-0110

Bellstam, G., Bhagat, S., & Cookson, J. A. (2020). A Text-Based Analysis of Corporate Innovation. *Management Science, 67*(7), 4004–4031. https://doi.org/10.1287/MNSC.2020.3682

Bose, S., Minnick, K., & Shams, S. (2021). Does carbon risk matter for corporate acquisition decisions? *Journal of Corporate Finance*, *70*, 102058. https://doi.org/10.1016/J.JCORPFIN.2021.102058

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. https://doi.org/10.1111/1475-679X.12294

Brown, N., & Deegan, C. (1998). The public disclosure of environmental performance information—a dual test of media agenda setting theory and legitimacy theory. *Accounting and Business Research*, *29*(1), 21–41. https://doi.org/10.1080/00014788.1998.9729564

Buchanan, B., Cao, C. X., & Chen, C. (2018). Corporate social responsibility, firm value, and influential institutional ownership. *Journal of Corporate Finance*, *52*, 73–95. https://doi.org/10.1016/J.JCORPFIN.2018.07.004

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. https://doi.org/10.1093/RFS/HHY007

Burchell, J., & Cook, J. (2006). It's good to talk? Examining attitudes towards corporate social responsibility dialogue and engagement processes. *Business Ethics: A European Review*, *15*(2), 154–170. https://doi.org/10.1111/J.1467-8608.2006.00439.X

Burritt, R. L., & Schaltegger, S. (2010). Sustainability accounting and reporting: Fad or trend? *Accounting, Auditing & Accountability Journal*, *23*(7), 829–846. https://doi.org/10.1108/09513571011080144

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*(4), 709–733. https://doi.org/10.1111/ABAC.12179

Chang, X., Fu, K., Low, A., & Zhang, W. (2015). Non-executive employee stock options and corporate innovation. *Journal of Financial Economics*, *115*(1), 168–188. https://doi.org/10.1016/J.JFINECO.2014.09.002

Cheng, S. F., De Franco, G., Jiang, H., & Lin, P. (2019). Riding the blockchain mania: Public firms' speculative 8-K disclosures. *Management Science*, *65*(12), 5901. https://doi.org/10.1287/mnsc.2019.3357

Chenhall, R. H., Kallunki, J.-P., & Silvola, H. (2011). Exploring the relationships between strategy, innovation, and management control systems: The roles of social networking, organic innovative culture, and formal controls. *Journal of Management Accounting Research*, *23*(1), 99–128. https://doi.org/10.2308/jmar-10069

Chen, J. W., Khoo, E. S., & Peng, Z. (2023). Climate change disclosure and the information environment in the initial public offering market. *Accounting & Finance*, *63*(S1), 907–952. https://doi.org/https://doi.org/10.1111/acfi.13085

Cho, C. H., Guidry, R. P., Hageman, A. M., & Patten, D. M. (2012). Do actions speak louder than words? An empirical investigation of corporate environmental reputation. *Accounting, Organizations and Society*, *37*(1), 14–25. https://doi.org/10.1016/J.AOS.2011.12.001

Cho, C. H., Laine, M., Roberts, R. W., & Rodrigue, M. (2015). Organized hypocrisy, organizational façades, and sustainability reporting. *Accounting, Organizations and Society*, *40*, 78–94. https://doi.org/10.1016/J.AOS.2014.12.003

Chokor, A., & Alfieri, E. (2021). Long and short-term impacts of regulation in the cryptocurrency market. *The Quarterly Review of Economics and Finance*, *81*, 157–173. https://doi.org/10.1016/J.QREF.2021.05.005

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is Corporate Virtue in the Eye of The Beholder? The Case of ESG Ratings. *Accounting Review*, *97*(1), 147–175. https://doi.org/10.2308/TAR-2019-0506

Ciftci, M., & Zhou, N. (2016). Capitalizing R&D expenses versus disclosing intangible information. *Review of Quantitative Finance and Accounting*, *46*(3), 661–689. https://doi.org/10.1007/S11156-014-0482-0/TABLES/11

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2010). Signaling Theory: A Review and Assessment. *Journal of Management*, *37*(1), 39–67. https://doi.org/10.1177/0149206310388419

Cooper, M. J., Knott, A. M., & Yang, W. (2020). Implications of Innovation Measurement. In *SSRN*. https://doi.org/10.2139/ssrn.2631655

Cui, J., Jo, H., & Na, H. (2018). Does Corporate Social Responsibility Affect Information Asymmetry? *Journal of Business Ethics*, *148*(3), 549–572. https://doi.org/10.1007/S10551-015-3003-8

Datt, R. R., Luo, L., & Tang, Q. (2019). Corporate voluntary carbon disclosure strategy and carbon performance in the USA. *Accounting Research Journal*, *32*(3), 417–435. https://doi.org/10.1108/ARJ-02-2017-0031/FULL/PDF

Deegan, C., Rankin, M., & Tobin, J. (2002). An examination of the corporate social and environmental disclosures of BHP from 1983-1997: A test of legitimacy theory. *Accounting, Auditing & Accountability Journal*, *15*(3), 312–343. https://doi.org/10.1108/09513570210435861/FULL/PDF

Deegan, C., Rankin, M., & Voght, P. (2000). Firms' disclosure reactions to major social incidents: Australian evidence. *Accounting Forum*, *24*(1), 101–130. https://doi.org/10.1111/1467-6303.00031

Deephouse, D. L. (1996). Does isomorphism legitimate? *Academy of Management Journal*, *39*(4), 1024–1039. https://doi.org/10.2307/256722

Deng, X., Kang, J. koo, & Low, B. S. (2013). Corporate social responsibility and stakeholder value maximization: Evidence from mergers. *Journal of Financial Economics*, *110*(1), 87–109. https://doi.org/10.1016/J.JFINECO.2013.04.014

Dhaliwal, D. S., Li, O. Z., Tsang, A., & Yang, Y. G. (2011). Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting. *Accounting Review*, *86*(1), 59–100. https://doi.org/10.2308/ACCR.00000005

Dhaliwal, D. S., Radhakrishnan, S., Tsang, A., & Yang, Y. G. (2012). Nonfinancial Disclosure and Analyst Forecast Accuracy: International Evidence on Corporate Social Responsibility Disclosure. *The Accounting Review*, *87*(3). https://doi.org/10.2308/accr-10218

DiMaggio, P. J., & Powell, W. W. (1983). The Iron Cage Revisited: Institutional Isomorphism and Collective Rationality in Organizational Fields. *American Sociological Review*, *48*(2), 147. https://doi.org/10.2307/2095101

Donovan, J., Jennings, J., Koharki, K., & Lee, J. (2021). Measuring credit risk using qualitative disclosure. *Review of Accounting Studies*, *26*(2), 815–863. https://doi.org/10.1007/s11142-020-09575-4

Dowling, J., & Pfeffer, J. (1975). Organizational Legitimacy: Social Values and Organizational Behavior. *The Pacific Sociological Review*, *18*(1), 122–136. https://doi.org/10.2307/1388226

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. https://doi.org/10.1016/j.jacceco.2017.07.002

Eliwa, Y., Aboud, A., & Saleh, A. (2021). ESG practices and the cost of debt: Evidence from EU countries. *Critical Perspectives on Accounting*, *79*, 102097. https://doi.org/10.1016/J.CPA.2019.102097

Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing. *Journal of Financial and Quantitative Analysis*, *52*(2), 811–836. https://doi.org/DOI: 10.1017/S0022109017000187

Fairhurst, D. (DJ), & Greene, D. T. (2022). Too much of a good thing? Corporate social responsibility and the takeover market. *Journal of Corporate Finance*, *73*, 102172. https://doi.org/10.1016/J.JCORPFIN.2022.102172

Feldman, R., Govindaraj, S., Livnat, J., & Segal, B. (2010). Management's tone change, post earnings announcement drift and accruals. *Review of Accounting Studies*, *15*(4), 915–953. https://doi.org/10.1007/S11142-009-9111-X/TABLES/12

Foley, S., Karlsen, J. R., & Putnins, T. J. (2019). Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies? *The Review of Financial Studies*, *32*(5), 1798–1853. https://doi.org/10.1093/RFS/HHZ015

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. https://doi.org/10.1016/J.JACCECO.2016.07.003

Frankel, R., Jennings, J., & Lee, J. (2021). Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science*, *68*(7), 5514–5532. https://doi.org/10.1287/MNSC.2021.4156

Frazier, K. B., Ingram, R. W., & Tennyson, B. M. (1984). A Methodology for the Analysis of Narrative Accounting Disclosures. *Journal of Accounting Research*, *22*(1), 318. https://doi.org/10.2307/2490713

Freeman, R. E. (1984). *Strategic management: a stakeholder approach*. Pitman.

Gao, L., Calderon, T. G., & Tang, F. (2020). Public companies' cybersecurity risk disclosures. *International Journal of Accounting Information Systems*, *38*, 100468. https://doi.org/10.1016/J.ACCINF.2020.100468

Garanina, T., Ranta, M., & Dumay, J. (2022). Blockchain in accounting research: current trends and emerging topics. *Accounting, Auditing and Accountability Journal*, *35*(7), 1507–1533. https://doi.org/10.1108/AAAJ-10-2020-4991

Gelb, D. S. (2000). Corporate Signaling with Dividends, Stock Repurchases, and Accounting Disclosures: An Empirical Study. *Journal of Accounting, Auditing and Finance, 15*(2), 99–120. https://doi.org/10.1177/0148558X0001500201

González, M., Guzmán, A., Tellez-Falla, D. F., & Trujillo, M. A. (2019). Governance, sentiment analysis, and initial public offering underpricing. *Corporate Governance: An International Review*, *27*(3), 226–244. https://doi.org/10.1111/CORG.12272

Gregory, A., Tharyan, R., & Whittaker, J. (2014). Corporate Social Responsibility and Firm Value: Disaggregating the Effects on Cash Flow, Risk and Growth. *Journal of Business Ethics*, *124*(4), 633–657. https://doi.org/10.1007/S10551-013-1898-5

GRI. (2021). *A Short Introduction to the GRI Standards*.

Guo, B., Paraskevopoulou, E., & Santamaría Sánchez, L. (2019). Disentangling the Role of Management Control Systems for Product and Process Innovation in Different Contexts. *European Accounting Review*, *28*(4), 681–712. 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. https://doi.org/10.1093/oep/gpt012

Hoberg, G., & Maksimovic, V. (2015). Redefining Financial Constraints: A Text-Based Analysis. *The Review of Financial Studies*, *28*(5), 1312–1352. https://doi.org/10.1093/RFS/HHU089

Holmstrom, B. (1989). Agency costs and innovation. *Journal of Economic Behavior & Organization*, *12*(3), 305–327. https://doi.org/10.1007/s00016-003-0167-x

Huang, F., Xiang, L., Liu, R., Su, S., & Qiu, H. (2019). The IPO corporate social responsibility information disclosure: Does the stock market care? *Accounting & Finance*, *59*(S2), 2157–2198. https://doi.org/https://doi.org/10.1111/acfi.12534

Huang, S., Ng, J., Ranasinghe, T., & Zhang, M. (2021). Do Innovative Firms Communicate More? Evidence from the Relation between Patenting and Management

Guidance. *Accounting Review*, *96*(1), 273–297. https://doi.org/10.2308/TAR-2017-0082

Investor.gov. (n.d.-a). *Form 10-K*. Retrieved July 5, 2024, from https://www.investor.gov/introduction-investing/investing-basics/glossary/form-10-k

Investor.gov. (n.d.-b). *Registration Statement*. Retrieved July 8, 2024, from https://www.investor.gov/introduction-investing/investing-basics/glossary/registration-statement

Jennings, P. D., & Zandbergen, P. A. (1995). Ecologically Sustainable Organizations: An Institutional Approach. *Academy of Management Review*, *20*(4), 1015–1052. https://doi.org/10.5465/AMR.1995.9512280034

Jones, M. J., & Shoemaker, P. A. (1994). Accounting narratives: A review of empirical studies of content and readability. *Journal of Accounting Literature*, *13*, 142.

Kaur, A., & Lodhia, S. (2018). Stakeholder engagement in sustainability accounting and reporting: A study of Australian local councils. *Accounting, Auditing and Accountability Journal*, *31*(1), 338–368. https://doi.org/10.1108/AAAJ-12-2014-1901/FULL/PDF

Kim, C. (Francis), Wang, K., & Zhang, L. (2019). Readability of 10-K Reports and Stock Price Crash Risk. *Contemporary Accounting Research*, *36*(2), 1184–1216. https://doi.org/10.1111/1911-3846.12452

Kogan, L., Papanikolaou, D., Seru, A., & Stoffman, N. (2017). Technological Innovation, Resource Allocation, and Growth. *The Quarterly Journal of Economics*, *132*(2), 665–712. https://doi.org/10.1093/QJE/QJW040

Kolk, A., Levy, D., & Pinkse, J. (2008). Corporate responses in an emerging climate regime: The institutionalization and commensuration of carbon disclosure. *European Accounting Review*, *17*(4), 719–745. https://doi.org/10.1080/09638180802489121

Lambertsen, N. N. (2024). Manipulation and obfuscation of financial reports. *Journal of Business Finance & Accounting*, *51*(1–2), 276–296. https://doi.org/10.1111/JBFA.12693

Lammers, J. C., & Barbour, J. B. (2006). An institutional theory of organizational communication. *Communication Theory*, *16*(3), 356–377. https://doi.org/10.1111/J.1468-2885.2006.00274.X

Lee, C., Lee, S., Kim, J., & Lim, J. S. (2024). Analyzing social media reactions to the FTX crisis: Unraveling the spillover effect on crypto markets. *Journal of Contingencies and Crisis Management*, *32*(2), e12577. https://doi.org/10.1111/1468-5973.12577

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. https://doi.org/10.2308/accr.00000043

Lewis, L., & Unerman, J. (1999). ETHICAL RELATIVISM: A REASON FOR DIFFERENCES IN CORPORATE SOCIAL REPORTING? *Critical Perspectives on Accounting*, *10*(4), 521–547. https://doi.org/10.1006/CPAC.1998.0280

Lewis, N. R., Parker, L. D., Pound, G. D., & Sutcliffe, P. (1986). Accounting Report Readability: The Use of Readability Techniques. *Accounting and Business Research*, *16*(63), 199–213.

Li, F. (2008). Annual report readability, current earnings, and earnings persistence. *Journal of Accounting and Economics*, *45*(2–3), 221–247. https://doi.org/10.1016/j.jacceco.2008.02.003

Li, F. (2010a). Textual Analysis of Corporate Disclosures: A Survey of the Literature. *Journal of Accounting Literature*, *29*, 143–165.

Li, F. (2010b). The information content of forward- looking statements in corporate filings-A naïve bayesian machine learning approach. *Journal of Accounting Research*, *48*(5), 1049–1102. https://doi.org/10.1111/J.1475-679X.2010.00382.X

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. *Review of Financial Studies*, *34*(7), 3265–3315. https://doi.org/10.1093/RFS/HHAA079

Li, S., Wang, G., & Luo, Y. (2022). Tone of language, financial disclosure, and earnings management: a textual analysis of form 20-F. *Financial Innovation*, *8*(1), 43. https://doi.org/10.1186/s40854-022-00346-5

Liu, W., Yuan, C., Wang, J., Lim, M. K., & Hou, J. (2024). Digital supply chain announcements and firm's stock market value: An empirical study from China. *Transportation Research Part E: Logistics and Transportation Review*, *187*, 103604. https://doi.org/10.1016/J.TRE.2024.103604

Li, Y., Gong, M., Zhang, X. Y., & Koh, L. (2018). The impact of environmental, social, and governance disclosure on firm value: The role of CEO power. *The British Accounting Review*, *50*(1), 60–75. https://doi.org/10.1016/J.BAR.2017.09.007

Loughran, T., & McDonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x

Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*(2), 307–326. https://doi.org/10.1016/J.JFINECO.2013.02.017

Loughran, T., & McDonald, B. (2016). Textual Analysis in Accounting and Finance: A Survey. *Journal of Accounting Research*, *54*(4), 1187–1230. https://doi.org/10.1111/1475-679X.12123

Lukka, K. (2010). The roles and effects of paradigms in accounting research. *Management Accounting Research*, *21*(2), 110–115. https://doi.org/10.1016/J.MAR.2010.02.002

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

Masulis, R. W., & Reza, S. W. (2015). Agency problems of corporate philanthropy. *Review of Financial Studies*, *28*(2), 592–636. https://doi.org/10.1093/RFS/HHU082

Ma, W., Wang, X., Wang, Y., & Wu, G. (2021). Measuring misleading information in IPO prospectuses. *Review of Quantitative Finance and Accounting*, *57*(3), 819–843. https://doi.org/10.1007/s11156-021-00964-7

Michael, A., & Dixon, R. (2019). Audit data analytics of unregulated voluntary disclosures and auditing expectations gap. *International Journal of Disclosure and Governance*, *16*(4), 188–205. https://doi.org/10.1057/S41310-019-00065-X/TABLES/4

Michael J. Casey, & Paul Vigna. (2018). *In blockchain we trust*. MIT Technology Review. https://www.technologyreview.com/2018/04/09/3066/in-blockchain-we-trust/

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*. https://doi.org/10.48550/arXiv.1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*. https://doi.org/10.48550/arXiv.1310.4546

Moneva, J. M., Rivera-Lirio, J. M., & Muñoz-Torres, M. J. (2007). The corporate stakeholder commitment and social and financial performance. *Industrial Management and Data Systems*, *107*(1), 84–102. https://doi.org/10.1108/02635570710719070

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

Nekhili, M., Boukadhaba, A., & Nagati, H. (2021). The ESG–financial performance relationship: Does the type of employee board representation matter? *Corporate Governance: An International Review*, *29*(2), 134–161. https://doi.org/10.1111/CORG.12345

Neu, D., Warsame, H., & Pedwell, K. (1998). Managing Public Impressions: Environmental Disclosures in Annual Reports. *Accounting, Organizations and Society*, *23*(3), 265–282. https://doi.org/10.1016/S0361-3682(97)00008-1

Nishant, R., Teo, T. S. H., & Goh, M. (2017). Do Shareholders Value Green Information Technology Announcements? *Journal of the Association for Information Systems*, *18*(8), 542–576. https://doi.org/10.17705/1jais.00466

Palazzo, G., & Scherer, A. G. (2006). Corporate legitimacy as deliberation: A communicative framework. *Journal of Business Ethics*, *66*(1), 71–88. https://doi.org/10.1007/S10551-006-9044-2

Patelli, L., & Pedrini, M. (2013). Is the Optimism in CEO's Letters to Shareholders Sincere? Impression Management Versus Communicative Action During the Economic Crisis. *Journal of Business Ethics 2013 124:1*, *124*(1), 19–34. https://doi.org/10.1007/S10551-013-1855-3

Patten, D. M. (1992). Intra-industry environmental disclosures in response to the Alaskan oil spill: A note on legitimacy theory. *Accounting, Organizations and Society*, *17*(5), 471–475. https://doi.org/10.1016/0361-3682(92)90042-Q

Ranta, M., Ylinen, M., & Järvenpää, M. (2023). Machine Learning in Management Accounting Research: Literature Review and Pathways for the Future. *European Accounting Review*, *32*(3), 607–636. https://doi.org/10.1080/09638180.2022.2137221

Ross, S. A. (1977). Determination of financial structure: The incentive-signalling approach. *Bell J Econ*, *8*(1), 23–40. https://doi.org/10.2307/3003485

Russo, M. V., & Fouts, P. A. (1997). A resource-based perspective on corporate environmental performance and profitability. *Academy of Management Journal*, *40*(3), 534–559. https://doi.org/10.2307/257052

Saidi, F., & Žaldokas, A. (2021). How Does Firms' Innovation Disclosure Affect Their Banking Relationships? *Management Science*, *67*(2), 742–768. https://doi.org/10.1287/mnsc.2019.3498

Sarkees, M. (2011). Understanding the links between technological opportunism, marketing emphasis and firm performance: Implications for B2B. *Industrial Marketing Management*, *40*(5), 785–795. https://doi.org/10.1016/J.INDMARMAN.2010.09.001

Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*(3), 355–374. https://doi.org/10.2307/1882010

Srinivasan, R., Lilien, G. L., & Rangaswamy, A. (2002). Technological opportunism and radical technology adoption: An application to e-business. *Journal of Marketing*, *66*(3), 47–60. https://doi.org/10.1509/JMKG.66.3.47.18508

Stratopoulos, T. C., Wang, V. X., & Ye, H. (2022). Use of Corporate Disclosures to Identify the Stage of Blockchain Adoption. *Accounting Horizons*, *36*(1), 197–220. https://doi.org/10.2308/HORIZONS-19-101

Suchman, M. C. (1995). Managing Legitimacy: Strategic and Institutional Approaches. *Academy of Management Review*, *20*(3), 571–610. https://doi.org/10.5465/AMR.1995.9508080331

Sun, Y., Zhao, D., & Cao, Y. (2024). The impact of ESG performance, reporting framework, and reporting assurance on the tone of ESG disclosures: Evidence from

Chinese listed firms. *Journal of Cleaner Production*, *466*, 142698. https://doi.org/10.1016/J.JCLEPRO.2024.142698

Tetlock, P. C., Saar-Tsechansky, M., & Macskassy, S. (2008). More Than Words: Quantifying Language to Measure Firms' Fundamentals. *The Journal of Finance*, *63*(3), 1437–1467. https://doi.org/10.1111/J.1540-6261.2008.01362.X

Thompson, J. D. (1967). *Organizations in Action*. McGraw-Hill Book Co.

Tian, X., & Wang, T. Y. (2014). Tolerance for failure and corporate innovation. *Review of Financial Studies*, *27*(1), 211–255. https://doi.org/10.1093/rfs/hhr130

Tilling, M. V., & Tilt, C. A. (2010). The edge of legitimacy: Voluntary social and environmental reporting in Rothmans' 1956-1999 annual reports. *Accounting, Auditing and Accountability Journal*, *23*(1), 55–81. https://doi.org/10.1108/09513571011010600/FULL/PDF

Tregidga, H., & Laine, M. (2022). On crisis and emergency: Is it time to rethink long-term environmental accounting? *Critical Perspectives on Accounting*, *82*, 102311. https://doi.org/10.1016/J.CPA.2021.102311

Tsang, R. C. W., Baldwin, A. A., Hair Jr., J. F., Affuso, E., & Lahtinen, K. D. (2023). The Informativeness of Sentiment Types in Risk Factor Disclosures: Evidence from Firms with Cybersecurity Breaches. *Journal of Information Systems*, *37*(3), 157–190. https://doi.org/10.2308/ISYS-2022-014

US Securities & Exchange Commission. (1934). *Form 10-K*.

Verrecchia, R. E. (1983). Discretionary disclosure. *Journal of Accounting and Economics*, *5*(C), 179–194. https://doi.org/10.1016/0165-4101(83)90011-3

Wasiuzzaman, S., Yong, F. L. K., Sundarasen, S. D. D., & Othman, N. S. (2018). Impact of disclosure of risk factors on the initial returns of initial public offerings (IPOs). *Accounting Research Journal*, *31*(1), 46–62. https://doi.org/10.1108/ARJ-09-2016-0122

Watson, A., Shrives, P., & Marston, C. (2002). Voluntary disclosure of accounting ratios in the UK. *The British Accounting Review*, *34*(4), 289–313. https://doi.org/10.1006/BARE.2002.0213

Yen, J.-C., & Wang, T. (2021). Stock price relevance of voluntary disclosures about blockchain technology and cryptocurrencies. *International Journal of Accounting Information Systems*, *40*, 100499. https://doi.org/https://doi.org/10.1016/j.accinf.2021.100499

Zhang, Y., & Wiersema, M. F. (2009). Stock market reaction to CEO certification: The signaling role of CEO background. *Strategic Management Journal, 30*(7), 693–710. https://doi.org/10.1002/SMJ.772

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	0.0274***	0.0520***	0.0362***	0.0482***	0.0368***	0.0537***	0.0500***	0.0646***
(topic dist.)	(0.0087)	(0.0118)	(0.0090)	(0.0125)	(0.0094)	(0.0130)	(0.0112)	(0.0156)
ROA	−0.0092***	−0.0105***	−0.0034***	−0.0037***	−0.0034***	−0.0037***	−0.0034***	−0.0037***
	(0.0006)	(0.0008)	(0.0003)	(0.0004)	(0.0003)	(0.0004)	(0.0003)	(0.0004)
R&D/Sales	0.00002	0.00003	0.00001	0.00002	0.00001	0.00002	0.00001	0.00002
	(0.00004)	(0.00005)	(0.00002)	(0.00003)	(0.00002)	(0.00003)	(0.00002)	(0.00003)
log (*Sales*)	0.0193	0.0163	−0.0368*	−0.0368	−0.0365*	−0.0376	−0.0368*	−0.0364
	(0.0211)	(0.0300)	(0.0197)	(0.0281)	(0.0196)	(0.0279)	(0.0197)	(0.0280)
log (*Assets*)	0.4874***	0.5865***	0.4873***	0.5714***	0.4852***	0.5700***	0.4862***	0.5696***
	(0.0209)	(0.0294)	(0.0205)	(0.0292)	(0.0196)	(0.0289)	(0.02094)	(0.0290)
log (*Age*)	0.2999***	0.2061***	0.2651***	0.1692***	0.2668***	0.1713***	0.2616***	0.1648***
	(0.0319)	(0.0414)	(0.0311)	(0.0406)	(0.0311)	(0.0406)	(0.0311)	(0.0406)
Total debt/Total capital	−0.0023***	−0.0031***	−0.0012***	−0.0017***	−0.0012***	−0.0017***	−0.0012***	−0.0016***
	(0.0005)	(0.0007)	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0004)	(0.0005)
Beta	−0.0675***	−0.0668**	−0.0434**	−0.0448*	−0.0438**	−0.0449*	−0.0424**	−0.0436*
	(0.0185)	(0.0265)	(0.0178)	(0.0254)	(0.0178)	(0.0253)	(0.0178)	(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:

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

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

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	−0.4340***	−0.5757***	1.4248***	1.9937***	1.3307***	1.8003***	1.3748***	1.8823***
(topic weight)	(0.1210)	(0.1685)	(0.0883)	(0.1233)	(0.0983)	(0.1358)	(0.0999)	(0.1394)
ROA	−0.0033***	−0.0103***	−0.0036***	−0.0040***	−0.0036***	−0.0039***	−0.0036***	−0.0039***
	(0.0003)	(0.0008)	(0.0003)	(0.0004)	(0.0003)	(0.0004)	(0.0003)	(0.0004)
R&D/Sales	0.00001	0.00001	0.00001	0.00002	0.00001	0.00003	0.00001	0.00003
	(0.00002)	(0.00005)	(0.00002)	(0.00003)	(0.00002)	(0.00003)	(0.00002)	(0.00003)
log (*Sales*)	0.0266	0.0259	−0.0387**	−0.0401	−0.0324*	−0.0310	−0.0328*	−0.0316
	(0.0194)	(0.0298)	(0.0191)	(0.0271)	(0.0193)	(0.0275)	(0.0191)	(0.0272)
log (*Assets*)	0.4715***	0.5703***	0.4902***	0.5764***	0.4853***	0.5690***	0.4874***	0.5720***
	(0.0200)	(0.0290)	(0.0200)	(0.0282)	(0.0202)	(0.0286)	(0.0200)	(0.0282)
log (*Age*)	0.2560***	0.1804***	0.2474***	0.1443***	0.2572***	0.1584***	0.2483***	0.1461***
	(0.0311)	(0.0412)	(0.0313)	(0.0408)	(0.0314)	(0.0409)	(0.0314)	(0.0409)
Total debt/Total	−0.0010***	−0.0031***	−0.0005***	−0.0008***	−0.0008*	−0.0011**	−0.0007*	−0.0010***
capital	(0.0004)	(0.0007)	(0.0004)	(0.0005)	(0.0004)	(0.0005)	(0.0004)	(0.0005)
Beta	−0.0435**	−0.0689***	−0.0559***	−0.0619**	−0.0582***	−0.0647***	−0.0536***	−0.0586***
	(0.0173)	(0.0264)	(0.0175)	(0.0249)	(0.0177)	(0.0252)	(0.0175)	(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	0.0137	0.0449***	0.0327***	0.0469***	0.0290***	0.0547***	0.0209	0.0286
(topic dist.)	(0.0101)	(0.0148)	(0.0114)	(0.0166)	(0.0110)	(0.0160)	(0.0141)	(0.0204)
ROA	−0.0077***	−0.0086***	−0.0077***	−0.0084***	−0.0077***	−0.0084***	−0.0077***	−0.0084***
	(0.0008)	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0012)
R&D/Sales	0.00005*	0.00005	0.00005*	0.00005	0.00005*	0.00005	0.00005*	0.00005
	(0.00003)	(0.00004)	(0.00003)	(0.00004)	(0.00003)	(0.00004)	(0.00003)	(0.00004)
log (*Sales*)	0.0746**	0.0702	0.0695***	0.0702	0.0698**	0.0667	0.0741**	0.0771*
	(0.0290)	(0.0445)	(0.0291)	(0.0446)	(0.0291)	(0.0443)	(0.0291)	(0.0445)
log (*Assets*)	0.4742***	0.5463***	0.4774***	0.5408***	0.4765***	0.5432***	0.4728***	0.5340***
	(0.0282)	(0.0445)	(0.0280)	(0.0442)	(0.0280)	(0.0440)	(0.0280)	(0.0441)
log (*Age*)	−0.0105	−0.1245**	−0.0174	−0.1445***	−0.0154	−0.1415**	−0.0180	−0.1453***
	(0.0390)	(0.0561)	(0.0385)	(0.0557)	(0.0386)	(0.0558)	(0.0387)	(0.0558)
Total debt/Total	−0.0004	−0.0006	−0.0003	−0.0006	−0.0003	−0.0005	−0.0003	−0.0006
capital	(0.0005)	(0.0007)	(0.0005)	(0.0007)	(0.0005)	(0.0007)	(0.0005)	(0.0007)
Beta	−0.0026	0.0005	−0.0004	−0.0039	−0.0023	−0.0047	−0.0031	−0.0080
	(0.0305)	(0.0471)	(0.0303)	(0.0467)	(0.0302)	(0.0465)	(0.0302)	(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***	−0.7660***	0.9202***	1.2617***	0.7943***	1.0295***	0.9175***	1.2452***
	(0.1459)	(0.2135)	(0.1005)	(0.1495)	(0.1030)	(0.1551)	(0.1032)	(0.1560)
ROA	−0.0077***	−0.0082***	−0.0080***	−0.0088***	−0.0080***	−0.0087***	−0.0079***	−0.0087***
	(0.0008)	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0012)
R&D/Sales	0.00006**	0.00005	0.00004	0.00004	0.00005*	0.00004	0.00004*	0.00004
	(0.00003)	(0.00004)	(0.00003)	(0.00004)	(0.00003)	(0.00004)	(0.00003)	(0.00004)
log (*Sales*)	0.0899***	0.0917**	0.0600**	0.0578	0.0680**	0.0695	0.0643**	0.0638
	(0.0285)	(0.0432)	(0.0289)	(0.0441)	(0.0292)	(0.0443)	(0.0288)	(0.0440)
log (*Assets*)	0.4560***	0.5060***	0.4983***	0.5690***	0.4883***	0.5538***	0.4936***	0.5622***
	(0.0275)	(0.0428)	(0.0282)	(0.0441)	(0.0285)	(0.0444)	(0.0279)	(0.0438)
log (*Age*)	−0.0303	−0.1594***	−0.0319	−0.1644***	−0.0297	−0.1602***	−0.0327	−0.1652***
	(0.0383)	(0.0547)	(0.0387)	(0.0563)	(0.0390)	(0.0564)	(0.0389)	(0.0564)
Total debt/Total capital	−0.0003	−0.0004	−0.00004	−0.00003	−0.00006	−0.0002	−0.000001	−0.00009
	(0.0005)	(0.0007)	(0.0005)	(0.0007)	(0.0005)	(0.0007)	(0.0005)	(0.0007)
Beta	−0.0031	−0.0027	−0.0073	−0.0137	−0.0155	−0.0242	−0.0062	−0.0122
	(0.0298)	(0.00461)	(0.0296)	(0.0459)	(0.0301)	(0.0465)	(0.0297)	(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***	1.0620***	−0.9125 (2.3115)	−1.2320 (3.1606)	−0.2100	0.4151	0.9685***	1.1144***
(0.1153)	(0.1771)			(0.1831)	(0.2787)	(0.1405)	(0.2170)
−0.0081***	−0.0090***	−0.0077***	−0.0082***	−0.0077***	−0.0083***	−0.0077***	−0.0086***
(0.0009)	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0012)	(0.0008)	(0.0012)
0.00005*	0.00005	0.00006**	0.00005	0.00005*	0.00005	0.00005*	0.00004
(0.00004)	(0.00004)	(0.00003)	(0.00004)	(0.00003)	(0.00004)	(0.00003)	(0.00004)
0.0718**	0.0742	0.0873***	0.0883**	0.0798***	0.0809*	0.0692**	0.0723
(0.0296)	(0.0444)	(0.0286)	(0.0434)	(0.0291)	(0.0438)	(0.0291)	(0.0442)
0.4856***	0.5509***	0.4589***	0.5098***	0.4684***	0.5194***	0.4840***	0.5461***
(0.0287)	(0.0444)	(0.0276)	(0.0430)	(0.0281)	(0.0434)	(0.0282)	(0.0441)
−0.0198	−0.1476***	−0.0269	−0.1548***	−0.0172	−0.1399**	−0.0382	−0.1681***
(0.0389)	(0.0562)	(0.0384)	(0.0549)	(0.0387)	(0.0560)	(0.0389)	(0.0562)
−0.0001	−0.0003	−0.0003	−0.0005	−0.0004	−0.0006	−0.0002	−0.0004
(0.0005)	(0.0007)	(0.0005)	(0.0007)	(0.0005)	(0.0007)	(0.0005)	(0.0007)
−0.0188	−0.0289	−0.0062	−0.0069	−0.0104	0.0051	−0.0165	−0.0241
(0.0304)	(0.0469)	(0.0300)	(0.0463)	(0.0303)	(0.0468)	(0.0303)	(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:

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

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

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***	1.0271***	0.08395***	0.8930***	0.3846*	0.1354	1.3876***	0.9754***		
(0.2649)	(0.2635)	(0.1545)	(0.1612)	(0.2078)	(0.2010)	(0.3311)	(0.3319)		
								−2.8468***	−2.4316***
								(0.2334)	(0.2398)
−0.0046***	−0.0050***	−0.0113***	−0.0116***	−0.0048***	−0.0052***	−0.0047***	−0.0051***	−0.0038***	−0.0043***
(0.0008)	(0.0008)	(0.0010)	(0.0011)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)	(0.0008)
5.0810***	4.5308***	4.3799***	4.0082***	5.0080***	4.4608***	5.0136***	4.4692***	6.1741***	5.4459***
(0.2412)	(0.2501)	(0.1576)	(0.1617)	(0.2400)	(0.2500)	(0.2397)	(0.2499)	(0.2982)	(0.3125)
1.6810***	1.6614***	1.8793***	1.7691***	1.7892***	1.7499****	1.6698***	1.6678***	2.5418***	2.3859***
(0.4947)	(0.4950)	(0.3658)	(0.3723)	(0.4936)	(0.4947)	(0.4928)	(0.4937)	(0.5001)	(0.5021)
−0.0568***	−0.0352	−0.0672***	−0.0360***	−0.0602***	−0.0382	−0.0582***	−0.0367	−0.0651***	−0.0398*
(0.0209)	(0.0237)	(0.0123)	(0.0129)	(0.0209)	(0.0237)	(0.0209)	(0.0237)	(0.0207)	(0.0232)
−4.5065***	−4.3097***	−4.3791***	−4.6162***	−4.5632***	−4.3659***	−4.5053***	−4.3197***	−4.5904***	−4.3802***
(0.5667)	(0.5722)	(0.4039)	(0.4211)	(0.5684)	(0.5738)	(0.5664)	(0.5737)	(0.5603)	(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***	0.0063**	0.0022	0.0041	0.0058*	0.0092**	0.0007	0.0039
	(0.0036)	(0.0031)	(0.0033)	(0.0036)	(0.0035)	(0.0037)	(0.0042)	(0.0045)
log (*Patents*)								
R&D/Sales	0.000002	0.0000003	−0.000003	0.000001	−0.000003	0.000002	−0.000004	0.000001
	(0.00002)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
log (*Sales*)	−0.0048	−0.0069**	−0.0091***	−0.0082***	−0.0092***	−0.0084***	−0.0092***	−0.0084***
	(0.0030)	(0.0030)	(0.0030)	(0.0031)	(0.0030)	(0.0030)	(0.0030)	(0.0031)
log (*Age*)	−0.0598**	−0.0529***	−0.0589***	−0.0567***	−0.0588***	−0.0565***	−0.0587***	−0.0570***
	(0.0087)	(0.0084)	(0.0085)	(0.0088)	(0.0085)	(0.0087)	(0.0086)	(0.0088)
Total debt/Total capital	0.00001	0.0006**	0.0002	0.0006**	0.0002	0.0006**	0.0002	0.0006**
	(0.0003)	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0003)	(0.0002)	(0.0003)
Beta	−0.0047	−0.0202**	−0.0100	−0.0132	−0.0097	−0.0128	−0.0103	−0.0132
	(0.0092)	(0.0085)	(0.0086)	(0.0088)	(0.0085)	(0.0088)	(0.0085)	(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	0.0035	−0.0007	0.0096	0.0107***	0.0137***	0.0176***	0.0233***		
(0.0046)	(0.0048)	(0.0169)	(0.0178)	(0.0035)	(0.0037)	(0.0058)	(0.0062)		
								0.0397***	0.0420***
								(0.0040)	(0.0042)
−0.000004	0.000001	0.0001	0.00001	−0.000002	0.000003	−0.000002	0.000003	−0.000002	0.00001
(0.00001)	(0.00001)	(0.0001)	(0.0001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
−0.0093***	−0.0083***	−0.0530***	−0.0645***	−0.0091***	−0.0083***	−0.0091***	−0.0083***	−0.0255***	−0.0256***
(0.0031)	(0.0031)	(0.0143)	(0.0161)	(0.0030)	(0.0030)	(0.0030)	(0.0030)	(0.0037)	(0.0037)
−0.0586***	−0.0568***	−0.2910***	−0.2638***	−0.0593****	−0.0570***	−0.0605***	0.0586***	−0.0690***	−0.0674***
(0.0086)	(0.0088)	(0.0422)	(0.0448)	(0.0085)	(0.0087)	(0.0085)	(0.0087)	(0.0086)	(0.0088)
0.0002	0.0006**	−0.0004	0.0019	0.0002	0.0006**	0.0002	0.0006**	0.0002	0.0006**
(0.0002)	(0.0003)	(0.0014)	(0.0015)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
−0.0103	−0.0133	−0.0402	−0.0615	−0.0098	−0.01303	−0.0093	−0.0123	−0.0100	−0.0132
(0.0085)	(0.0088)	(0.0421)	(0.0456)	(0.0085)	(0.0088)	(0.0085)	(0.0088)	(0.0085)	(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*	−0.0380	−1.6816***	−1.6249***	0.2650***	0.3273***	−0.0508	−0.0149		
(0.0359)	(0.0370)	(0.3775)	(0.5020)	(0.0482)	(0.0497)	(0.0423)	(0.0439)		
								0.0397***	0.0420***
								(0.0040)	(0.0042)
0.0000001	0.000004	0.000006	0.000003	0.000002	0.000006	0.000001	0.000005	−0.000002	0.00001
(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
−0.0043	−0.0036	−0.0007	−0.0011	−0.0049*	−0.0045	−0.0039	−0.0033	−0.0255***	−0.0256***
(0.0030)	(0.0030)	(0.0028)	(0.0029)	(0.0030)	(0.0030)	(0.0029)	(0.0029)	(0.0037)	(0.0037)
−0.0576***	−0.0553***	−0.0579***	−0.0543***	−0.0566***	−0.0536***	−0.0577***	−0.0555***	−0.0690***	−0.0674***
(0.0084)	(0.0086)	(0.0080)	(0.0083)	(0.0084)	(0.0086)	(0.0084)	(0.0086)	(0.0086)	(0.0088)
0.00003	0.0005*	0.00001	0.0005**	0.00001	0.0005**	0.00005	0.0005*	0.0002	0.0006**
(0.0002)	(0.0003)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0003)	(0.0002)	(0.0002)
−0.0120	−0.0170**	−0.0197**	−0.0254***	−0.0111	−0.0157*	−0.0124	−0.0173**	−0.0100	−0.0132
(0.0084)	(0.0086)	(0.0081)	(0.0084)	(0.0084)	(0.0087)	(0.0084)	(0.0086)	(0.0085)	(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.

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

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.

Alternative text no.	Dependent variable				
	Alternative innovation book text innovation **(topic distribution, 40-topic)**				
	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

Table 2. Corrected descriptive statistics

Variable	**log(*Patents*)**	**log(*Citations*)**	**log(*Sales*)**	**log(*Assets*)**	**log(*Age*)**	***R&D/Sales***	***Total Debt/Total Capital***	***ROA***	**log(*Q*)**	***Beta***
Count	30,641	30,641	39,829	39,888	33,958	17,154	39,155	39,157	35,880	33,239
Mean	0.61	0.74	11.88	13.07	9.21	215.81	33.54	-20.42	0.98	1.93
Standard Deviation	1.29	1.66	3.75	2.81	0.60	1125.22	45.74	107.78	0.41	35.09
Median	0.00	0.00	12.57	13.40	9.19	5.51	31.10	2.43	0.85	1.22

Essay II: ESG Disclosure and ESG Performance of Seeking Buyer Companies

Tatiana King, Essi Nousiainen & Mikko Ranta

School of Accounting & Finance

University of Vaasa

Abstract

This paper explores the involvement in ESG practices of companies that publicly seek a buyer. We analyse whether these companies increase ESG disclosure and performance before the announcement and whether these actions influence the acquisition outcome. The second essay extends the literature by exploring ESG activities in firms seeking buyers, finding that these firms disclose more ESG information but do not differ in ESG performance compared to peers, suggesting this disclosure may be superficial. Further, our results indicate that ESG practices are not associated with the likelihood of being acquired. We conclude that acquiring companies might consider seeking buyers' involvement in enhanced ESG disclosure 'window dressing.' The deals might also be closed based on targets' financial characteristics, especially when buyers lack a shared understanding of disclosed ESG information and when this information is not supported by ESG performance. This paper contributes to ongoing debates about the necessity for clear, formalized regulatory guidelines linking ESG disclosures to performance indicators, potentially enhancing decision-making efficiency in M&A transactions.

Keywords: ESG disclosure; ESG performance; M&A; seeking buyer firms; machine learning

JEL: C45, G34, G41, M14, M40,

1. Introduction

Involvement in environmental, social and governance (ESG) activities has become important in making investment decisions in financial markets (Christensen et al., 2022). Involvement in ESG practices may improve a company's value creation (Gregory et al., 2014; Li et al., 2018; Dhaliwal et al., 2012), firms' financial performance (Li et al., 2018; Nekhili et al., 2021), lead to lower cost of equity and debt (Dhaliwal et al., 2011; Eliwa et al., 2021; Erragragui, 2018), improve firms' competitive advantage (Shrivastava, 1995; Russo and Fouts, 1997) and help establish long-term relations with stakeholders (Donaldson and Preston, 1995; Rezaee, 2016; Tregidga and Laine, 2022).

Several studies have investigated the impact of firms' engagement in ESG activities on the outcomes of M&A deals (e.g. Aktas et al., 2011; Bose et al., 2021; Boone and Uysal, 2020; Deng et al., 2013; Fairhurst and Greene, 2022). Researchers mainly conclude that involvement in ESG practices leads to better announcement returns (Deng et al., 2013), higher takeover premiums (Gomes and Marsat, 2018), and value creation for shareholders in the context of M&A (Aktas et al., 2011). Companies are more often considered as targets when they have high or low levels of CSR (corporate social responsibility, CSR, often used interchangeably with ESG) performance (Gomes, 2019; Fairhurst and Greene, 2022). Recently, ESG engagement has been considered as a mitigating factor for the diversification discount in cross-border acquisitions (Kim et al., 2021) that may influence the choice of a target firm in leveraged buyouts (Onukar, 2021). To the best of our knowledge, all the existing studies investigating involvement in ESG activities in corporate acquisitions focus on a situation when the buyer initiates the deal and takes ownership of another entity's share capital, equity interests or assets or when two companies combine to form a new enterprise together.

In practice, a significant proportion of M&A deals are target-initiated, which is largely neglected in prior research. Mulherin and Simsir (2015) calculated the proportion of search-for-buyer cases may be as high as 38.7% of all deals with the original reported date disclosed. Murdock (2010) also reconfirms that issuing a so-called 'seeking buyer announcement' to facilitate a sale process has become a persistent phenomenon over the years.

Companies openly looking for a buyer represent an interesting context for examining a firm's motivation to engage in ESG activities as these types of firms have time and incentives to exercise changes related to involvement in ESG practices. We build our assumptions on two theories – legitimacy (Ashforth and Gibbs, 1990; Dowling and Pfeffer, 1975; Eliwa et al., 2021) and signalling (Dhaliwal et al., 2011; Macias et al., 2011; Wong and Zhang, 2021). In line with legitimacy theory, firms aim to ensure that they are perceived as operating within the bounds and norms of their societies (Deegan and Unerman, 2011) meaning that firms attempt to ensure that their activities are perceived by external parties as being legitimate. Therefore, firms may be more involved in ESG activities because they are able to influence societal appraisal to increase their legitimacy (Deephouse, 1996; Suchman, 1995). According to signalling theory, companies engage in ESG initiatives to minimise information asymmetry, lessen uncertainty, influence share price response (Ramchander et al., 2012) and improve trust among different stakeholders (Feng et al., 2022; Dunn and Sainty, 2009).

Utilizing the frameworks of legitimacy and signalling theories, our study examines whether firms actively seeking buyers intensify their engagement in ESG activities as a strategy to streamline their sale process. Stolowy and Paugam (2018) claim that between 2006 and 2016, the percentage of firms reporting sustainability information increased from less than 10% to more than 80% in Europe and to more than 60% in the US. However, the existing US Generally Accepted Accounting Principles (GAAP) require relatively few aspects of ESG information to be disclosed in 10-K filings (e.g., asset retirement obligations, environmental remediation liabilities, etc.), and a significant part of ESG-related information remains voluntary and unregulated in the US market (De Villiers and Marques, 2016; Fatemi et al., 2018; Clarkson et al., 2020). Therefore, there are significant differences in the quantity of ESG information voluntarily disclosed by US companies.

The lack of regulation on ESG disclosure may potentially lead to a situation where some managers opportunistically manipulate the amount of ESG disclosure to signal high ESG involvement without real results supporting that (see Talpur et al., 2023 for review). Therefore, in our study, we examine whether disclosed ESG information is related to the ESG performance of seeking buyer companies to assess whether they 'window dress' their annual reports or genuinely invest in ESG activities. Our assumptions are based on the idea that seeking buyer firms initiate deals themselves and, therefore, have enough time to engage in ESG activities to make themselves more attractive to potential buyers. We also investigate whether involvement in ESG activities of seeking buyer firms impacts the outcome of the M&A deal and whether the buyers take into consideration involvement in ESG activities of target firms while making the final decision about the deal.

Our findings suggest that seeking buyer firms disclose more ESG information (based on ML ESG disclosure and Bloomberg ESG disclosure - 2-stage least squares). Thus, this indicates that firms seeking a buyer enhance ESG disclosure when preparing for the official announcement in line with legitimacy theory. At the same time, we do not observe higher ESG performance for the seeking buyer sample in comparison to the matched sample. Our results also suggest that there is no relationship between enhanced ESG disclosure before the announcement and the likelihood of being acquired. In other words, the likelihood of being bought is not higher for seeking buyer companies that are more involved in enhanced ESG disclosure, suggesting that buyers might consider these activities as a 'window dressing' strategy (Macias et al., 2011; Delmas and Burbano, 2011). Acquiring companies might also consider seeking buyers' involvement in ESG disclosure as a waste of time and resources (Bénabou and Tirole, 2010; Masulis and Reza, 2015) or as social activities that promote managers' personal reputations as good citizens (Barnea and Rubin, 2010; Cheng et al., 2013) rather than genuine involvement in ESG. Moreover, the results can be explained by differing preferences the buying companies have, like some having no interest in ESG issues and focusing on financial characteristics of seeking buyer companies, like sales growth and leverage, when making the decisions (Datta et al., 2001; Harrison et al., 2014). Target companies with higher leverage and sales growth might also prefer to finish the deals quicker not to be exposed to interest payments and not to have additional pressure from capital reimbursement outflows (Aktas et al., 2010).

Our contribution to the literature is threefold. First, building on the results of previous research on companies' involvement in ESG activities in the context of M&As (see, e.g., Aktas et al., 2011; Deng et al., 2013; Gomes and Marsat, 2018; Gomes, 2019; Boone and Uysal, 2020), our work examines ESG in a specific group of seeking buyer firms and adds to our understanding of the motives behind involvement in ESG activities in M&As. We add to the current research by showing that seeking buyer companies are involved only in enhanced ESG disclosure compared to our matched sample, which is not supported by higher ESG performance.

Second, we apply a novel approach to the evaluation of ESG information in companies' reports. We apply a machine learning technique, namely a word embedding model, to collect and measure the quantity of ESG disclosure in seeking buyer firms' 10-K forms, a document that the US Securities and Exchange Commission (SEC) requires all public companies to file each year. Previous research has mainly used ESG disclosure measures prepared by third parties (e.g. Bloomberg). However, Kimbrough et al. (2022) argue that the provision of ESG information by management has the potential to reduce the discrepancies observed in ESG indices created by external parties. In our machine-learning approach, we draw insight from Li et al. (2021) and create a measure that takes advantage of the word embedding model, which represents words in such a manner as how humans understand words. Unlike more well-defined concepts (e.g. the tone of business outlook), ESG can be described using less frequent words, phrases and idioms that make sense only in a particular context. Our approach allows us to identify hundreds or even thousands of words and phrases related to each ESG dimension based on the context in which they appear. As such, a firm's ESG score is determined by a combination of all these words and phrases, not just the keywords used to define each dimension.

Finally, we contribute to previous research analysing the role of ESG activities in M&A deals with evidence that enhanced ESG disclosure is not associated with the likelihood of being acquired. Based on our findings the probability of acquisition of seeking buyer firms is positively and significantly related to financial characteristics, like leverage and 3-year average sales growth, bringing us to the conclusion that acquirers pay more attention to and consider more important financial performance indicators (Aktas et al., 2010; Datta et al., 2001; Harrison et al., 2014), rather than ESG practices that are hard to assess and analyse. Therefore, the results of the paper call for formalised rules and regulations by standard setters on how ESG practices should be reported in order to clearly reflect ESG outcomes that are easy to interpret and that are linked to ESG performance.

The rest of the study is organised as follows: Section 2 reviews prior literature and develops hypotheses, focusing on the specifics of seeking buyer firms. Section 3 describes the empirical setting and proxies used for evaluating ESG disclosure and performance. In Section 4, we describe the sample selection process and descriptive statistics. Section 5 reports the results. We go on with additional empirical tests in Section 6 and conclude in Section 7.

2. Literature Review and Hypotheses Development

2.1 Involvement of Seeking Buyer Firms in ESG Activities

In earlier literature devoted to the context of mergers and acquisitions, acquiring firms are predominantly assumed to be deal initiators. However, for various reasons companies may wish to sell themselves, their assets or change ownership, initiating the announcement regarding searching for a buyer. While the information that a company is open to bids is often conveyed 'behind closed doors' and is communicated only to a limited number of potential buyers (Boone and Mulherin, 2008), some companies prefer to make a public announcement of their intentions (Anagnostopoulou and Tsekrekos, 2013; Anagnostopoulou and Tsekrekos, 2015). This may be due to strategic reasons, leverage, growth or distress (see Murdock and Madura, 2011; Aktas et al., 2010) indicating that the board of directors has solid intentions to sell the company (Mulherin and Simsir, 2015).

Prior research documents the existence of earnings management in financial statements by targets that are seeking buyers (e.g. Anagnostopoulou and Tsekrekos, 2015; Macias et al., 2011; Murdock, 2010). Anagnostopoulou and Tsekrekos (2015, p. 352) explain that seeking buyer companies have 'the time, motivation and opportunity to influence their reported numbers in a way that can affect the contractual outcome of their scope.' This contrasts with bidder-initiated deals which naturally deprive the target firm of an opportunity to undertake any accounting-related manipulation to influence contractual outcomes. However, according to Erickson and Wang (1999), the party that is aware of its intentions long before the deal is more motivated and hence, more likely to engage in manipulations (Anagnostopoulou and Tsekrekos, 2015).

Among the incentives for detected earnings management in seeking buyer firms could be to create a more attractive picture of management efficiency and the value of the company, thus reducing the expected cost of a future debt (DeFond and Jiambalvo, 1994) or attracting a buyer by masking inefficiencies or managing public impressions to achieve a more appealing bid price (Alexandridis et al., 2010; Anagnostopoulou and Tsekrekos, 2015). In this case, seeking buyer companies behave similarly to companies preparing for an IPO that have been observed engaging in earnings management characterized by high accruals (Teoh et al., 1998; Alhadab et al., 2015). However, companies' involvement in earnings management should not be considered a positive long-term strategy as these companies are characterised by higher failure of an IPO, lower survival rates in subsequent periods (Alhadab et al., 2015) and poor stock return performance in three years after the IPO (Teoh et al., 1998). Similarities are also observed for companies going through a seasoned equity offering (SEO) as managers of these companies actively engage in more opaque channels to overstate earnings to look more appealing on the market (Kothari et al., 2016; Cohen and Zarowin, 2010).

As involvement in ESG activities is also tailored towards external parties, it is perceived as one of the important instruments for building a positive impression (Neu et al., 1998). Further, according to the 'reputation-building hypothesis', higher involvement in ESG activities may lead to a reduction in the cost of equity capital, better analyst coverage and lower absolute forecast errors (Dhaliwal et al., 2012), lower cost of capital (Dhaliwal et al., 2011; Eliwa et al., 2021) and higher firms' value (Buchanan et al., 2018; Cho et al., 2012). Given the evidence from previous research (e.g. Anagnostopoulou and Tsekrekos, 2013; Anagnostopoulou and Tsekrekos, 2015) that seeking buyer firms have the time and opportunity to manipulate their earnings, we assume that these companies may also have opportunities and motives to improve their engagement in ESG activities – both ESG disclosure and ESG performance – closer to the announcement date. That may be done in order to signal a positive image to stakeholders, influence societal appraisal and increase their legitimacy (Eliwa et al., 2021; Hummel and Schlick, 2016). Involvement in ESG activities may also help seeking buyer firms decrease information asymmetry in the market by signalling their compliance with societal expectations (Cui et al., 2018).

Due to the importance of ESG activities for sustaining competitive advantages, companies disclose information about these activities in their annual reports providing additional information to the stakeholders for decision making (Ryou et al., 2022). From the perspective of signalling theory, good companies would use disclosure to differentiate from bad companies (Spence, 1973). ESG disclosure, among other types of disclosure, can be used to reassure the potential investors about the benefits of the future M&A. As involvement in ESG activities is tailored towards external parties, it is perceived as one of the important instruments for formulating impressions about the company and building a positive image for stakeholders (Jones and Murrell, 2001; Neu et al., 1998). However, signalling requires tangible commitments, for the signal to be effective, it must be backed by observable actions (Connelly, 2010). Accordingly, if ESG disclosure is a believable commitment, the seeking buyer companies that engage in ESG disclosure should also be characterized by higher ESG performance.

From the legitimacy theory point of view, companies are expected to disclose information and to act in accordance with societal norms. However, organizations may be using voluntary reporting as a legitimacy tool to mask actions that are not aligned with the reported values (Tilling and Tilt, 2010). In addition, previous research argues that companies tend to legitimize through CSR reporting following unfavourable press or incidents (Deegan et al., 2002; Patten, 1992), or use environmental disclosure to respond to environmental concerns in the media (Brown & Deegan, 1998). Unlike ESG disclosures, which can be superficial or misleading, actual ESG performance involves measurable changes and improvements in practices and outcomes related to environmental, social, and governance factors. ESG performance is considered an important factor for shareholders' value creation (Buchanan et al., 2018), as socially responsible companies experience positive abnormal returns (Takahashi and Yamada, 2021) and solve the conflicts between managers and stakeholders more easily (Jo and Harjoto, 2012; El Ghoul et al., 2017). Managers engaged in ESG activities reduce the negative externalities of corporations on society that helps companies to legitimate themselves (Moser and Martin, 2012; Baldini et al., 2018). Hence, according to legitimacy theory, ESG disclosure may not always be related to corresponding performance.

Given that seeking buyer companies have the opportunity to influence their reporting prior to the M&A, and the argument that companies may use ESG disclosure as a legitimacy building action to mitigate negative concerns, seeking buyer firms may be engaged in enhanced ESG disclosure, but not equivalent ESG performance. However, if ESG disclosure is a credible commitment, the ESG performance of the seeking buyer companies should also be higher compared to their peers. Against this background we propose our first set of hypotheses:

H1.1: *Seeking buyer firms are characterised by higher ESG disclosure before the announcement date.*

H1.2: *Seeking buyer firms are characterised by higher ESG performance before the announcement date.*

2.2 Seeking Buyers' Involvement in ESG Activities as a Factor in Completing an M&A Deal

Previous research on the role of firms' involvement in ESG in M&A deals is relatively scarce and mainly examines the effect of the acquirer's involvement in CSR activities on announcement returns (Deng et al., 2013), takeover premiums (Gomes and Marsat, 2018), empire building (Gul et al., 2020) and acquisition decisions (Bose et al., 2021). Some studies also focus on analysing target companies' ESG performance and deal outcomes (Aktas et al., 2011; Bereskin et al., 2018; Lin and Wei, 2006; Fairhurst and Greene; 2022). In this paper, we analyse ESG disclosure and performance of seeking buyer firms and the way these activities influence the acquisition outcome that, to the best of our knowledge, has not been studied before.

Prior research documents that a firm's involvement in ESG activities is valuable for potential buyers (Aktas et al., 2011; Deng et al., 2013) and may signal a positive corporate image, revenue-generating opportunities, more predictable earnings, and lower risks (e.g. Brooks and Oikonomou, 2018; Kimbrough and Louis, 2011; Kim et al., 2012). Involvement in ESG is also considered to be an important factor for shareholder value creation (Buchanan et al., 2018; Dimson et al., 2015; Aktas et al., 2011) as companies performing well in ESG activities have a lower cost of equity and debt (Dhaliwal et al., 2011; Eliwa et al., 2021), face positive abnormal returns (Aktas et al., 2011; Dimson et al., 2015) and easier solve the conflicts between managers and stakeholders (Dhaliwal et al., 2012).

Companies involved in ESG practices also tend to have higher voluntary disclosure of these activities (e.g., Dhaliwal et al., 2011; Bilyay-Erdogan, 2022) as high ESG performers signal their quality to the market to positively influence future expectations (Clarkson et al., 2008; Lys et al., 2015). The information disclosed about ESG practices is considered an important means of communication between top managers and stakeholders (Dhaliwal et al., 2011; Dhaliwal et al., 2012; Roberts, 2003) that helps to conduct a moral dialogue between them (Gray, 2002). In line with previous findings, companies with a more aggressive approach towards ESG disclosure are better at enforcing contracts in cross-border acquisitions, therefore, implying that ESG disclosure serves as a conduit for forming mutual trust among stakeholders (Kim et al., 2021). As involvement in ESG activities improves a firm's reputation and transparency (Tsang et al., 2023), management has an opportunity to lower or mitigate the information asymmetry problem by behaving in a socially responsible way and revealing voluntary non-financial disclosure to stakeholders (Cui et al., 2018; Cho et al., 2013). Thus, we argue that involvement in ESG activities – disclosure and performance – can be used by seeking buyer firms as a valuable factor for attracting an acquiring company to initiate an M&A deal and increase the likelihood of being acquired.

At the same time, another stream of research claims that engagement in ESG activities can be treated by stakeholders as unfavourable. In line with agency theory, ESG practices are generally not in the interest of stakeholders. Brown et al. (2006) and Kruger (2015) find that managers engaging in corporate philanthropy benefit themselves at the expense of shareholders. Moreover, engagement in ESG may lead to an increase in agency costs as managers might only tend to promote their personal reputation as good citizens (Barnea and Rubin, 2010; Cheng et al., 2013). Furthermore, managers engaged in time-consuming CSR activities can lose focus from their core managerial responsibilities (Jensen, 2002). This discussion leads us to an assumption that acquiring companies may consider targets' investments in ESG activities as a waste of resources for unjustified purposes related to insider-initiated corporate philanthropy (Bénabou and Tirole, 2010; Masulis and Reza, 2015).

Further, seeking buyer companies may be engaged in ESG activities to affect stakeholders' perceptions (Ashforth and Gibbs, 1990). Levitt (1958, p.47) said that 'welfare and society are not the corporation's business. Its business is making money, not sweet music'. This point of view is supported later by Hemingway and Maclagan (2004), who claim that motivation for engaging in ESG practices is 'greenwashing' and is mainly used to cover up some corporate misbehaviour. These 'window dressing' or 'cheap talk' actions (Delmas and Burbano, 2011) may be used by companies to maintain legitimacy and to have a positive impression on stakeholders pretending to be committed to societal requirements (Ashforth and Gibbs, 1990). In line with this approach, firms 'greenwash' or 'window dress' by enhancing the quantity of ESG disclosure while not having a strong ESG performance. Therefore, this extrapolated ESG disclosure may 'camouflage' real practices to maintain legitimacy and create a more positive image on the market (Eliwa et al., 2021; Michelon et al., 2016). As acquiring companies might not have full trust in ESG information presented by seeking buyer companies, this may negatively influence the likelihood of acquisition.

Understanding buyer preferences is key to why certain companies are targeted for acquisition, particularly regarding their ESG performance and disclosure. However, not all buyers prioritize ESG factors; for many, financial performance, strategic fit, or market position may weigh more heavily in their decision-making. Companies with strong ESG performance can attract buyers focused on sustainability, hoping to bolster their reputational and regulatory standing. Yet, if a buyer's strategic priorities differ, the significance of ESG diminishes. For these buyers, robust financial metrics and strategic synergies are often more critical. While high ESG disclosure can signal commitment to social and environmental values, without tangible performance to support it, such disclosures can also raise doubts about authenticity, potentially deterring acquirers.

Against this multifaceted setting, our second hypothesis is:

H2.1: *A seeking buyer's involvement in ESG activities [disclosure and performance] has a positive relationship with the likelihood of being acquired.*

H2.2: *A seeking buyer's involvement in ESG activities [disclosure and performance] has a negative relationship with the likelihood of being acquired.*

3. Empirical Setting

3.1 Involvement of Seeking Buyer Companies in ESG Activities

In order to test our hypothesis of whether seeking-buyer firms are more involved in ESG activities than their industry peers, we estimate the following logistic regression equation:

$$SeekingBuyer_{it} = \beta_0 + \beta_1 ESGDisclosure_{it-1} + \beta_2 ESGPerformance_{it-1} + \beta_3 ESGDisclosure_{it-1} * ESGPerformance_{it-1} + \beta_4 SizeFirm_{it-1} + \beta_5 ROA_{it-1} + \beta_6 Leverage_{it-1} + \beta_7 AgeFirm_{it-1} + \beta_8 EarningsManagement_{it-1} + \beta_9 DocumentLength_{it-1} + \gamma_t + \mu_i + \varepsilon_{it} \quad (1)$$

The above equation includes all the ESG-related terms. In practice, we evaluate separate models that include the ESG disclosure variable, the ESG performance variable, and their multiplication. The dependent variable *SeekingBuyer* is equal to 1 if the company has issued a seeking buyer announcement during the years 2000-2021 and 0 otherwise. We use logit regression with year and industry -fixed effects for estimations. Following prior literature, we include a series of controls for firm-specific characteristics (e.g. Anagnostopoulou and Tsekrekos, 2015; Fairhurst and Greene, 2022; Levi et al., 2014; Murdock and Madura, 2011), such as return on assets (variable *ROA*) and leverage (variable *Leverage*), as companies with negative or low profitability and higher debt tend to issue an official announcement for seeking for a buyer more often (Anagnostopoulou and Tsekrekos, 2015; Chen and Findlay, 2003; Aktas et al., 2010). Previous studies consider firm size (variable *SizeFirm*) as an important control factor in M&A deals as larger firms naturally present a more formidable defence against takeovers due to the increased resource requirements necessary for their acquisition (Masulis et al., 2007). Moreover, larger companies are usually acquired by hostile acquisition (Schwert, 2000), which is not related to seeking buyer firms that are more often characterised by smaller size (Anagnostopoulou and Tsekrekos, 2013). We also control for company age (variable *AgeFirm*), since Levi et al. (2014) claim that older firms are less likely to initiate an acquisition. Furthermore, we control for document length (*DocumentLength*) in the models with ML measure, as we observe an increasing trend in the overall length of 10-K forms over the years. Finally, this study incorporates controls for earnings management (*EarningsManagement*), in response to prior research findings that indicate a heightened tendency for seeking buyer firms to engage in accruals-based earnings manipulation prior to the formal announcement (Anagnostopoulou and Tsekrekos, 2013; Anagnostopoulou and Tsekrekos, 2015). All the models also include time and industry -fixed effects. The variables are described in detail in Appendix A.

We have opted to use t-1 controls in all our models. By employing lagged values of these controls, we aim to mitigate potential endogeneity issues. Furthermore, we hypothesize that the control variables affect the dependent variable primarily through their values in the previous period, similar to the independent variables. Thus, using t-1 values is deemed more appropriate for our analysis.

3.2 Involvement in ESG Activities and the Acquisition Outcome

To evaluate our hypothesis of whether the involvement of seeking buyer firms in ESG activities leads to a higher likelihood of being acquired, we estimate the following logistic regression model:

$$AcquisitionOutcome_{it} = \beta_0 + \beta_1 ESG_{it-1} + \beta_2 SizeFirm_{it-1} + \beta_3 ROA_{it-1} + \beta_4 Leverage_{it-1} + \beta_5 AgeFirm_{it-1} + \beta_6 EarningsManagement_{it-1} + \beta_7 DocumentLength_{it-1} + \beta_8 SalesGrowth_{it-1} + \varepsilon_{it} \quad (2)$$

The main explanatory variables are one of the measures reflecting involvement in ESG activities – disclosure or performance before the official announcement, and the dependent variable is *AcquisitionOutcome* – equal to 1 if a seeking buyer company was acquired in 2 years after issuing an official announcement. We incorporate the same firm-specific attributes as utilized in our preceding model, recognizing their integral connection to the dynamics of the seeking buyer firm's status (e.g. Arouri et al., 2019; Aktas et al., 2011; Deng et al., 2013; Fairhurst and Greene, 2022; Levi et al., 2014). Additionally, a company's financial characteristics—specifically, profitability and leverage—directly influence the likelihood of an acquisition (Anagnostopoulou and Tsekrekos, 2015; Deng et al., 2013; Masulis et al., 2007). Higher profitability may influence different deal outcomes (Datta et al., 2001; Deng et al., 2013; Levi et al., 2014), while higher leverage may add additional pressure on managers to perform well, making them spend more time and energy on making the deal a success (Jensen, 1986). Moreover, targets with higher leverage usually show higher eagerness to sell and close the deal (Aktas et al., 2010), which makes controlling for these characteristics important in our study.

4. Data and Sample Selection

4.1 ESG Disclosure and Performance Proxies

Prior studies have highlighted inconsistencies in the measurement of ESG metrics, as evidenced in the works of Berg et al. (2022), Christensen et al. (2022), Kimbrough et al. (2022), and Serafeim and Yoon (2022). In light of this, our research employs a comprehensive approach to evaluate the engagement of companies' ESG practices. This approach includes the use of several proxies, of both disclosure and performance, to provide a more comprehensive assessment.

We employ the Bloomberg ESG disclosure score as our first proxy to gauge a firm's engagement in ESG disclosure, a metric that has been validated and utilized in previous research (e.g., Buchanan et al., 2018; Christensen et al., 2022; Li et al., 2018). This disclosure score is based on information that companies report through annual reports, sustainability reports, corporate websites, and other sources. It ranges from 0.1 for companies with the minimum amount of disclosed ESG information to 100 for firms that disclose the complete set of indicators included in the Bloomberg ESG disclosure score.

However, Kimbrough et al. (2022) claim that ESG information provided by managers directly can decrease the existing disagreement in ESG indexes created by third parties, e.g. Bloomberg. Moreover, as the Bloomberg ESG disclosure score does not directly reflect the amount of ESG-related information in companies' reports, we draw insight from Li et al. (2021) to construct a purely reporting-based measure of

ESG disclosure. Li et al. (2021) used an ML textual approach, namely a word embedding model, to build a measure for corporate culture and we use a similar approach to create a proxy for ESG disclosure from annual reports.

Word embedding language models (Mikolov et al., 2013a, 2013b) can be used to make dictionaries more objective. This approach uses neural networks to model words or phrases with low-dimensional vectors (usually 100–300 elements) so that they preserve the semantic information of words. The networks are trained using extensive textual datasets, in our case 10-K forms, to efficiently learn word representations. This learning process focuses on predicting a central word based on surrounding words or vice versa, as established by Mikolov et al. (2013a). In the resulting vector space, words and phrases that frequently appear in similar contexts within 10-K forms exhibit proximate representations. This proximity in representation reflects their contextual and semantic associations, effectively quantifying these relationships. It represents words and phrases in a manner that aligns closely with human understanding, thereby facilitating a more intuitive interpretation of text. Consequently, this objective, data-driven method yields improved dictionaries tailored for 10-K forms, when compared to purely researcher-designed subjective dictionaries (Bhatia et al., 2021).

We briefly outline the process of constructing our machine learning model, with a more comprehensive description available in Li et al. (2021). The training dataset of the language model comprises 162,155 10-K forms sourced from the Edgar database (https://www.sec.gov/edgar/). Initially, we pre-process this data by removing extraneous elements such as HTML tags, and substituting numerical values with a designated symbol (#). Subsequently, we employ a pre-trained deep-learning algorithm from the spaCy library (www.spacy.io) to identify and label named entities within the text. These entities are annotated with the tag 'NER' followed by the entity type. The same model is also utilized to detect and categorize 'noun chunks' — phrases comprising nouns and their related words. In the final phase, we apply the word2vec neural network architecture, as described by Mikolov et al. (2013a), to develop word embedding representations for the identified phrases in the 10-K documents. This modelling is executed using the Gensim library (www.radimrehurek.com/gensim/).

The developed model is then utilized to identify 100 phrases that are contextually similar to our preselected seed words in 10-K forms, addressing each dimension of ESG independently. The phrases are presented in Appendix B. For the environmental dimension, the seed words are 'environmental' and 'sustainability'; for the social dimension, 'social' and 'our people'; and for governance, 'governance' and 'our_governance'. We compute the frequency of these identified words/phrases in the 10-K forms of both the target companies and their industry peers. This calculation serves as a metric for assessing the extent of ESG disclosure. Given our observation of an increasing trend in the overall length of 10-K reports, we opt for an absolute measure, as the relative measure could potentially misrepresent the actual extent of ESG disclosure.

Nevertheless, no textual analysis method is fool-proof, as is the case with our ML measure (Lewis and Young, 2019). While the vector space model, which encapsulates the semantic information of phrases, enhances the objectivity of our measure, the initial selection of seed words by researchers introduces a potential source of error. The ML-augmented dictionary might be biased right from the start if the seed words are inappropriately chosen. Therefore, the selection of these seed words necessitates particular attention. To mitigate this issue, we rigorously validated the chosen seed words through a consensus among all co-authors of this paper and an external accounting expert. Additionally, we established the robustness of our measure by juxtaposing the ML-derived results with those obtained from a manual coding of twenty 10-K forms, employing Cohen's Kappa as a statistical measure of agreement. This coding was performed by a seasoned accounting expert. The resultant Kappa value of 0.48 suggests a reasonably high level of reliability for our measure, particularly considering the inherent ambiguity in ESG factors.

Our third measure aims to estimate the ESG performance of companies. Although Michelon et al. (2020) claim that investors are more concerned about ESG disclosure rather than the real implementation of ESG performance, we anticipate the performance to be an important factor in M&A deals of seeking buyer firms. Our proxy of ESG performance is the LSEG ASSET4 rating that reflects environmental, social and governance performance metrics (e.g., Cheng et al., 2014; Dyck et al., 2019; Eliwa et al., 2021). LSEG's ESG framework goes beyond mere disclosure by incorporating advanced analytics and peer benchmarking, enabling a comprehensive evaluation of a company's real-world ESG impact in comparison to its industry peers (www.lseg.com). They incorporate both qualitative and quantitative data, including third-party verification and on-the-ground performance metrics. LSEG's ESG assessment is dynamic by nature, which involves regular updates and adjustments based on evolving standards and practices, ensuring that the focus remains on measurable and progressive ESG performance, rather than static or outdated disclosures. The scale of the ESG rating is from 0 to 100, where 0 is the worst ESG performance, while 100 characterises the best ESG performance.

4.2 Sample

We acquired the data of all US firms that officially announced seeking a buyer during the period of 2000 to 2021 from the LSEG Eikon M&A database (www.lseg.com). The database defines a seeking buyer firm as a case 'where the target company has announced plans to seek out a buyer or buyers for its assets or the company itself.' We only keep publicly listed firms in the sample due to data availability constraints for private companies.

To be included in our sample, a seeking buyer company needs to meet two requirements: the company needs to have issued a 10-K form (available in the SEC Edgar database) in the same fiscal year as the official announcement date, and its financial data needs to be available in LSEG Datastream. The final sample of seeking buyer companies includes 246 individual deals. The industry distribution of the deals based on primary SIC codes (US Securities and Exchange Commission, 2021) is provided in Table 1. The manufacturing sector has the highest representation, 70.9 % with 175 companies. This is followed by transportation and public utilities at 13.0 % (32 companies), services industry at 10.1 % (27 companies), wholesale and retail trade at 2.4 % (6 companies), and mining at 2.4 % (6 companies).

<TABLE 1 SHOULD BE INSERTED AROUND HERE>

4.3 Descriptive Statistics

Table 2 provides descriptive statistics for the main variables. We use propensity score matching to collect a matching sample of companies for further comparative analysis. Propensity score matching is a technique that is commonly used in accounting research to collect a control group with similar characteristics to the test group (Cram et al., 2010; Kothari et al., 2005). We base the propensity score matching on financial characteristics, namely, market value, revenues, total assets, EBIT and current liabilities. The data requirements are the same as for the seeking buyer companies: availability of financial data and 10-K filings from the SEC. The final sample consists of 233 matching companies. The data for the matching companies is also year-matched to each deal.

The 'Seeking buyer' sample generally shows higher Bloomberg ESG and ML ESG disclosure scores, and LSEG ESG performance scores compared to the matching sample. In terms of firm characteristics, the 'Seeking buyer' sample has higher average total assets, ROA, and age. There is no significant difference between leverage, sales growth and discretionary accruals.

<TABLE 2 SHOULD BE INSERTED AROUND HERE>

Table 3 presents the correlations for the variables. There is a strong positive correlation between the Bloomberg ESG disclosure score and the LSEG ESG performance score, suggesting similar tendencies in these scores. However, the ML ESG disclosure score shows weaker and mixed correlations with the other two. In terms of financial and operational metrics, company size is positively correlated with Bloomberg and LSEG ESG scores but negatively correlated with the ML ESG disclosure score. Company age also shows a positive, albeit weaker, relationship with Bloomberg and LSEG ESG scores. Leverage, on the other hand, has very weak correlations with all ESG scores, indicating minimal influence. The earnings management variable has weak correlations with the ESG scores, except a slightly more distinguishable positive correlation with the ML ESG score.

<TABLE 3 SHOULD BE INSERTED AROUND HERE>

5. Results

5.1 Involvement of Seeking Buyer Companies in ESG Activities

We estimate equation (1) as a logistic panel regression model. The results are provided in Table 4. The dependent variable *SeekingBuyer* has a statistically significant and positive relationship to the main explanatory variable, *ESG,* for the ML ESG measure. The relationship is not statistically significant for Bloomberg ESG and LSEG ESG.

<TABLE 4 SHOULD BE INSERTED AROUND HERE>

Thus, our findings indicate that seeking buyer companies are more involved in ESG disclosure than their industry peers, tentatively supporting our H1.1. However, as the ESG performance of seeking-buyer companies does not differ from their peers, our results indicate that seeking-buyer companies might use enhanced ESG disclosure just for 'window dressing' to improve their legitimacy in the market (Eliwa et al., 2021; Hummel and Schlick, 2016).

Reflecting on these results, H1.2, on the other hand, is rejected as our results do not support it. In line with signalling theory (Dhaliwal et al., 2011; Macias et al., 2011; Spence, 1973; Wong and Zhang, 2021), higher ESG disclosure should be a signal of corresponding higher ESG performance. Since there is no significant difference in ESG performance, the higher ML ESG disclosure does not likely serve as a high-quality signal of the company's ESG involvement in the seeking buyer setting.

To get more support for these findings, we also examine the interactions between ESG performance and disclosure, reported in columns (4) and (5) of Table 4. A positive coefficient would suggest that high levels of ESG disclosure amplify the positive effects of strong ESG performance on the firm's propensity to seek acquisitions. Conversely, a negative coefficient would indicate that increased disclosure may not appear simultaneously to increased ESG performance, and vice versa. The last two columns of Table 4 provide the results of the empirical analysis. The primary interest variable is an interaction term designed to represent firm ESG practices, with the analysis pairing the effects of Bloomberg and ML ESG disclosure scores with the LSEG ASSET4 ESG performance score. The ESG performance and disclosure exhibit a positive relationship with the status of being a seeking buyer company in both columns. The interaction term between the Bloomberg ESG score and the LSEG ESG score on the other hand, is negative and statistically significant at the 10% level. In addition, the interaction term between the ML ESG score and the LSEG ESG score is not significant. The results strengthen the view that simultaneous higher ESG disclosure and performance scores are less likely to be associated with a firm being in the seeking buyer category, and the interaction between ESG disclosure and performance is negative, thereby, emphasizing a 'high disclosure – low performance' observation. This result supports the hypothesis that ESG disclosure would be a legitimacy building action for the seeking buyer companies, rather than a signal of the companies' ESG involvement, further supporting the acceptance of hypothesis H1.1, but rejection of H1.2.

5.2 Involvement in ESG Activities and the Likelihood of Being Acquired

To test our second set of hypotheses, we estimate equation (2) as a logistic regression model. We report the regression results of the relationship between seeking buyers' involvement in ESG practices, both disclosure and performance, and the acquisition outcome in Table 5.

<TABLE 5 SHOULD BE INSERTED AROUND HERE>

We find no statistically significant association with any of the ESG metrics and the likelihood of being acquired. These results reflect that acquirers might consider ESG disclosure by seeking buyer firms as 'cheap talk' (Delmas and Burbano, 2011) that is used for 'window dressing' (Macias et al., 2011) if it is not properly supported by clear ESG performance outcomes. Indeed, our results reflect that in some models there is a positive and significant association between such financial characteristics, as average sales growth and leverage, and the acquisition outcome, indicating that these factors are considered more important when making acquisition decisions. This outcome is consistent with the results of previous research reflecting that these financial characteristics are positively related to the deal outcomes (Harrison et al., 2014; Datta et al., 2001; Aktas et al., 2010). We conduct a similar analysis using individual ESG dimensions reported in Appendix C.

6. Additional Analysis

To get more verification for the opposite roles of ESG performance and disclosure characterising seeking buyer companies, we repeat the analysis for the Bloomberg measures using a 2SLS model to diminish endogeneity issues. This analysis utilizes logit regression where the dependent variable is binary, indicating whether a firm is a part of the seeking buyer sample or not. The main explanatory variable, 'ESG hat', is estimated using 2SLS as an instrumental variable to address potential endogeneity issues. The chosen instrument is the year when a company's ESG score is first reported in the Bloomberg database (Chen and Xie, 2022), where an earlier year would instrument higher ESG disclosure and be negatively correlated with the Bloomberg ESG disclosure variable. We test the first hypothesis with the equation:

$$\widehat{SeekingBuyer}_{it} = \hat{\beta}_0 + \hat{\beta}_1 ESGDisclosure_{it-1} + \hat{\beta}_2 SizeFirm_{it-1} + \hat{\beta}_3 ROA_{it-1} + \hat{\beta}_4 Leverage_{it-1} + \hat{\beta}_5 AgeFirm_{it-1} + \hat{\beta}_6 EarningsManagement_{it-1} \quad (3)$$

The empirical results are provided in Table 6. The key finding is that the 'ESG hat' variable has a negative and statistically significant relationship at 10% level with a firm being in the seeking buyer category. This suggests that higher ESG disclosure, as measured and instrumented in this study with an *earlier* year of being reported in the database, hence resulting in a negative coefficient, is associated with an increased likelihood of a firm being categorized as actively seeking a buyer. Thus, the results support the findings of equation (1) and the conclusions presented in section 5.1.

<TABLE 6 SHOULD BE INSERTED AROUND HERE >

7. Conclusions

Existing literature exploring the relationship between a company's engagement in ESG activities and the outcomes of M&A is relatively scarce and has examined the effect of ESG involvement on announcement returns (Deng et al., 2013), takeover premiums (Gomes and Marsat, 2018), empire building (Gul et al., 2020) and acquisition decisions (Bose et al., 2021; Boone and Uysal, 2020). Furthermore, all the existing studies have been focused on M&A deals where acquiring companies initiated the deals.

We extend academic literature by examining the involvement in ESG activities of firms that officially issue an announcement that they seek a buyer. Consistent with previous research, (e.g. Anagnostopoulou and Tsekrekos, 2013; Anagnostopoulou and Tsekrekos, 2015) we observe that seeking-buyer firms have both time and motives for the actions that may facilitate a deal, e.g. involvement in enhanced ESG disclosure, as these actions may help gain legitimacy, improve the company's image on the market and send positive signals to potential buyers.

Our results reflect that seeking buyer companies disclose more ESG information in comparison to the matching peers based on the ML ESG disclosure measure. However, their ESG performance in comparison to the matched sample does not significantly differ and the results support the notion that neither ESG disclosure variable is complementary to ESG performance within the seeking buyer sample. Therefore, this enhanced ESG disclosure by seeking buyer firms can be characterised as merely 'cheap talk' (Delmas and Burbano, 2011). Another possible explanation is that the companies are engaging in more disclosure to avoid litigation, as previous research has found that strong disclosure and the disclosure of bad news alike decreases litigation risk (Hanley & Hoberg, 2012).

Furthermore, we find no evidence that enhanced ESG practices are associated with the likelihood of being acquired. Thus, this indicates that acquiring companies do not buy the seeking buyers' enhanced ESG disclosure and might consider it as 'window dressing' (Macias et al., 2011; Delmas and Burbano, 2011), as a waste of time and resources (Bénabou and Tirole, 2010; Masulis and Reza, 2015) or as social activities that just promote managers' personal reputation as good citizens (Barnea and Rubin, 2010; Cheng et al., 2013). Furthermore, our results indicate that acquiring companies do not consider ESG performance to be relevant in their decision-making. We conclude that the acquiring firms may place greater importance on the financial characteristics when making acquisition decisions, as opposed to ESG activities, which can be challenging to comprehend and interpret (Berg et al., 2022; Christensen et al., 2022; Kimbrough et al., 2022).

Our study offers valuable contributions to the field. From the theoretical point of view, our research contributes to a deeper comprehension of corporate disclosures made by firms seeking buyers, an area that has not been extensively explored in existing literature. From the practical perspective, we suggest that acquiring companies verify the real integration of ESG in a target company's strategy. Only in this case, extensive disclosure of ESG may reflect the actual adoption of ESG that in the future will lead to positive material effects and not whistle-blowing. The findings of our study offer also valuable insights with practical implications for policymakers and regulators in their decision-making processes. First, some regulatory bodies (e.g., the SEC and the European Commission) have already issued recommendations on disclosing ESG-related information for publicly listed companies. The results of this study are also in line with Directive 2014/95/EU and the proposed Corporate Sustainability Reporting Directive (CSRD). These regulations will further integrate the use of involvement in ESG practices in decision-making by different stakeholders. Our study contributes to the ongoing discussion regarding the critical need for standardizing ESG reporting, as our findings underscore the significant influence of companies' engagement in ESG activities on the outcomes of M&A deals.

However, this study is not without limitations, thus pointing to potential future research directions. First, we only consider seeking buyer companies from one country – the US. Future research could explore the behaviours of firms seeking buyers in various other markets. Studies may also explore whether different factors and institutions influence the relationship between involvement in ESG activities and the likelihood of being acquired as country-level investor protection and potential regulatory oversight are important factors that might influence ESG disclosure (Hummel et al., 2024) and M&A outcomes (Bose et al., 2021). There might be some differences between different industries as some are considered 'greener' than others making industry-level analysis an interesting avenue for future M&A research. The study does not offer an exhaustive explanation on the motivations of seeking buyer companies to disclose more ESG information compared to their peers. A future research path would be to delve into the quality of the ESG disclosures.

Second, our conclusions about the motivations and preferences of acquiring firms must be interpreted with caution, given that we do not control for the specific interests of buyers in ESG issues, or whether there are ESG-driven motivations behind the acquisition decisions. The lack of association between ESG performance and the likelihood of acquisition does not necessarily imply a disinterest in ESG dimensions per se, but rather reflects a complex interplay of ESG and financial considerations. Indeed, the statistical significance of financial outcomes, such as sales growth and leverage, suggests that these factors carry considerable weight in acquisition decisions. Consequently, a comprehensive analysis that integrates both ESG posture and financial variables of both seeking buyer companies and potential buyers is essential for understanding the dynamics of seeking buyer behaviour and acquisition outcomes. Such an approach might reveal alternative explanations, suggesting that while ESG factors are disclosed, they may not be pivotal in the final acquisition decision unless accompanied by strong financial performance.

Furthermore, the scope of this research was limited to examining seeking buyers' involvement in ESG practices prior to the announcement, and the impact of these practices on the likelihood of acquisition. Expanding the analysis to include other dimensions of M&A transactions – such as deal value, payment methods, and the resultant synergy between the acquiring and acquired firms – would enable a more holistic examination of M&As.

References

Aktas, N., De Bodt, E., & Cousin, J. G. (2011). Do financial markets care about SRI? Evidence from mergers and acquisitions. *Journal of Banking & Finance*, *35*(7), 1753-1761.

Aktas, N., De Bodt, E., & Roll, R. (2010). Negotiations under the threat of an auction. *Journal of Financial Economics, 98*(2), 241–255.

Alexandridis, G., Petmezas, D., & Travlos, N. G. (2010). Gains from mergers and acquisitions around the world: New evidence. *Financial Management*, *39*(4), 1671-1695.

Alhadab, M., Clacher, I., & Keasey, K. (2015). Real and accrual earnings management and IPO failure risk. *Accounting and Business research*, *45*(1), 55-92.

Anagnostopoulou, S. C., & Tsekrekos, A. E. (2013). Do firms that wish to be acquired manage their earnings? Evidence from major European countries. *International Review of Financial Analysis*, 30, 57-68.

Anagnostopoulou, S. C., & Tsekrekos, A. E. (2015). Earnings management in firms seeking to be acquired. *The British Accounting Review*, *47*(4), 351-375.

Arouri, M., Gomes, M., & Pukthuanthong, K. (2019). Corporate social responsibility and M&A uncertainty. *Journal of Corporate Finance*, *56*, 176-198.

Ashforth, B. E., & Gibbs, B. W. (1990). The double-edge of organisational legitimation. *Organization Science*, *1*(2), 177-194.

Barnea, A., & Rubin, A. (2010). Corporate social responsibility as a conflict between shareholders. *Journal of Business Ethics*, *97*, 71-86.

Bénabou, R., & Tirole, J. (2010). Individual and corporate social responsibility. *Economica*, *77*(305), 1-19.

Bereskin, F., Byun, S. K., Officer, M. S., & Oh, J. M. (2018). The effect of cultural similarity on mergers and acquisitions: Evidence from corporate social responsibility. *Journal of Financial and Quantitative Analysis*, *53*(5), 1995-2039.

Berg, F., Koelbel, J. F., & Rigobon, R. (2022). Aggregate confusion: The divergence of ESG ratings. *Review of Finance*, *26*(6), 1315-1344.

Bhatia, S., Olivola, C. Y., Bhatia, N., & Ameen, A. (2021). Predicting leadership perception with large-scale natural language data. *The Leadership Quarterly*, 101535.

Bilyay-Erdogan, S., Danisman, G. O., & Demir, E. (2023). ESG Performance and Investment Efficiency: The Impact of Information Asymmetry. *Journal of International Financial Markets, Institutions and Money*, 101919.

Boone, A. L., & Mulherin, J. H. (2008). Do auctions induce a winner's curse? New evidence from the corporate takeover market. *Journal of Financial Economics, 89*(1), 1-19.

Boone, A., & Uysal, V. B. (2020). Reputational concerns in the market for corporate control. *Journal of Corporate Finance*, *61*, 101399.

Bose, S., Minnick, K., & Shams, S. (2021). Does carbon risk matter for corporate acquisition decisions?. *Journal of Corporate Finance*, *70*, 102058.

Brooks, C., & Oikonomou, I. (2018). The effects of environmental, social and governance disclosures and performance on firm value: A review of the literature in accounting and finance. *The British Accounting Review*, *50*(1), 1-15.

Brown, N., & Deegan, C. (1998). The public disclosure of environmental performance information—a dual test of media agenda setting theory and legitimacy theory. *Accounting and Business Research, 29*(1), 21–41.

Brown, W. O., Helland, E., & Smith, J. K. (2006). Corporate philanthropic practices. *Journal of Corporate Finance*, *12*(5), 855-877.

Buchanan, B., Cao, C. X., & Chen, C. (2018). Corporate social responsibility, firm value, and influential institutional ownership. *Journal of Corporate Finance*, *52*, 73-95.

Chen, C., & Findlay, C. (2003). A Review of Cross-border Mergers and Acquisitions in APEC. *Asian-Pacific Economic Literature*, *17*(2), 14-38.

Chen, Z. & Xie, G. (2022). ESG disclosure and financial performance: Moderating role of ESG investors, *International Review of Financial Analysis*, 83, 102291,

Cheng, B., Ioannou, I., & Serafeim, G. (2014). Corporate social responsibility and access to finance. *Strategic Management Journal*, *35*(1), 1-23.

Cheng, H., Hong, H., & Shue, K. (2013). *Do managers do good with other people's money?* (No. w19432). National Bureau of Economic Research. Dartmouth College, Princeton University, and University of Chicago.

Cho, C. H., Guidry, R. P., Hageman, A. M., & Patten, D. M. (2012). Do actions speak louder than words? An empirical investigation of corporate environmental reputation. *Accounting, Organizations and Society*, *37*(1), 14-25.

Cho, S. Y., Lee, C., & Pfeiffer Jr, R. J. (2013). Corporate social responsibility performance and information asymmetry. *Journal of Accounting and Public Policy*, *32*(1), 71-83.

Christensen, D. M., Serafeim, G., & Sikochi, A. (2022). Why is corporate virtue in the eye of the beholder? The case of ESG ratings. *The Accounting Review*, *97*(1), 147-175.

Clarkson, P. M., Li, Y., Richardson, G. D., & Vasvari, F. P. (2008). Revisiting the relation between environmental performance and environmental disclosure: An empirical analysis. *Accounting, Organizations and Society*, *33*(4-5), 303-327.

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*(1), 3-34.

Cohen, D. A., & Zarowin, P. (2010). Accrual-based and real earnings management activities around seasoned equity offerings. *Journal of Accounting and Economics*, *50*(1), 2-19.

Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2010). Signaling Theory: A Review and Assessment. Journal of Management, 37(1), 39–67.

Cram, D. P., Karan, V., & Stuart, I. (2009). Three threats to validity of choice-based and matched-sample studies in accounting research. *Contemporary Accounting Research*, 26(2), 477-516.

Cui, J., Jo, H., & Na, H. (2018). Does corporate social responsibility affect information asymmetry?. *Journal of Business Ethics*, *148*(3), 549-572.

Datta, S., Iskandar-Datta, M., & Raman, K. (2001). Executive compensation and corporate acquisition decisions. *The Journal of Finance*, *56*(6), 2299-2336.

Dechow, P. M., Sloan, R. G., & Sweeney, A. P. (1995). Detecting earnings management. *Accounting Review*, *70*(2), 193-225.

De Villiers, C., & Marques, A. (2016). Corporate social responsibility, country-level predispositions, and the consequences of choosing a level of disclosure. *Accounting and Business Research, 46(*2), 167-195.

Deegan, C., Rankin, M., & Tobin, J. (2002). An examination of the corporate social and environmental disclosures of BHP from 1983-1997: A test of legitimacy theory. *Accounting, Auditing & Accountability Journal, 15*(3), 312–343.

Deegan, C. and Unerman, J. (2011). EBOOK: Financial Accounting Theory: European Edition. McGraw Hill.

Deephouse, D. L. (1996). Does isomorphism legitimate?. *Academy of Management Journal*, *39*(4), 1024-1039.

DeFond, M. L., & Jiambalvo, J. (1994). Debt covenant violation and manipulation of accruals. *Journal of Accounting and Economics, 17*(1-2), 145-176.

Delmas, M. A., & Burbano, V. C. (2011). The drivers of greenwashing. *California Management Review*, *54*(1), 64-87.

Deng, X., Kang, J. K., & Low, B. S. (2013). Corporate social responsibility and stakeholder value maximisation: Evidence from mergers. *Journal of Financial Economics*, *110*(1), 87-109.

Dhaliwal, D. S., Li, O. Z., Tsang, A., & Yang, Y. G. (2011). Voluntary nonfinancial disclosure and the cost of equity capital: The initiation of corporate social responsibility reporting. *The Accounting Review*, *86*(1), 59-100.

Dhaliwal, D. S., Radhakrishnan, S., Tsang, A., & Yang, Y. G. (2012). Nonfinancial disclosure and analyst forecast accuracy: International evidence on corporate social responsibility disclosure. *The Accounting Review*, *87*(3), 723-759.

Dimson, E., Karakaş, O., & Li, X. (2015). Active ownership. *The Review of Financial Studies*, *28*(12), 3225-3268.

Donaldson, T., & Preston, L. E. (1995). The stakeholder theory of the corporation: Concepts, evidence, and implications. *Academy of Management Review*, *20*(1), 65-91.

Dowling, J., & Pfeffer, J. (1975). Organisational legitimacy: Social values and organisational behavior. *Pacific Sociological Review*, *18*(1), 122-136.

Dunn, P., & Sainty, B. (2009). The relationship among board of director characteristics, corporate social performance and corporate financial performance. *International Journal of Managerial Finance*, *5*(4), 407-423.

Dyck, A., Lins, K. V., Roth, L., & Wagner, H. F. (2019). Do institutional investors drive corporate social responsibility? International evidence. *Journal of Financial Economics*, *131*(3), 693-714.

Eliwa, Y., Aboud, A., & Saleh, A. (2021). ESG practices and the cost of debt: Evidence from EU countries. *Critical Perspectives on Accounting*, *79*, 102097.

Erickson, M., & Wang, S. W. (1999). Earnings management by acquiring firms in stock for stock mergers. *Journal of Accounting and Economics*, *27*(2), 149-176.

Erragragui, E. (2018). Do creditors price firms’ environmental, social and governance risks?. *Research in International Business and Finance*, *45*, 197-207.

Fairhurst, D. D., & Greene, D. T. (2022). Too much of a good thing? Corporate social responsibility and the takeover market. *Journal of Corporate Finance*, *73*, 102172.

Fatemi, A., Glaum, M., & Kaiser, S. (2018). ESG performance and firm value: The moderating role of disclosure. *Global Finance Journal*, *38*, 45-64.

Feng, J., Goodell, J. W., & Shen, D. (2022). ESG rating and stock price crash risk: Evidence from China. *Finance Research Letters*, *46*, 102476.

Gomes, M. (2019). Does CSR influence M&A target choices?. *Finance Research Letters*, *30*, 153-159.

Gomes, M., & Marsat, S. (2018). Does CSR impact premiums in M&A transactions?. *Finance Research Letters*, *26*, 71-80.

Gray, R. (2002). The social accounting project and Accounting Organizations and Society Privileging engagement, imaginings, new accountings and pragmatism over critique?. *Accounting, Organizations and Society*, *27*(7), 687-708.

Gregory, A., Tharyan, R., & Whittaker, J. (2014). Corporate social responsibility and firm value: Disaggregating the effects on cash flow, risk and growth. *Journal of Business Ethics*, *124*(4), 633-657.

Gul, F. A., Krishnamurti, C., Shams, S., & Chowdhury, H. (2020). Corporate social responsibility, overconfident CEOs and empire building: Agency and stakeholder theoretic perspectives. *Journal of Business Research*, *111*, 52-68.

Hackbarth, D., & Morellec, E. (2008). Stock returns in mergers and acquisitions. *The Journal of Finance*, *63*(3), 1213-1252.

Hanley, K. W., & Hoberg, G. (2012). Litigation risk, strategic disclosure and the underpricing of initial public offerings. *Journal of Financial Economics*, *103*(2), 235–254. https://doi.org/10.1016/J.JFINECO.2011.09.006

Harrison, J. S., Hart, M., & Oler, D. K. (2014). Leverage and acquisition performance. *Review of Quantitative Finance and Accounting*, *43*, 571-603.

Hemingway, C. A., & Maclagan, P. W. (2004). Managers' personal values as drivers of corporate social responsibility. *Journal of Business Ethics*, *50*(1), 33-44.

Hsu, A., Koh, K., Liu, S., & Tong, Y. H. (2019). Corporate social responsibility and corporate disclosures: An investigation of investors' and analysts' perceptions. *Journal of Business Ethics*, *158*(2), 507-534.

Hummel, K., & Schlick, C. (2016). The relationship between sustainability performance and sustainability disclosure–Reconciling voluntary disclosure theory and legitimacy theory. *Journal of Accounting and Public Policy, 35*(5), 455-476.

Hummel, K., Mittelbach-Hoermanseder, S., Cho, C. H., & Matten, D. (2024). Corporate social responsibility disclosure: a topic-based approach. *Accounting and Business Research*, *54*(1), 87-124.

Hussain, T., & Shams, S. (2022). Pre-deal differences in corporate social responsibility and acquisition performance. *International Review of Financial Analysis*, *81*, 102083.

Jensen, M. C. (1986). Agency costs of free cash flow, corporate finance, and takeovers. *The American Economic Review*, *76*(2), 323-329.

Jensen, M. C. (2002). Value maximisation, stakeholder theory, and the corporate objective function. *Business Ethics Quarterly*, 235-256.

Kau, J. B., Linck, J. S., & Rubin, P. H. (2008). Do managers listen to the market?. *Journal of Corporate Finance*, *14*(4), 347-362.

Kim, B. J., Jung, J. Y., & Cho, S. W. (2021). Can ESG mitigate the diversification discount in cross-border M&A?. *Borsa Istanbul Review.* In press.

Kim, Y., Park, M. S., & Wier, B. (2012). Is earnings quality associated with corporate social responsibility?. *The Accounting Review*, *87*(3), 761-796.

Kimbrough, M. D., & Louis, H. (2011). Voluntary disclosure to influence investor reactions to merger announcements: An examination of conference calls. *The Accounting Review*, *86*(2), 637-667.

Kimbrough, M. D., Wang, X., Wei, S., & Zhang, J. (2022). Does voluntary ESG reporting resolve disagreement among ESG rating agencies?. *European Accounting Review*, 1-33.

Kothari, S. P., Leone, A. J., & Wasley, C. E. (2005). Performance matched discretionary accrual measures. *Journal of Accounting and Economics*, 39(1), 163-197.

Kothari, S. P., Mizik, N., & Roychowdhury, S. (2016). Managing for the moment: The role of earnings management via real activities versus accruals in SEO valuation. *The Accounting Review*, *91*(2), 559-586.

Kruger, P. (2015). Corporate goodness and shareholder wealth. *Journal of Financial Economics*, *115*(2), 304-329.

Levi, M., Li, K., & Zhang, F. (2014). Director gender and mergers and acquisitions. *Journal of Corporate Finance*, *28*, 185-200.

Levitt, T. (1958). The dangers of social-responsibility. *Harvard Business Review*, *36*(5), 41-50.

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.

Li, K., Mai, F., Shen, R., & Yan, X. (2021). Measuring Corporate Culture Using Machine Learning. *The Review of Financial Studies*. *34*(7), 3265-3315.

Li, Y., Gong, M., Zhang, X. Y., & Koh, L. (2018). The impact of environmental, social, and governance disclosure on firm value: The role of CEO power. *The British Accounting Review*, *50*(1), 60-75.

Lin, C. Y. Y., & Wei, Y. C. (2006). The role of business ethics in merger and acquisition success: An empirical study. *Journal of Business Ethics*, *69*(1), 95-109.

Luo, Y. (2005). Do insiders learn from outsiders? Evidence from mergers and acquisitions. *The Journal of Finance*, *60*(4), 1951-1982.

Lys, T., Naughton, J. P., & Wang, C. (2015). Signaling through corporate accountability reporting. *Journal of Accounting and Economics*, *60*(1), 56-72.

Macias, A. J., Anilowski C., & Sanchez, J.M. (2011) *Can Targets Benefit from Negotiations? Evidence from Auctions and Negotiations*. Available online: https://ssrn.com/abstract=1787313.

Masulis, R. W., & Reza, S. W. (2015). Agency problems of corporate philanthropy. *The Review of Financial Studies*, *28*(2), 592-636.

Masulis, R. W., Wang, C., & Xie, F. (2007). Corporate governance and acquirer returns. *The Journal of Finance, 62*(4), 1851-1889.

Michelon, G., Pilonato, S., Ricceri, F., & Roberts, R. W. (2016). Behind camouflaging: Traditional and innovative theoretical perspectives in social and environmental accounting research. *Sustainability Accounting, Management and Policy Journal*, *7*(1), 2-25.

Michelon, G., Rodrigue, M., & Trevisan, E. (2020). The marketisation of a social movement: Activists, shareholders and CSR disclosure. *Accounting, Organizations and Society*, *80*, 101074.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013a). Efficient Estimation of Word Representations in Vector Space. *ArXiv:1301.3781 [Cs]*. http://arxiv.org/abs/1301.3781

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013b). Distributed Representations of Words and Phrases and their Compositionality. *ArXiv:1310.4546 [Cs, Stat]*. http://arxiv.org/abs/1310.4546

Mulherin, H., & Simsir, A.S. (2015). Measuring deal premiums in takeovers. *Financial Management*, *44*(1), 1–14.

Murdock, M. (2010). *Essays on Corporate Restructuring*. Florida Atlantic University. Available online: http://fau.digital.flvc.org/islandora/object/fau%3A3592/datastream/OBJ/view/Essays_in_corporate_restructuring.pdf

Murdock, M., & Madura, J. (2011). Should prospective target firms publicly seek buyers. *Working Paper, Florida Atlantic University*.

Nekhili, M., Boukadhaba, A., & Nagati, H. (2021). The ESG–financial performance relationship: Does the type of employee board representation matter?. *Corporate Governance: An International Review*, *29*(2), 134–161.

Neu, D., Warsame, H., & Pedwell, K. (1998). Managing public impressions: environmental disclosures in annual reports. *Accounting, Organizations and Society*, *23*(3), 265-282.

Onukar, S. R. (2021). Untangling ESG in Private Equity: Do ESG Commitments Affect Target Firm Choice in LBOs? Wharton Working Paper. Available online: https://repository.upenn.edu/cgi/viewcontent.cgi?article=1228&context=wharton_research_scholars

Patten, D. M. (1992). Intra-industry environmental disclosures in response to the Alaskan oil spill: A note on legitimacy theory. *Accounting, Organizations and Society, 17*(5), 471–475. https://doi.org/10.1016/0361-3682(92)90042-Q

Ramchander, S., Schwebach, R. G., & Staking, K. I. M. (2012). The informational relevance of corporate social responsibility: Evidence from DS400 index reconstitutions. *Strategic Management Journal*, *33*(3), 303–314.

Rezaee, Z. (2016). Business sustainability research: A theoretical and integrated perspective. *Journal of Accounting literature*, *36*(1), 48–64.

Roberts, J. (2003). The manufacture of corporate social responsibility: Constructing corporate sensibility. *Organisation*, *10*(2), 249–265.

Russo, M. V., & Fouts, P. A. (1997). A resource-based perspective on corporate environmental performance and profitability. *Academy of management Journal*, *40*(3), 534–559.

Schwert, G. W. (2000). Hostility in takeovers: in the eyes of the beholder?. *The Journal of Finance*, 55(6), 2599–2640.

Serafeim, G., & Yoon, A. (2022). Stock price reactions to ESG news: The role of ESG ratings and disagreement. *Review of Accounting Studies*, 1-31.

Shrivastava, P. (1995). Ecocentric management for a risk society. *Academy of management review*, *20*(1), 118–137.

Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87(3), 355–374. https://doi.org/10.2307/1882010

Stolowy, H., & Paugam, L. (2018). The expansion of non-financial reporting: an exploratory study. *Accounting and Business Research*, *48*(5), 525–548.

Suchman, M. C. (1995). Managing legitimacy: Strategic and institutional approaches. *Academy of Management Review*, *20*(3), 571–610.

Talpur, S., Nadeem, M., & Roberts, H. (2023). Corporate social responsibility decoupling: a systematic literature review and future research agenda. *Journal of Applied Accounting Research*. Published-ahead-of-print.

Teoh, S. H., Welch, I., & Wong, T. J. (1998). Earnings management and the long-run market performance of initial public offerings. *The Journal of Finance*, *53*(6), 1935-1974.

Tregidga, H., & Laine, M. (2022). On crisis and emergency: Is it time to rethink long-term environmental accounting?. *Critical Perspectives on Accounting*, *82*, 102311.

Tsang, A., Frost, T., & Cao, H. (2023). Environmental, social, and governance (ESG) disclosure: A literature review. *The British Accounting Review*, *55*(1), 101149.

Tunyi, A. A. (2021). Revisiting acquirer returns: Evidence from unanticipated deals. *Journal of Corporate Finance*, *66*, 101789.

US Securities and Exchange Commission. (2021). *Division of Corporation Finance: Standard Industrial Classification (SIC) Code List.* Available online:

https://www.sec.gov/corpfin/division-of-corporation-finance-standard-industrial-classification-sic-code-list

Waldman, D. A., Siegel, D. S., & Javidan, M. (2006). Components of CEO transformational leadership and corporate social responsibility. *Journal of Management Studies*, *43*(8), 1703-1725.

Wong, J. B., & Zhang, Q. (2021). Stock market reactions to adverse ESG disclosure via media channels. *The British Accounting Review*, 101045.

Tables

TABLE 1

Industry distribution of seeking buyer companies

Industry distribution	N. of companies	%
Mining	6	2.4 %
Manufacturing	175	70.9 %
Transportation and public utilities	32	13.0 %
Wholesale and Retail Trade	6	2.4 %
Services	27	10.1 %
Total	246	100.00 %

TABLE 2

Descriptive Statistics

The descriptive statistics of the ESG variables for the 'Seeking buyer' sample and the matching sample. All ESG variables and control variables are lagged at t-1. Welch's t-test due to unequal variances is reported in the last column. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively. All variables are described in Appendix A.

	Full sample			Seeking buyer			Matching sample			t-Test	
Variable	Obs.	Mean	Std. Dev	Obs	Mean	Std. Dev	Obs	Mean	Std. Dev	Diff.	p-value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(5)-(8)	
ESG Variables											
Bloomberg ESG	276	44.222	13.589	147	46.798	13.412	129	41.226	13.026	5.572	***
ML ESG	366	0.006	0.005	159	0.007	0.006	207	0.005	0.004	0.002	***
LSEG ESG	443	37.787	31.106	235	43.041	30.732	208	31.845	30.510	11.20	***
Firm Characteristics											
SizeFirm	479	15.730	2.822	246	16.490	2.140	233	14.911	3.258	1.580	***
ROA	479	5.844	38.782	246	13.053	29.546	233	-1.824	46.168	14.877	***
Leverage	479	0.286	0.196	246	0.298	0.179	233	0.273	0.210	0.025	-
AgeFirm	479	1.599	0.318	246	1.634	0.320	233	1.562	0.312	0.073	**
EarningsManagement	479	-0.028	0.107	246	-0.029	0.114	233	-0.027	0.097	-0.002	-
AverageSalesGrowth	427	26.799	70.219	220	24.373	58.648	207	27.752	73.757	-3.378	-
DocumentLength	366	12.731	0.817	159	12.953	0.586	207	12.568	0.899	0.385	***
AcquisitionOutcome				246	0.199	0.400					

TABLE 3

Correlations

Correlations of the ESG variables and control variables for the 'Seeking buyer' sample. All ESG variables and control variables are lagged at t-1. All variables are described in Appendix A.

	ML ESG	Bloomberg ESG	LSEG ESG	Leverage	SizeFirm	AgeFirm	ROA	Earnings Managem ent	Average Sales Growth	Acquisitio n Outcome	Document Length
ML ESG	1.000										
Bloomberg ESG	-0.182	1.000									
LSEG ESG	0.086	0.795	1.000								
Leverage	0.160	0.134	-0.074	1.000							
SizeFirm	-0.218	0.580	0.579	0.016	1.000						
AgeFirm	-0.127	0.083	0.138	-0.126	0.242	1.000					
ROA	-0.190	0.066	0.187	-0.175	0.400	0.337	1.000				
EarningsManagement	0.179	-0.041	0.011	-0.034	-0.031	-0.095	-0.056	1.000			
AverageSalesGrowth	-0.223	-0.035	0.001	-0.177	0.094	-0.084	0.403	-0.111	1.000		
AcquisitionOutcome	0.028	-0.044	-0.098	0.083	-0.251	-0.159	-0.178	-0.070	0.003	1.000	
DocumentLength	-0.130	0.268	0.030	0.223	0.521	-0.213	-0.135	-0.068	-0.081	-0.304	1.000

TABLE 4

The ESG practices of Seeking Buyer companies compared to peers

This table presents the results of the analysis of the ESG practices of the 'Seeking buyer' company from logit regressions, where the dependent variable equals 1 if the firm is from the seeking buyer sample, and 0 otherwise. The main explanatory variable is the ESG variable, or an interaction variable to proxy for firm ESG practices. The columns report the ESG scores and the scores interacted: the ML ESG disclosure score or the Bloomberg ESG disclosure score, and the LSEG ESG performance score. All the variables are described in Appendix A. Standard errors are clustered by industry, and are reported below the coefficients. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Dependent variable	Seeking Buyer				
	ESG_{t-1} variable				
	(1) ML ESG	(2) Bloomberg ESG	(3) LSEG ESG	(4) ML*LSEG	(5) Bloomberg*LSEG
ESG Disclosure	66.2582***	0.0270		0.2836**	0.4935***
	(20.4840)	(0.0203)		(0.1382)	(0.1124)
ESG Performance			-0.0010	0.9957***	0.3635**
			(0.0050)	(0.1875)	(0.1837)
ESG Disclosure*Performance				-0.2990	-0.5248*
				(0.4723)	(0.2786)
SizeFirm	0.3034***	0.1623	0.2074***	0.2623***	0.2108**
	(0.0537)	(0.1401)	(0.0693)	(0.0445)	(0.0941)
ROA	0.0018	-0.0036	0.0052	-0.0041	-0.0074**
	(0.0018)	(0.0038)	(0.0057)	(0.0030)	(0.0031)
Leverage	0.4601	1.1589*	0.9563**	1.0121	1.2008
	(0.8207)	(0.6276)	(0.4342)	(0.8609)	(0.9196)
AgeFirm	0.0223	-0.1822	0.2161	-0.9509***	-0.8726***
	(0.4673)	(0.2715)	(0.1315)	(0.2970)	(0.3336)
EarningsManagement	-2.8284**	-2.2436	-0.5579	-4.9344***	-4.5288***
	(1.1782)	(2.3002)	(1.8939)	(1.6566)	(1.6669)
DocumentLength	0.2831			-0.0620	
	(0.2021)			(0.4510)	
Year fixed effects	X	X	X	X	X
Industry fixed effects	X	X	X	X	X
N	366	276	447	224	224
Pseudo R^2	0.221	0.108	0.078	0.243	0.244

TABLE 5

ESG practices and acquisition

This table presents the results of the analysis of the likelihood of subsequent acquisition of the 'Seeking buyer' company from logit regressions, where the dependent variable equals 1 if the firm was acquired within two years and 0 otherwise. The main explanatory variables are the ESG variables. We additionally add to all the models a 3-year rolling average of sales growth as a financial characteristic of sample companies. All variables are described in Appendix A. Standard errors are clustered by industry, and are reported below the coefficients. The Pseudo R-squared is McFadden's. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Dependent variable	AcquisitionOutcome					
Independent variables	(1) ML ESG		(2) Bloomberg ESG		(3) LSEG ESG	
ESG	-20.3736	-7.0960	-0.0090	-0.0097	0.0014	0.0039
	(69.3109)	(69.9133)	(0.0315)	(0.0343)	(0.0126)	(0.0132)
SizeFirm	-0.4136**	-0.4370**	-0.0665	-0.0605	-0.2125	-0.2544
	(0.2003)	(0.1780)	(0.2063)	(0.2100)	(0.1921)	(0.1979)
ROA	0.0094	0.0023	-0.0008	-0.0003	-0.0021	-0.0031
	(0.0062)	(0.0115)	(0.0191)	(0.0199)	(0.0090)	(0.0088)
Leverage	1.0577	1.3086	0.9747**	0.9573**	0.8361	0.9030
	(0.8497)	(0.9002)	(0.4377)	(0.4160)	(0.5616)	(0.5881)
AgeFirm	-0.8728	-0.5066	-1.1943	-1.2243	-1.5101	-1.3537
	(1.4962)	(1.4002)	(1.3371)	(1.3884)	(1.0600)	(1.0820)
EarningsManagement	0.0137	0.0145	0.1426	0.1436	0.1158	0.1006
	(0.2135)	(0.2139)	(0.1302)	(0.1316)	(0.1204)	(0.1223)
DocumentLength	-0.0785	-0.0433				
	(0.03607)	(0.3610)				
AverageSalesGrowth		0.0057		-0.0010		0.0023**
		(0.0072)		(0.0076)		(0.0009)
N	95	95	106	106	152	152
Pseudo R^2	0.161	0.168	0.062	0.062	0.084	0.102

TABLE 6

2-stage least squares: ESG practices of the 'Seeking buyer' companies

This table presents the results of the analysis of the ESG practices of the 'Seeking buyer' company from logit regressions, where the dependent variable equals 1 if the firm is from the seeking buyer sample, and 0 otherwise. The main explanatory variable 'ESG hat' is estimated as a 2-stage least squares instrumental variable to address endogeneity. The instrumental variable used is the year the company's ESG score is first reported in the Bloomberg database. All variables are described in Appendix A. Standard errors are reported below the coefficients. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

	Bloomberg ESG disclosure
ESG hat	-0.9897*
	(0.5583)
SizeFirm	5.0649*
	(2.6271)
ROA	-0.0570
	(0.0366)
Leverage	9.0133*
	(4.6720)
AgeFirm	1.1645
	(1.0581)
EarningsManagement	-5.4065
	(2.4580)
N	224
Pseudo R^2	0.122

Appendix

Appendix A – Variable definition

Variables	**Definition**
ESG_i	One of the proxies reflecting the involvement of firm i in ESG activities. We estimate *ESG disclosure* using the Bloomberg ESG disclosure score (Bloomberg ESG) and a machine learning-based ESG disclosure score (ML ESG). We evaluate *ESG performance* using the LSEG ASSET4 score.
$SeekingBuyer_i$	A dummy variable which takes the value 1 if a firm *i* has issued a 'seeking buyer' announcement in 2000-2021, and 0 otherwise
$AcquisitionOutcome_i$	A dummy variable which takes the value 1 if a firm *i* was acquired in a period of 2 years after issuing an official announcement about seeking a buyer and 0 otherwise
$Leverage_i$	Leverage of firm *i*, calculated as a share of total liabilities over total assets
ROA_i	Net income before extraordinary items (t-1) / Total assets (t-1)
$SizeFirm_i$	The logarithm of total assets of a firm *i*
$AgeFirm_i$	The logarithm of the age of a firm *i* in years
$DocumentLength_i$	The logarithm of the number of phrases in a 10-K form identified by the language model.
$EarningsManagement_i$	Earnings management estimated with the help of the modified Jones model (Dechow et al., 1995). We estimate regressions in the models for all years according to 2-digit SIC codes
$AverageSalesGrowth_i$	Three-year average sales growth before the seeking buyer announcement
ESG hat	2-stage least squares instrumental variable. The instrumental variable used is the year the company's ESG score is first reported in the Bloomberg database.

Appendix B

ESG phrases of the ML-based measure

The phrases used to create the ML-based measure for the environment, social and governance dimensions of ESG.

Environment	Social	Governance
global_warming, food_security, air_quality, global_issues, water_scarcity, the_planet, energy_policy, industry, landscape, environments, health, resource, dynamics, conservation, geologic, sustainability, environmental, climate, ecological, consumer, maritime, atmospheric, regulatory, legislative, agriculture, air, hygiene, technological, political, marketplace, geotechnical, agriculturally, bioforensic, fisheries, ecosystems, smarter_cities, metallurgical, envir, circuit, logistical, modernisation, atomic, cloud_activity, water_quality, public_safety, arctic, the_mega_trends, extreme_designs, forestry, oil_processes, shale, geospatial, intelligent, food_safety, fire_prevention, ecobiotic, people_safety, our_nation, safety_health, wildlife, ecology, airline, geopolitical, enduring, our_environment, mexico_oil, biodiversity, change_policies, critical_health, nuclear_safety, animal_habitat, standards, environmentally, fuel_leaks, haze, temperature, risk_technology, ecosystem, to_long_term, our_bamboo, agricultural, geophysically, infrastructures, automation, gas_shale, electrical, air_land, chemical, an_environment, innovative, geographic, deepwater, logistics, shale_project, dynamic_nature, geologic_basins, innovation, meteorological, polluted_waters, farm_efficiency	leadership, excellence, creativity, our_people, our_values, wellness, inspiring, the_culture, our_culture, literacy, compassion, embraces, teamwork, our_world, the_habits, innovation, ethnic, learning, mentoring, inspires, individuality, coaching, kindness, mutual_respect, this_spirit, personal_growth, mindfulness, philosophy, wellbeing, glamour, better_sleep, sustainability, fostering, favorite, well_being, talent, celebrating, society, popular_culture, vocabulary, championing, these_tenets, the_theme, welcoming, thoughts, moments, thrives, strong_culture, fun, workplace, wisdom, passion, lifelong, empowerment, philanthropy, unwavering, volunteerism, the_passion, early_learning, inclusivity, empathy, fosters, religion, business_skills, success_stories, behavioral, teacher, embracing, nurtures, online_teaching, admiration, your_leadership, humor, visual, intellectually, reading_skills, an_open_culture, living_life, everyone, adventure, fluency, school_spirit, lifestyles, multicultural, foster, our_innovation, comprehension, our_teams, spirited, the_vision, health, mentorship, respectful, warm_inviting, enthusiasm, strong_leaders, intelligence, teams, team_building, motivational	our_governance, our_ethics, oversight, risk_management, organisational, internal_audit, independence, goverance, committees, oversees, nominating, accountability, coporate, employee_code, board, the_governance, auditing, audit_committee, supervisory, policymaking, practices, leadership, board_affairs, conduct_policy, business_ethics, governance_risk, director_code, regulatory, boards, the_committees, the_board_role, our_committees, human_resource, advisory, compliance_code, legislative, conduct_code, auditors, ethics_code, human_capital, corporate_risk, risk_control, board_oversight, human_resources, clear_roles, conduct_board, board_processes, risk_committee, integrity, analogous_rules, our_code, the_oversight, insider, policies, standards, policy_code, risk_oversight, board_role, rule_changes, guidelines, frameworks, ceo_report, communications, strong_program, ethical_conduct, these_key_risks, defined_roles, oversee, audit_functions, supervision, good_governance, securities_code, committees_code, insider_filings, business_code, ethics_board, varying_norms, communication, officers_code, our_codes, our_oversight, risk_governance, ethics, directorates, risk_policies, key_policies, public_affairs, disciplinary, joint_code, oversight_rules, overseeing, security_policy, directors_codes, the_management, policy_plans, reporting, board_structure, risk_compliance, formal_policies, risk_assessment

Appendix C

TABLE C.1

ESG practices and acquisition

This table presents the results of the analysis of the likelihood of subsequent acquisition of the 'Seeking buyer' company from logit regressions, where the dependent variable equals 1 if the firm was acquired within two years and 0 otherwise. The main explanatory variables are the ESG variables. All variables are described in Appendix A. Standard errors are clustered by industry, and are reported below the coefficients. The Pseudo R-squared is McFadden's. *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

Dependent variable	AcquisitionOutcome								
Independent variables	ML ESG			Bloom ESG			LSEG ESG		
	E	S	G	E	S	G	E	S	G
ESG Disclosure	-111.6095	-337.3107	64.0494	0.0143	-0.0219	-0.0669*			
	(69.7460)	(367.5355)	(45.2548)	(0.0219)	(0.0302)	(0.0354)			
ESG Performance							0.0001	-0.0004	-0.0064
							(0.0097)	(0.0092)	(0.0089)
SizeFirm	-0.0929	-0.1631	-0.2452	-0.2069	-0.0051	0.0303	-0.2375	-0.2334	-0.1790
	(0.1465)	(0.1466)	(0.1797)	(0.2232)	(0.1982)	(0.1921)	(0.1701)	(0.1608)	(0.1530)
ROA	0.0030	0.0033	0.0039	-0.0003	-0.0007	0.0005	-0.0025	-0.0025	-0.0017
	(0.0064)	(0.0062)	(0.0071)	(0.0171)	(0.0214)	(0.0151)	(0.0082)	(0.0082)	(0.0082)
Leverage	0.5564	1.0290	0.5226	0.6762	1.0359**	1.2193***	0.8932	0.9006	0.8883*
	(0.4276)	(0.8619)	(0.4427)	(0.4437)	(0.4024)	(0.4437)	(0.5681)	(0.5659)	(0.4957)
AgeFirm	-1.5146	-0.9587	-0.9306	-1.2941	-1.1100	-1.4838	-0.9821	-0.9747	-0.9208
	(0.9923)	(1.0483)	(1.0506)	(1.2050)	(1.3415)	(1.1808)	(1.0528)	(1.0659)	(1.0670)
Earnings-Management	0.0524	-0.0409	0.0672	0.1728	0.1270	0.2169	0.0768	0.0756	0.0687
	(0.1328)	(0.1935)	(0.1227)	(0.1405)	(0.1342)	(0.1515)	(0.1171)	(0.1178)	(0.1168)
DocumentLength	0.4435	0.1219	0.6899						
	(0.6660)	(0.7318)	(0.8790)						
N	167	167	167	106	106	106	159	159	159
Pseudo R^2	0.084	0.060	0.065	0.069	0.071	0.094	0.075	0.075	0.078

Essay III: A Survey of Blockchain Disclosure by Firms in the Initial Public Offering (IPO) Process

Essi Nousiainen & Mikko Ranta

School of Accounting & Finance

University of Vaasa, Vaasa, Finland

Abstract

This paper investigates blockchain-related disclosure strategies in Initial Public Offerings (IPOs) by analyzing 712 S-1 filings from 2016-2022. Using text mining techniques, we aim to uncover trends and patterns in blockchain disclosure among emerging companies, providing unique insights into how blockchain is integrated into business models. The results reveal five key disclosure topics, which have experienced significant changes over time in sentiment, uncertainty, and modal language. This study contributes to the literature on accounting disclosure by offering a comprehensive examination of the evolving trends, challenges, and opportunities in blockchain disclosure during IPOs. The implications of our findings extend to investors, regulators, and industry stakeholders.

Keywords: Initial Public Offering, S-1, Latent Dirichlet Allocation, Disclosure

1. Introduction

The listing of the cryptocurrency exchange platform Coinbase on Nasdaq in 2021 marked a significant milestone for the emergence of cryptocurrency-based businesses in the traditional financial markets. With a high opening price of $381 per share, Coinbase became a trailblazer as the first major crypto-based business to debut on a U.S. exchange[1]. This emergence underscores a broader trend of blockchain-utilizing companies seeking public listings or Initial Public Offering (IPO) prospects throughout the 2010s and 2020s. Understanding how these companies communicate their blockchain-related activities during the IPO process is crucial, as it provides insights into the integration of emerging technologies within traditional market structures.

Blockchain technology has evolved significantly since its invention. Initially focused on cryptocurrency transactions, it progressed to Blockchain 2.0 with the introduction of smart contracts (Bhutta et al., 2021). Building upon this foundation, Blockchain 3.0 represents a broader spectrum of applications beyond finance, encompassing areas such as supply chain management, decentralized finance, and autonomous organizations. This evolution highlights the dynamic nature of blockchain technology and its capacity to reshape industries. Consequently, investigating how companies disclose their blockchain integration during the IPO process becomes pertinent.

In this paper, we study the blockchain-related disclosures of firms undergoing the IPO process. Specifically, we focus on the S-1 filings of these firms—documents submitted to the U.S. Securities and Exchange Commission (SEC) that provide detailed information about a company's financials, business model, management team, and other critical aspects. Recent studies have underlined the value of data analytical methods in business research (Delen & Zolbanin, 2018), and highlighted the value of descriptive research (Gow et al., 2016). Thus, we employ text mining techniques to descriptively analyze the content of the S-1 disclosures, allowing us to gain profound insights into the underlying themes and patterns beyond the surface level. We aim to provide unique insights into how blockchain is integrated into business models and communicated to potential investors during the IPO process.

Prior research on blockchain disclosure includes studies by Cheng et al. (2019), Stratopoulos et al. (2022), and Yen & Wang (2021). However, the existing research lacks a comprehensive investigation on the disclosure sentiment and trends from the disclosing companies' point of view. Prior research has underlined the controversies and uncertainties surrounding blockchain and cryptocurrencies. For instance,

[1] https://www.nytimes.com/live/2021/04/14/business/stock-market-today

Chokor & Alfieri (2021) found that investors react negatively to regulations in the cryptocurrency market, while Foley et al. (2019) identified the extensive use of Bitcoin for criminal activities, which is a significant concern. Given these findings, gaining insights into how companies involved with cryptocurrencies and blockchain perceive them is valuable. Accordingly, this paper has two main objectives: first, to analyse the blockchain disclosure content and textual topics in S-1 filings, and second, to examine the time evolution of disclosure sentiment.

Studying IPO companies offers a unique opportunity to observe early-stage engagement with emerging technologies, such as blockchain, as these firms are in growth-focused phases and emphasize innovation to attract investors. These companies are often at the forefront of technological innovation, and their approach to blockchain can offer valuable insights into how this technology is being integrated into various business models and strategies. IPO firms' agility in adopting new technologies provide fresh, dynamic insights compared to the more stable, incremental approaches of mature companies. In addition, S-1 filings are highly regulated, creating an interesting context for studying blockchain disclosure compared to other marketing materials that lack such scrutiny. Although the disclosure of blockchain-related information within S-1 filings is largely voluntary, the mandatory nature of the filings requires firms to disclose truthful information and address associated risks.

Cheng et al. (2019) examined speculative and existing blockchain disclosure types, along with investor responses to strategic disclosures. The literature has also addressed the stage of blockchain adoption as inferred from these disclosures, as well as its relevance to stock prices (Stratopoulos et al., 2022; Yen & Wang, 2021b). Our study adds to this body of work by conducting an in-depth analysis of blockchain disclosure content, particularly in the context of emerging companies. S-1 filings are especially valuable for this analysis, as they are inherently forward-looking, offering insights into future market trends and potential developments. Companies disclose blockchain-related information in various contexts, ranging from speculative buzzwords to established business models and products, as well as associated risks. We study the linguistic features of blockchain disclosure to evaluate changes in the disclosure sentiment trends. A body of research has also raised controversies surrounding cryptocurrencies (Chokor & Alfieri, 2021; Foley et al., 2019), which we contribute to as our results measure the sentiment from the point of view of those companies that are associated with them.

This research also contributes to the literature on accounting disclosure by providing insights into disclosure strategies within statutory and regulated filings, which offer a relatively consistent dataset for analysis. Additionally, we enhance the understanding of blockchain technology adoption and implementation by shedding light on its various business use cases. Our research adopts a descriptive approach, which is particularly beneficial for developing new theories based on forward-looking company disclosures (Gow et al., 2016). By not imposing a priori theoretical assumptions, our findings reveal intriguing strategies for integrating blockchain into business models, as well as the general sentiment toward blockchain among firms engaging with its applications. This methodology lets us capture the dynamic and evolving nature of blockchain disclosures, providing deeper insights into how emerging companies use this technology to shape their market positioning and strategic direction. As a result, our study establishes a foundation for future research that can further explore and theorize the implications of these innovative strategies.

The paper is organized as follows: Section 2 covers background literature, Section 3 discusses the sample and research methods. Findings are presented in Section 4, followed by a discussion, and the paper concludes in Section 5.

2. Background

2.1 Initial Public Offering

When going public in the United States, a company must file a registration statement known as Form S-1.[2] This form includes the IPO prospectus, which is the legal offering document containing information about the company's business operations, financial condition, results of operations, risk factors, and management.

The language used in S-1 filings has been the subject of several studies in relation to IPO returns (i.e. Loughran & McDonald, 2013). More recently, Ma et al. (2021) conducted a study to measure the extent of misleading information in IPO disclosures. They found that firms tend to report more misleading information in their IPO when they face difficulties in obtaining regulatory approval for the IPO. In a different study, Wasiuzzaman et al. (2018) explored the relationship between the disclosure of IPO risk factors and the initial returns. Chen et al. (2023) took a more specific approach by studying the disclosures related to climate change in IPO prospectuses. They investigated how these disclosures could affect the pricing of the IPO, providing insights into the growing importance of environmental factors in financial markets. Furthermore, IPO disclosures have also been studied in the context of Chinese IPO prospectuses (Zhao et al., 2022; Huang et al., 2019). In summary, the

[2] https://www.sec.gov/education/smallbusiness/goingpublic/registrationstatement

language used in S-1 filings and IPO disclosures has been a subject of interest for researchers. They have explored various elements such as the presence of misleading information, the disclosure of risk factors, and specific topics like climate change. The findings from these inquiries highlight the significance of transparency and the necessity for precise information disclosure during the public offering process. We aim to extend this literature by studying the language in Form S-1 filings in the context of blockchain disclosure.

2.2 Blockchain Technologies and Accounting

Blockchain has emerged as a significant area of interest in accounting literature due to its potential to transform traditional accounting practices. As companies increasingly adopt blockchain solutions, understanding how these technologies impact accounting practices becomes crucial for practitioners and regulators alike. The technology's inherent features—decentralization, immutability, transparency, and security—offer promising applications in areas such as auditing, financial reporting, and regulatory compliance. For instance, Dai and Vasarhelyi (2017) discuss how blockchain can enable real-time accounting and continuous auditing through automated verification of transactions, reducing the need for intermediaries and enhancing trust in financial data. Furthermore, Schmitz and Leoni (2019) explore how smart contracts, a facet of blockchain technology, can automate contractual agreements, thereby streamlining processes and reducing the potential for human error or fraud.

Garanina et al. (2022) identified ten distinct topics related to blockchain that have been addressed in accounting literature. These topics encompass a wide range of areas, including the opportunities and challenges presented by the application of blockchain technology, the trading and uncertainty associated with blockchain and cryptocurrency, and the nature of cryptocurrencies and crypto assets, among others. Luo and Yu (2022) explored accounting practices for cryptocurrency, reflecting the growing relevance of digital currencies in today’s financial landscape. Similarly, the implications of blockchain technology on audit practices have been examined by Lombardi et al. (2022), highlighting the transformative potential of this technology in traditional auditing processes. One stream of accounting research related to blockchain is concentrating on how the technology impacts accounting and the opportunities and challenges that may arise in the future. This future-oriented approach is evident in recent research efforts aimed at introducing blockchain to the accounting profession and the scientific field (see e.g. Calderón & Stratopoulos, 2020; Jayasuriya & Sims, 2023). A segment of accounting literature has specifically focused on exploring the audit implications of blockchain. For instance, Desplebin et al.

(2021), Dyball and Seethamraju (2022), and Kend and Nguyen (2020) are all focusing on how blockchain may impact the accounting and audit industry.

2.3 Blockchain Disclosure

In the context of blockchain, firms may have incentives to engage in technological opportunism, such as adopting and integrating blockchain-related activities into their operations. This is primarily driven by the potential benefits that blockchain technology can offer. According to Sarkees (2011), engaging in technological opportunism has been proven to positively impact profits and firm value. This suggests that the adoption of blockchain technology could lead to increased profitability and enhance the overall value of a firm. Blockchain technology, with its inherent features of transparency, security, and immutability, can streamline business processes, reduce operational costs, and improve efficiency, thereby potentially boosting profits. Furthermore, firms also engage in technological opportunism to gain a competitive advantage, as noted by Srinivasan et al. (2002). In the highly competitive business landscape, adopting blockchain technology can provide firms with a distinct edge. Blockchain can enable real-time tracking of transactions, enhance data security, and foster trust among stakeholders, which can differentiate a firm from its competitors.

The speculative blockchain disclosure in 8-K filings has been studied previously by Cheng et al. (2019), who found that especially speculative blockchain disclosure increased with the rise of Bitcoin prices. Stratopoulos et al. (2022), on the other hand, used corporate disclosures (i.e. 10-K and S-1 filings) to study the stage of blockchain adoption, and found that while blockchain disclosure was initially concentrated on cryptocurrency in the earlier years, there has been a shift towards focusing on blockchain business applications. This suggests that as companies become more familiar with the technology, they are beginning to explore its broader applications beyond just cryptocurrency. This trend indicates an increasing recognition of the transformative potential of blockchain technology across various business domains.

Gao et al. (2020) studied cybersecurity risk disclosures, highlighting the new types of risks that have emerged with the spread of new technology. In a similar vein, our current study identifies a blockchain risk topic, emphasizing that companies recognize the risks carried by blockchain-based assets or business models. Moreover, the potential security risks associated with blockchain and digital assets have been a topic of discussion in accounting literature (Castonguay & Stein Smith, 2020). Chokor and Alfieri (2021) also find that the increase in the probability of regulation has a negative effect on cryptocurrency returns, increasing the risks associated with them.

3. Sample and Research Methods

3.1 Sample

We begin by collecting all S-1 filings from the years 2016-2022. The starting year was determined based on Stratopoulos et al. (2022), who found that the first year that blockchain words were mentioned in S-1 filings was 2016. The scraping process retrieved 17,009 individual filings from the SEC EDGAR database. To be included in the final sample, the criterion was that the filing should mention at least one of the following blockchain related words: blockchain, crypto, cryptocurrency, digital currency, Bitcoin, Litecoin, Dogecoin, Ethereum, web3, non-fungible token, or NFT. Some disclosures mentioning non-blockchain related uses of the word "crypto" were deleted in the later phase. After filtering, we arrived at the final sample of 712 filings containing at least one of the specified blockchain-related words.

After collecting the full S-1 filings, they were split into paragraphs to study the blockchain words in their context. The sentence mentioning one of the keywords and the following four sentences, were extracted into individual text segments. The final sample includes 18,455 text segments from form S-1 filings.

3.2 Topic Model

Similarly to Stratopoulos et al. (2022), we conduct a descriptive analysis of blockchain disclosure. We use text mining techniques and sentiment analysis combined with statistical models to get profound insights of the content and shifts in blockchain-related topics. Before training the topic model, the paragraphs are pre-processed by lowercasing, removing punctuation, and converting the words into tokens. In addition, stop words and those words appearing in more than 99% or less than 4 individual paragraphs are removed and of the remaining, the most common 20,000 words are kept.

We employ Latent Dirichlet allocation (LDA) (Blei et al., 2003) for topic modelling, a probabilistic generative model that identifies latent thematic structures within textual data. It is widely employed for topic modeling in accounting research (e.g. Bellstam et al., 2020; Dyer et al., 2017; Garanina et al., 2022; Nousiainen et al., 2024). LDA assumes that documents are mixtures of topics, and topics are mixtures of words. Through iterative processing, LDA probabilistically assigns words to topics, revealing the underlying topical composition of documents. This method enables the identification of established themes within the dataset, enhancing our comprehension of the latent semantic structures that characterize blockchain disclosure in IPOs.

Determining the optimal number of topics is critical for meaningful analysis. We experimented with different models, evaluating them based on coherence scores and interpretability. A five-topic model provided the best balance between granularity and clarity, allowing us to identify distinct themes within the blockchain disclosures.

3.3 Research Methods

Following the research methodology of Gao et al. (2020), we study the changes in various linguistic features of the blockchain disclosure topics over time. We use the word lists from Loughran et al. (2011) to analyze the linguistic content of the filings, focusing specifically on negative sentiment, positive sentiment, uncertainty, litigiousness, weak modality, strong modality, and constraining language. We select a "main topic" for each text paragraph based on the topic with the highest share within the text. Then we build regression models with subsamples of each topic, where the dependent variable is the linguistic feature and the independent variable is the year of disclosure, to analyze the yearly sentiment evolution within each blockchain topic. The following regression equation is used to analyze the changes in sentiment:

$$Sentiment = \beta_0 + \beta_1\, year + \varepsilon. \qquad (1)$$

The accounting literature offers a wealth of evidence suggesting that the sentiment conveyed in disclosures holds considerable informational value. For example, Tsang et al. (2023) found that risk disclosure sentiment types have informative value and investors respond to those disclosure sentiments. Also, considering IPO sentiment, González et al. (2019) found that uncertain sentiment in IPO prospectuses was associated with IPO underpricing.

In the context of blockchain disclosure, the use of uncertain, strong or weak modal words could potentially indicate a higher degree of uncertainty surrounding blockchain technologies. Firms may strategically manipulate the tone of their disclosures to create a specific impression for the reader. For instance, Li et al. (2022) discovered that firms engaging in earnings management tend to use more positive and modal words, possibly as a tactic to obscure their earnings management behavior. This suggests that the tone of disclosure can be a tool used by firms to shape perceptions and manage impressions. Similarly, Ertugrul et al. (2017) found a correlation between the use of uncertain or weak modal words and stricter loan terms, as well as a higher risk of future stock price crashes. This implies that the tone of disclosure, particularly the use of modal words, can have significant implications for a firm's financial prospects. On the other hand, Patelli and Pedrini (2013) found that optimism in CEO letters was associated with higher past and future performance. The positive tone in blockchain disclosure might be a sign of sincere optimism regarding blockchain technology. In summary, the tone of blockchain disclosures can serve as a barometer of a firm's uncertainty about blockchain technologies. Furthermore, firms may deliberately adjust the tone of their disclosures to influence reader perceptions, highlighting the strategic role of disclosure tone in corporate communication.

4. Results

4.1 Disclosure Topics and Descriptive Statistics

4.1.1 Disclosure Topics

Table 1 presents the most representative words for each of the five topics generated by the LDA model. We labeled each topic based on these words and a thorough review of the associated text segments. Eventually we identified five topics labeled Risk Factors, General Business, Cryptocurrency Mining, Cryptocurrency Trading & Investment, and Blockchain Technology Solutions. We also assign each text paragraph a main topic based on the topic with the highest share in the specific paragraph.

[TABLE 1]

Topic 1 is labeled "Risk Factors" and mainly includes risk disclosures containing mentions of blockchain words. This topic mainly consists of cryptocurrency risks, for example, companies mention risk factors affecting the costs of bitcoin mining, such as power costs. On the other hand, some companies discuss Bitcoin related risks, such as price fluctuations and changes in the regulatory environment, which has also been previously proven to affect crypto returns (Chokor & Alfieri, 2021).

Topic 2 is labeled "General Financial", where firms discuss their finances and holdings of crypto assets. This topic is mainly statutory information and financial jargon and holds less interest in the context of this study.

Topic 3, on the other hand, discusses cryptocurrency and crypto mining, and we label it "Bitcoin Mining". This topic includes overviews of Bitcoin or Ethereum and explanations of the fundamentals of Bitcoin or other cryptocurrencies that the company operates with. Some disclosures in this topic discuss Bitcoin mining in a very general manner and not in relation to the specific company's business. This topic also discusses the technicalities of blockchain mining. The inclusion of a mining topic in the sample of companies undergoing the IPO process points to a significant prevalence of cryptocurrency mining practices among these emerging businesses.

Topic 4 includes disclosures with discussions of cryptocurrency trading and exchange, and we label it “Crypto Trading and Investment.”

Topic 5 is labeled “Blockchain Technology Solution,” and includes descriptions of blockchain based business models and technologies. This topic also includes speculative disclosures of blockchain, where the company has hired a blockchain expert as a board member.

Figure 1 illustrates the topic distribution over the sample period, showing trends and shifts in emphasis among the topics. In the early years, The Crypto Trading and Investment -topic is clearly distinguishable in the sample as a strong topic, but its importance decreases over time. In contrast, the shares of Topics 1-3 remain relatively stable throughout the sample period. The Blockchain Technology topic has the highest share in 2018, but excluding 2018 it has a relatively steadily increasing trend. Figure 2 shows the number of blockchain disclosure paragraphs over the sample period. The number of disclosures steadily increases throughout the period, with a noticeable spike in 2021, followed by a decline in 2022. This volume may also be influenced by the overall number of IPOs during the sample period. The number of paragraphs for each individual topic generally aligns with the overall trend, although certain years and topics stand out. Notably, Topic 4 – Crypto Trading and Investment – has a high number of disclosures in 2017, surpassed only by the record year 2021. For most other topics, the 2022 levels remain higher than pre-2021 levels, despite the overall decline compared to 2021. Additionally, the high proportion of the Blockchain Disclosure topic in 2018, as shown in Figure 1, is also evident in Figure 2.

Previous literature has identified somewhat different blockchain disclosure topics compared to present research (see e.g. Stratopoulos et al., 2022), when studying company filings. For example, Bitcoin transactions and Bitcoin mining topics had high shares in previous literature, whereas we find a more even distribution among the topics. On the other hand, similarly to Stratopoulos et al. (2022), the blockchain technology solution topic is steadily increasing its share among all topics. In this topic classification we find that topics 1-4 (Risk Factors, General Financial, Mining, and Trading & Investment) mainly refer to the first generation of blockchain considering crypto currencies, whereas in topic 5 Blockchain Technology Solution applications related to the later generations are discussed. We conclude that the LDA algorithm seems to have successfully distinguished these different levels of blockchain application.

[Figure 1]

[Figure 2]

4.1.2 Industry Distribution

Table 2 presents an overall look into the industry distribution. The last column also presents the average number of text extracts per company in that industry group. The largest industry groups on Table 2 are Manufacturing (95), Blank Check (243), and Services (231). Less obvious industries for blockchain disclosure in the sample include Agriculture, Forestry and Fishing (4), Mining (8), Construction (1), Transportation and Public Utilities (17), Wholesale (7), and Public Administration (1). The industries with the highest average numbers of blockchain paragraphs are Construction (119) and Public Administration (100), but as both of these industries only have one occurrence, they can be counted as outliers. Next, by far, comes Finance, Insurance and Real Estate with an average blockchain paragraph count of 77.54. The number is also quite high for the Service industry with 34.77 paragraphs on average, whereas the for the rest of the industries the average number is approximately ten or lower.

[TABLE 2]

Table 3 presents the ten most common 4-digit SIC industries in the sample. The most common individual SIC code is 6770 Blank Check. These companies are mainly special purpose acquisition companies (SPAC's), which are companies that are taken public to make an acquisition in the future (Rodrigues & Stegemoller, 2014). 2021 was a record year for SPAC IPO's, which may also be reflected in our data, when there were the highest number of blockchain disclosures in 2021 as well, as could be seen from Figure 2.

Within the top ten industries are also multiple computing and software industries (codes 7372, 7370, 7374, and 7371), which are expected since blockchain has its foundations in computer science. SIC code 2834 Pharmaceutical Preparations stands out with 24 occurrences. Whereas blockchain could potentially be a transforming technology in the pharmaceuticals supply chain (see e.g. Abdallah & Nizamuddin, 2023), closer inspection reveals that these are mainly speculative disclosures (as per the Cheng et al. (2019) definition of speculative blockchain disclosure). Lastly, there are SIC codes 6221, 7371, 6199 representing financial services and management consulting. These industries are most likely to do with Bitcoin and cryptocurrencies. See the Appendix for S-1 text extracts for each of the most common industries.

[TABLE 3]

4.1.3 Correlation coefficients

Correlation coefficients between the topics and their linguistic features are presented on Table 4. Especially the Risk Factors and Blockchain technology solution topics exhibit relatively high correlation coefficients with different linguistic features. The Risk Factors topic exhibits positive correlations with negative (0.421), uncertainty (0.449), litigious (0.381), and weak modal (0.454). The Blockchain Solution topic exhibits negative correlations with negative (-0.303) and litigious (-0.202) languages. It has a positive correlation with positive sentiment (0.229). General Business, Mining, and Trading and Investment topics exhibit mainly weak correlations with the linguistic features, except for a more pronounced positive correlation (0.121) between the negative words and the Mining topic and a negative correlation between the positive words and the Trading & Investment topic.

[TABLE 4]

4.2 Regression Results

4.2.1 Risk Factor Topic

The regression results for the Risk Factor topic (Table 5) indicate a statistically significant increase in negative sentiment, positive sentiment, uncertain language, and weak modal sentiment, along with a statistically significant decrease in litigious and constraining languages over time in the risk factor topic. While the change in strong modal sentiment is not statistically significant, there is an increase in weak modal sentiment, which could reflect a trend with the companies increasingly using weak modal words to reflect risks while simultaneously avoiding implicit language. This would be supported by the increase in uncertain language as well. On the contrary, the increase in positive sentiment might signal a growing sense of optimism regarding blockchain technologies or a reduction in perceived risks. Previous literature has also found a connection between positive tone and post-IPO performance (Bian et al., 2021). The increase in positive textual tone could also be a generally positive signal regarding company risks in this context. Alternatively, it could be a deliberate effort to temper the reader's impression. The simultaneous rise in negative and positive sentiments suggests that firms are providing a more nuanced discussion of risks—acknowledging potential pitfalls while also highlighting possible opportunities associated with blockchain technologies.

[TABLE 5]

4.2.2 General Financial Topic

The regression results on the general financial topic are reported on the "General Financial" -column on Table 5. These findings suggest a significant decrease in optimism, an increase in the expression of uncertainty, and an increase in weak modal and constraining sentiment over time in the context of general business activities related to blockchain. All of these findings together could indicate an increase in uncertainty and cautiousness regarding the declaration of cryptocurrency returns and profits.

4.2.3 Mining Topic

The regression results on the mining topic are presented on the "Mining" -column on Table 5. These findings indicate that there is a notable increase in positive sentiment, uncertain language, and weak modal sentiment over time in the context of blockchain mining. Additionally, there is a significant decrease in strong modal sentiment. However, negative sentiment and litigious language do not exhibit statistically significant changes. These linguistic changes may reflect evolving perceptions, developments, and industry trends in blockchain mining. Blockchain mining has become more common as a business model during the investigation period, and the increase in positive sentiment could mean good returns and profitability in the said industry. Previous research has found connections between optimistic tone and company performance alike (Bian et al., 2021; Patelli & Pedrini, 2013). On the other hand, the increase in uncertain language and weak modal sentiment, in addition to the decrease in strong modal language, are a sign of increased cautiousness within the companies. For example, when describing cryptocurrency mining, they may be adding remarks on the related risks and challenges that previously would have been seen unnecessary. Also new regulatory demands regarding crypto assets could require the businesses to be more cautious in their disclosure. Cryptocurrency mining requires a lot of electricity, and since energy prices have surged after the beginning of the war in Ukraine[3], it could affect the sentiment especially in the latest sample year 2022. In conclusion, companies must adapt their disclosure language to address evolving risks and opportunities related to blockchain mining, as well as to meet the expectations of investors and regulators.

[3]https://www.weforum.org/agenda/2023/02/russia-ukraine-war-energy-costs/

https://www.ecb.europa.eu/press/economic-bulletin/focus/2022/html/ecb.ebbox202204_01~68ef3c3dc6.en.html

4.2.4 Trading and Investment Topic

The regression results on the Trading and Investment topic are presented on the "Trading & Investment" -column Table 5. There is a significant increase in the use of uncertain language over time. This suggests that companies may be addressing growing uncertainties and complexities in blockchain trading and investment. There is a significant increase in weak modal sentiment over time, indicating a shift towards more cautious language when discussing blockchain trading and investment. There is also a decrease in litigious language and an increase in constraining language.

The U.S. Securities and Exchange Commission published the first report regarding its legal authority over cryptocurrency,[4] and has since enforced multiple cryptocurrency cases.[5] Companies trading cryptocurrencies having to increasingly comply with SEC regulation, and anticipating SEC declaring cryptocurrency as a security, is an issue that has emerged within the rise of blockchain technology. The increased SEC supervision and the anticipation of regulation increase are possible reasons behind the increasingly cautious sentiment in trading-related disclosure.

4.2.5 Blockchain Solution Topic

The regression results on the Blockchain solution topic are displayed on the "Blockchain Solution" -column on Table 5. These findings indicate that there have been notable changes in sentiment and linguistic features related to blockchain solutions as well over time. In contrast to the previous topics, there is a statistically significant decrease in negative and uncertain language. In addition, there is a decrease in strong and weak modal sentiment, which suggests a shift toward more moderate language in discussing these solutions. In most of the other topics, there was an increase in uncertain and weak modal languages, and such a difference for the blockchain solution topic seems interesting. It may be that the cautiousness related to cryptocurrencies is not reflected similarly in other blockchain technologies and firms have a more positive outlook on them overall. To conclude, there is a clear shift towards more confident language by the decrease in the share of uncertain, negative and modal words in this topic compared to the other topics.

One explanation to the differing results for the blockchain solution topic is that first generation blockchain applications – cryptocurrencies – are essentially financial instruments as opposed to the latter generations encompassing a wider array of applications. Companies may engage in "marketing speak" when discussing

[4] https://www.sec.gov/news/press-release/2017-131

[5] https://www.sec.gov/spotlight/cybersecurity-enforcement-actions

blockchain technologies—an informal style of disclosure compared to that of cryptocurrencies. For example, the disclosures in this topic also include speculative statements intended to suggest future endeavors to investors, and these types of disclosures are not subject to the same level of regulatory scrutiny, compared to financial assets. As Michael and Dixon (2019) argue that the credibility of voluntary non-financial disclosure is questionable, it could also be the case here. The regulation does not require the companies to practice cautiousness in their disclosure and the information is not audited, hence they lack the incentive to practice similar cautiousness as observed in the other topics.

4.2.6 Full Sample Results

We also analyze the model used for topic subsamples with the full S-1 blockchain sample. The regression results are presented on the "Full Sample" column on Table 5. The full sample regression results follow the majority of the subsample results, with a statistically significant increase in uncertain and weak modal language, and a statistically significant decrease in strong modal language, as well as a statistically significant decrease in litigious language and an increase in positive sentiment.

4.2.7 Industry Results

Next, we focus on how the previous findings apply to different industries engaging in blockchain technologies. We analyze the uncertainty-related sentiments (uncertainty, strong modal, and weak modal) in different industry subgroups, since those sentiments were the most constant trend among the topic-based subsamples. An increase (decrease) in uncertainty and weak modal language would indicate an increase (decrease) in overall uncertainty, and an increase (decrease) in strong modal language might indicate a decrease (increase) in uncertainty, or a shift towards moderate language.

The industry-specific topic distributions are presented in Figure 3. Topic 1 Risk Factors is clearly distinguishable in all industries except for the Mining and Transportation and Public Utilities. Topic 2 General Financial is predictably present in all industries as well. Topic 3 Mining is the most pronounced in Wholesale, Public Administration, Manufacturing, and Finance, Insurance and Real Estate industries. Topic 4 Trading & Investment is only clearly present in the Finance, Insurance and Real Estate industry, and somewhat in Wholesale as well. Topic 5 Blockchain Solutions can be found in every industry, but least for Finance, Insurance and Real Estate.

The regression results for the industry subsample are presented in Table 6. There is a statistically significant increase in uncertain and weak modal language observed for the Finance, Insurance and Real Estate, Services, and Retail industries. There is also a statistically significant decrease in strong modal language for the Manufacturing, Finance, Insurance and Real Estate, and Retail industries. There are no changes in sentiment for the Transportation and Public Utilities industry. Also, the results for the Manufacturing industry do not indicate an increase in uncertainty, unlike the rest.

[Table 6]

[Figure 3]

4.3 Discussion

Our findings reveal significant shifts in the sentiment and linguistic features of blockchain disclosures in IPO filings from 2016 to 2022. Every blockchain-related S-1 topic, except Topic 5—Blockchain Solutions—exhibited a statistically significant increase in the use of uncertain and weak modal language during the period. One might expect that as new technologies become more established, the uncertainty surrounding them would decrease. However, our findings suggest that this is not the case with blockchain technologies, particularly cryptocurrencies. Increased volatility and regulatory changes in the cryptocurrency markets during the later years of the sample period may have contributed to a more cautious tone compared to earlier years. Companies are clearly reflecting the controversies related to cryptocurrencies in their disclosures, possibly to align with public sentiment and regulatory expectations.

Previous literature has found a connection between uncertain or modal disclosure language and stricter loan terms, as well as a higher risk of future stock price crashes (Ertugrul et al., 2017). This could indicate that blockchain-disclosing companies may also be facing financial challenges. González et al. (2019) also found that uncertain sentiment in IPO prospectuses was associated with IPO underpricing, and hence the disclosure sentiment is a relevant signaling mechanism for IPO firms. The increase in uncertainty, especially over the crypto-related topics could signal that cryptocurrencies were met with enthusiasm in the early adoption period, but were somewhat unable to live up to the expectations as challenges have started to emerge. The public debate on the challenges related to cryptocurrencies and the threat of increasing regulation may also factor in the uncertainty in disclosures. On the other hand, we find that the blockchain technology solutions disclosure is being reported with an increasingly confident, or at least less uncertain, language, supporting the findings of Stratopoulos et al. (2022). However, this can also be a consequence of the poor credibility of less regulated voluntary non-financial disclosure, as Michael and Dixon (2019) argue.

The increase in uncertainty was the most pronounced in the Finance, Insurance and Real Estate, Services, and Retail industries. Interestingly, despite the Retail sector holding a substantial majority in the Blockchain Solutions topic (over fifty percent), where there was no increase in uncertainty, the industry still experienced an increase in uncertainty. In contrast, the Finance, Insurance and Real Estate sectors featured topics that were notably marked by heightened uncertainty. It is also noteworthy that the main topic distribution for the Manufacturing industry is very similar to the Services industry, but an increase in uncertainty was only present in the Services industry. Hence, there seems to be also other factors to the topics affecting the changes in sentiment industry-wise.

5. Conclusion

In this comprehensive survey, we explored the landscape of blockchain disclosure by firms in the initial public offering (IPO) process, focusing on S-1 filings submitted to the U.S. Securities and Exchange Commission (SEC) from 2016 to 2022. Employing text mining techniques, we investigated the content, topics, and sentiment evolution of blockchain disclosures to understand the patterns and trends shaping the integration of blockchain technology into emerging businesses.

Previous research reveals that firms use speculative blockchain disclosure in times of heightened bitcoin prices (Cheng et al., 2019) and that as firms become more accustomed to blockchain technologies, they begin to explore its applications beyond cryptocurrency (Stratopoulos et al., 2022). Disclosure sentiment has also been associated with various financial consequences (see e.g. Ertugrul et al., 2017; González et al., 2019; Li et al., 2022; Tsang et al., 2023).

The empirical results of our analysis showcase distinct topics within blockchain disclosures: Risk Factors, General Business, Cryptocurrency Mining, Cryptocurrency Trading & Investment, and Blockchain Technology Solutions. These topics provide a nuanced understanding of how firms communicate their involvement with blockchain technology in the IPO process. The industry analysis reveals a diverse range of sectors engaging with blockchain, from computing and software to pharmaceuticals and financial services.

Furthermore, the regression analyses highlight significant changes in sentiment and language use over time within each blockchain-related topic. Notably, there is an overall increase in uncertainty and cautious language, reflecting the evolving dynamics, challenges, and regulatory landscape associated with blockchain technologies. These findings contribute new insights into the literature researching accounting disclosure and provide a future research stream to study the aftermarket performance of these IPO companies engaging in blockchain disclosure. We also provide a method for quantifying and grouping the blockchain disclosure for further study. The fundamentals of blockchain-disclosing companies could be further inspected, for example, to examine their leverage or profitability, and how the uncertain sentiment is related to the financial performance of the company.

The implications of our findings are significant for both investors and regulators. The increased uncertainty on a general level may signal heightened risk in the cryptocurrency market, which investors should carefully consider. Those investing in IPOs should also scrutinize these disclosures and critically evaluate any references to blockchain technologies. Our findings suggest that companies are adopting a more cautious approach toward cryptocurrencies, possibly due to market conditions or concerns about rising regulation. In addition, managers and practitioners could assess the tone of their competitors' or potential customers' blockchain disclosure using our approach combining text analysis methods. Similarly to Stratopoulos et al. (2022), the LDA approach could help narrow down competitors or customers, but combining it with the sentiment trends technique from Gao et al. (2020b) would be useful for further insights on the group of companies.

The regulatory implications of the research results could be assessed further to determine, whether more regulation of cryptocurrency-related issues would be necessary. First generation blockchain applications – cryptocurrencies – are essentially financial instruments, as opposed to the later generations encompassing a wider array of applications. Although financial disclosure is already heavily regulated, regulators should assess whether firms are accurately reporting the risks and financial consequences associated with cryptocurrency holdings, as we observe the increase in uncertain language. The same level of uncertainty does not apply to the blockchain technology topic, raising questions about the underlying reasons. Scientific literature does not criticize blockchain technologies to a similar extent to cryptocurrencies. Companies may also see cryptocurrencies as a specifically difficult asset or business model, depending on the extent that it is integrated into the company (cf. holding crypto assets and crypto exchange as a business model). Another possible explanation is that companies may engage in "marketing speak" when discussing blockchain technologies—an informal style of disclosure compared to that of cryptocurrencies. For example, blockchain disclosures might include speculative statements intended to suggest future endeavors to investors, and these types of disclosures are not subject to the same level of regulatory scrutiny or litigation risk. A further research path would be to study the extent and credibility of speculative disclosure in the blockchain context. Moreover, further investigation into the impact of these disclosures on investor behavior and post-IPO performance would also be beneficial. Examining how market participants interpret and react to the sentiment and content of blockchain disclosures could provide deeper insights into the economic consequences of corporate communication strategies.

While this study provides valuable insights into blockchain disclosure practices during IPOs, it is not without limitations. Our analysis is confined to S-1 filings, which, although rich in information, represent just one facet of corporate disclosure. Additionally, our study focuses on U.S. firms; extending the analysis to other jurisdictions could offer comparative perspectives on global blockchain disclosure practices.

In conclusion, this study enhances our understanding of blockchain disclosure practices in the IPO context, providing valuable insights for investors, regulators, and researchers alike. As blockchain technology continues to reshape industries, our findings contribute to the discourse on transparency, risk communication, and strategic positioning of firms adopting blockchain in their operations.

References

Abdallah, S., & Nizamuddin, N. (2023). Blockchain-based solution for Pharma Supply Chain Industry. *Computers & Industrial Engineering*, *177*, 108997. https://doi.org/10.1016/J.CIE.2023.108997

Bellstam, G., Bhagat, S., & Cookson, J. A. (2020). A Text-Based Analysis of Corporate Innovation. *Management Science*, *67*(7), 4004–4031. https://doi.org/10.1287/MNSC.2020.3682

Bhutta, M. N. M., Khwaja, A. A., Nadeem, A., Ahmad, H. F., Khan, M. K., Hanif, M. A., Song, H., Alshamari, M., & Cao, Y. (2021). A Survey on Blockchain Technology: Evolution, Architecture and Security. *IEEE Access*, *9*, 61048–61073. https://doi.org/10.1109/ACCESS.2021.3072849

Bian, S., Jia, D., Li, R., Sun, W., Yan, Z., & Zheng, Y. (2021). Can management tone predict IPO performance? – Evidence from mandatory online roadshows in China. *Pacific-Basin Finance Journal*, *68*, 101588. https://doi.org/10.1016/J.PACFIN.2021.101588

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.

Calderón, J., & Stratopoulos, T. C. (2020). What Accountants Need to Know about Blockchain*. *Accounting Perspectives*, *19*(4), 303–323. https://doi.org/10.1111/1911-3838.12240

Castonguay, J. J., & Stein Smith, S. (2020). Digital Assets and Blockchain: Hackable, Fraudulent, or Just Misunderstood?*. *Accounting Perspectives*, *19*(4), 363–387. https://doi.org/10.1111/1911-3838.12242

Cheng, S. F., De Franco, G., Jiang, H., & Lin, P. (2019). Riding the blockchain mania: Public firms' speculative 8-K disclosures. *Management Science*, *65*(12), 5901. https://doi.org/10.1287/mnsc.2019.3357

Chen, J. W., Khoo, E. S., & Peng, Z. (2023). Climate change disclosure and the information environment in the initial public offering market. *Accounting & Finance*, *63*(S1), 907–952. https://doi.org/https://doi.org/10.1111/acfi.13085

Chokor, A., & Alfieri, E. (2021). Long and short-term impacts of regulation in the cryptocurrency market. *The Quarterly Review of Economics and Finance*, *81*, 157–173. https://doi.org/10.1016/J.QREF.2021.05.005

Dai, J., & Vasarhelyi, M. A. (2017). Toward Blockchain-Based Accounting and Assurance. *Journal of Information Systems*, *31*(3), 5–21. https://doi.org/10.2308/ISYS-51804

Delen, D., & Zolbanin, H. M. (2018). The analytics paradigm in business research. *Journal of Business Research*, *90*, 186–195. https://doi.org/10.1016/J.JBUSRES.2018.05.013

Desplebin, O., Lux, G., & Petit, N. (2021). To Be or Not to Be: Blockchain and the Future of Accounting and Auditing*. *Accounting Perspectives*, *20*(4), 743–769. https://doi.org/10.1111/1911-3838.12265

Dyball, M. C., & Seethamraju, R. (2022). Client use of blockchain technology: exploring its (potential) impact on financial statement audits of Australian accounting firms. *Accounting, Auditing and Accountability Journal*, *35*(7), 1656–1684. https://doi.org/10.1108/AAAJ-07-2020-4681

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. https://doi.org/10.1016/j.jacceco.2017.07.002

Ertugrul, M., Lei, J., Qiu, J., & Wan, C. (2017). Annual Report Readability, Tone Ambiguity, and the Cost of Borrowing. *Journal of Financial and Quantitative Analysis*, *52*(2), 811–836. https://doi.org/10.1017/S0022109017000187

Foley, S., Karlsen, J. R., & Putnins, T. J. (2019). Sex, Drugs, and Bitcoin: How Much Illegal Activity Is Financed through Cryptocurrencies? *The Review of Financial Studies*, *32*(5), 1798–1853. https://doi.org/10.1093/RFS/HHZ015

Gao, L., Calderon, T. G., & Tang, F. (2020a). Public companies' cybersecurity risk disclosures. *International Journal of Accounting Information Systems*, *38*, 100468. https://doi.org/10.1016/J.ACCINF.2020.100468

Garanina, T., Ranta, M., & Dumay, J. (2022). Blockchain in accounting research: current trends and emerging topics. *Accounting, Auditing and Accountability Journal*, *35*(7), 1507–1533. https://doi.org/10.1108/AAAJ-10-2020-4991

González, M., Guzmán, A., Tellez-Falla, D. F., & Trujillo, M. A. (2019). Governance, sentiment analysis, and initial public offering underpricing. *Corporate Governance: An International Review*, *27*(3), 226–244. https://doi.org/10.1111/CORG.12272

Gow, I. D., Larcker, D. F., & Reiss, P. C. (2016). Causal Inference in Accounting Research. *Journal of Accounting Research*, *54*(2), 477–523. https://doi.org/10.1111/1475-679X.12116

Huang, F., Xiang, L., Liu, R., Su, S., & Qiu, H. (2019). The IPO corporate social responsibility information disclosure: Does the stock market care? *Accounting & Finance*, *59*(S2), 2157–2198. https://doi.org/https://doi.org/10.1111/acfi.12534

Jayasuriya, D. D., & Sims, A. (2023). From the abacus to enterprise resource planning: is blockchain the next big accounting tool? *Accounting, Auditing and Accountability Journal*, *36*(1), 24–62. https://doi.org/10.1108/AAAJ-08-2020-4718

Kend, M., & Nguyen, L. A. (2020). Big Data Analytics and Other Emerging Technologies: The Impact on the Australian Audit and Assurance Profession. *Australian Accounting Review*, *30*(4), 269–282. https://doi.org/10.1111/auar.12305

Li, S., Wang, G., & Luo, Y. (2022). Tone of language, financial disclosure, and earnings management: a textual analysis of form 20-F. *Financial Innovation*, *8*(1), 43. https://doi.org/10.1186/s40854-022-00346-5

Lombardi, R., de Villiers, C., Moscariello, N., & Pizzo, M. (2022). The disruption of blockchain in auditing – a systematic literature review and an agenda for future research. *Accounting, Auditing and Accountability Journal*, *35*(7), 1534–1565. https://doi.org/10.1108/AAAJ-10-2020-4992

Loughran, T., & McDonald, B. (2013). IPO first-day returns, offer price revisions, volatility, and form S-1 language. *Journal of Financial Economics*, *109*(2), 307–326. https://doi.org/10.1016/J.JFINECO.2013.02.017

Loughran, T., Mcdonald, B., Battalio, R., Easton, P., Fuehrmeyer, J., Gao, P., Harvey, C., Hirschey, N., Marietta-Westberg, J., & Schultz, P. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal Of Finance*, *LXVI*(1). https://doi.org/10.1111/j.1540-6261.2010.01625.x

Luo, M., & Yu, S. (2022). Financial reporting for cryptocurrency. *Review of Accounting Studies*. https://doi.org/10.1007/s11142-022-09741-w

Ma, W., Wang, X., Wang, Y., & Wu, G. (2021). Measuring misleading information in IPO prospectuses. *Review of Quantitative Finance and Accounting*, *57*(3), 819–843. https://doi.org/10.1007/s11156-021-00964-7

Michael, A., & Dixon, R. (2019). Audit data analytics of unregulated voluntary disclosures and auditing expectations gap. *International Journal of Disclosure and Governance*, *16*(4), 188–205. https://doi.org/10.1057/S41310-019-00065-X/TABLES/4

Nousiainen, E., Ranta, M., Ylinen, M., & Järvenpää, M. (2024). Using machine learning and 10-K filings to measure innovation. *Accounting & Finance, 64(*4*), 3211-3239*. https://doi.org/10.1111/ACFI.13245

Patelli, L., & Pedrini, M. (2013). Is the Optimism in CEO's Letters to Shareholders Sincere? Impression Management Versus Communicative Action During the Economic Crisis. *Journal of Business Ethics 2013 124:1*, *124*(1), 19–34. https://doi.org/10.1007/S10551-013-1855-3

Rodrigues, U., & Stegemoller, M. (2014). What all-cash companies tell us about IPOs and acquisitions. *Journal of Corporate Finance*, *29*, 111–121. https://doi.org/10.1016/J.JCORPFIN.2014.07.003

Sarkees, M. (2011). Understanding the links between technological opportunism, marketing emphasis and firm performance: Implications for B2B. *Industrial Marketing Management*, *40*(5), 785–795. https://doi.org/10.1016/J.INDMARMAN.2010.09.001

Schmitz, J., & Leoni, G. (2019). Accounting and Auditing at the Time of Blockchain Technology: A Research Agenda. *Australian Accounting Review*, *29*(2), 331–342. https://doi.org/10.1111/AUAR.12286

Srinivasan, R., Lilien, G. L., & Rangaswamy, A. (2002). Technological opportunism and radical technology adoption: An application to e-business. *Journal of Marketing*, *66*(3), 47–60. https://doi.org/10.1509/JMKG.66.3.47.18508

Stratopoulos, T. C., Wang, V. X., & Ye, H. (2022). Use of Corporate Disclosures to Identify the Stage of Blockchain Adoption. *Accounting Horizons*, *36*(1), 197–220. https://doi.org/10.2308/HORIZONS-19-101

Tsang, R. C. W., Baldwin, A. A., Hair Jr., J. F., Affuso, E., & Lahtinen, K. D. (2023). The Informativeness of Sentiment Types in Risk Factor Disclosures: Evidence from Firms with Cybersecurity Breaches. *Journal of Information Systems*, *37*(3), 157–190. https://doi.org/10.2308/ISYS-2022-014

Wasiuzzaman, S., Yong, F. L. K., Sundarasen, S. D. D., & Othman, N. S. (2018). Impact of disclosure of risk factors on the initial returns of initial public offerings (IPOs). *Accounting Research Journal*, *31*(1), 46–62. https://doi.org/10.1108/ARJ-09-2016-0122

Yen, J. C., & Wang, T. (2021). Stock price relevance of voluntary disclosures about blockchain technology and cryptocurrencies. *International Journal of Accounting Information Systems*, *40*, 100499. https://doi.org/10.1016/J.ACCINF.2021.100499

Zhao, M., Ke, Y., & Yi, Y. (2022). The effects of risk factor disclosure on analysts' earnings forecasts: evidence from Chinese IPOs. *Asia-Pacific Journal of Accounting & Economics*, *29*(4), 866–895. https://doi.org/10.1080/16081625.2020.1772089

Tables

Table 1. Blockchain disclosure topics

1 Risk Factors	2 General Financial	3 Bitcoin Mining	4 Crypto Trading and Investment	5 Blockchain Technology Solution
Asset	Asset	Bitcoin	Bitcoin	Technology
Digital	Cash	Network	Trust	Digital
Result	Common	Transaction	Price	Service
Security	Expense	Mining	Exchange	Platform
Operation	Agreement	Miner	Market	Customer
Crypto	Cost	Asset	Fund	Blockchain
Currency	Loss	Digital	Sponsor	Currency
Market	Cryptocurrency	Block	Investment	Bank
Regulatory	Statement	User	Participant	Industry
include	Mining	Blockchain	Trading	Include

Table 2. Overall industry distribution of the sample firms

Industry	Count	Average number of text extracts per company
Agriculture, Forestry and Fishing	4	3.5
Mining	8	1.38
Construction	1	119
Manufacturing	95	9.56
Transportation and Public Utilities	17	2.71
Wholesale	7	13.14
Retail	16	10.94
Blank Check	243	5.88
Finance, Insurance and Real Estate	56	77.54
Services	231	34.77
Public Administration	1	100
NA	35	91

Table 3. Most common SIC 4-digit industries found in the sample

Industry	Count
6770 – Blank Check	243
7372 – Prepackaged Software	61
7389 – Business Services	38
7370 – Computer Programming, Data Processing and Other Services	28
7374 – Computer Processing and Data Preparation and Processing Services	25
2834 – Pharmaceutical Preparations	20
6221 – Commodity Contracts Brokers and Dealers	14
8742 – Management Consulting Services	12
7371 – Computer Programming Services	11
6199 – Publicly Traded Finance Services	10

Table 4. Correlation coefficients

	1 - Risk Factors	2 - General business	3 - Mining	4 - Trading & Investment	5 - Blockchain Solution
Negative	0.421	-0.167	0.121	-0.043	-0.303
Positive	-0.021	-0.121	0.088	-0.181	0.229
Uncertainty	0.449	-0.233	0.078	0.016	-0.289
Litigious	0.381	-0.097	-0.152	0.073	-0.202
Strong modal	-0.091	-0.111	-0.029	0.249	-0.033
Weak modal	0.454	-0.276	0.132	0.028	-0.314
Constraining	0.238	0.015	-0.078	0.020	-0.185

Note: This table contains the correlation coefficients between each blockchain disclosure topic and sentiment style.

Table 5. Regression results

Dependent variable	Independent variable	Risk Factors	General Financial	Mining	Trading & Investment	Blockchain Solution	Full Sample
Negative Sentiment	Intercept	0.0206**	0.0110*	0.0283***	0.0135**	0.0226***	0.0213***
		(0.010)	(0.006)	(0.009)	(0.005)	(0.004)	(0.003)
	Year	0.0011**	0.0001	0.00007	0.0003	-0.0008***	0.0001
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Positive Sentiment	Intercept	-0.0067**	0.0182***	-0.0103**	0.0038*	0.0214***	0.0013
		(0.003)	(0.005)	(0.004)	(0.002)	(0.005)	(0.002)
	Year	0.0008***	-0.0006**	0.0012***	0.0001	-0.0002	0.0004***
		(0.000)	(0.000)	(0.000)	(0.0001)	(0.000)	(0.0001)
Uncertainty	Intercept	-0.0104	-0.0040	0.0032	0.0119***	0.0259***	0.0115***
		(0.007)	(0.005)	(0.006)	(0.004)	(0.004)	(0.002)
	Year	0.0024***	0.0007***	0.0011***	0.0005**	-0.0008***	0.0005***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Litigious Language	Intercept	0.0588***	0.0152***	0.0040	0.0251***	0.0035	0.0241***
		(0.007)	(0.005)	(0.003)	(0.004)	(0.003)	(0.002)
	Year	-0.0017***	-0.0003	0.0001	-0.0005**	0.00005	-0.0006***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
Strong Modal	Intercept	0.0024	0.0067**	0.0147***	0.0093***	0.0130***	0.0163***
		(0.002)	(0.003)	(0.003)	(0.003)	(0.003)	(0.001)
	Year	0.0001	-0.0002	-0.0005***	0.0001	-0.0003**	-0.0005***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.0001)
Weak Modal	Intercept	0.0013	-0.0022	0.0011	0.0016	0.0103***	0.0076***
		(0.005)	(0.003)	(0.005)	(0.003)	(0.002)	(0.002)
	Year	0.0012***	0.0003**	0.0008***	0.0006***	-0.0003***	0.0003***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.0001)
Constraining	Intercept	0.0214***	-0.0019	0.0055**	0.0015	0.0043***	0.0056***
		(0.004)	(0.004)	(0.002)	(0.002)	(0.002)	(0.001)
	Year	-0.0005***	0.0004**	-0.0001	0.0002**	-0.0001	0.00002
		(0.000)	(0.000)	(0.000)	(0.000)	(0.0001)	(0.00006)

Note: The table presents the regression results where the dependent variable is the sentiment type, and the independent variable is the year. The columns report the coefficients for each subsample with asterisks indicating statistical significance: *, **, and *** representing 10%, 5% and 1% level, respectively. The standard errors are reported in brackets below the coefficients.

Table 6. *Industry sentiment*

		Dependent variable		
Industry	Independent variable	Uncertainty	Strong Modal	Weak Modal
Manufacturing	Intercept	0.0219*	0.0200***	0.0194**
		(0.012)	(0.007)	(0.009)
	Year	-0.0003	-0.0008**	-0.0005
		(0.001)	(0.000)	(0.000)
Finance, insurance, real estate	Intercept	0.0120***	0.0177***	0.0018
		(0.004)	(0.002)	(0.003)
	Year	0.0005***	-0.0005***	0.0006***
		(0.000)	(0.000)	(0.000)
Services	Intercept	-0.0062	0.0027	0.0023
		(0.004)	(0.002)	(0.003)
	Year	0.0014***	0.0001	0.0006***
		(0.000)	(0.0001)	(0.000)
Transportation and public utilities	Intercept	0.0483*	0.0124	-0.0051
		(0.027)	(0.021)	(0.008)
	Year	-0.0017	-0.0002	0.0005
		(0.001)	(0.001)	(0.000)
Retail	Intercept	-0.0317	0.0960***	-0.0304**
		(0.020)	(0.016)	(0.014)
	Year	0.0022**	-0.0044***	0.0018**
		(0.001)	(0.001)	(0.001)

Note: The table presents the regression results of industry subsamples, where the dependent variable is the sentiment type and the independent variable is the year. The rows report the coefficients for each industry subsample with asterisks indicating statistical significance: *, **, and *** representing 10%, 5% and 1% level, respectively. The standard errors are reported in brackets below the coefficients.

Figures

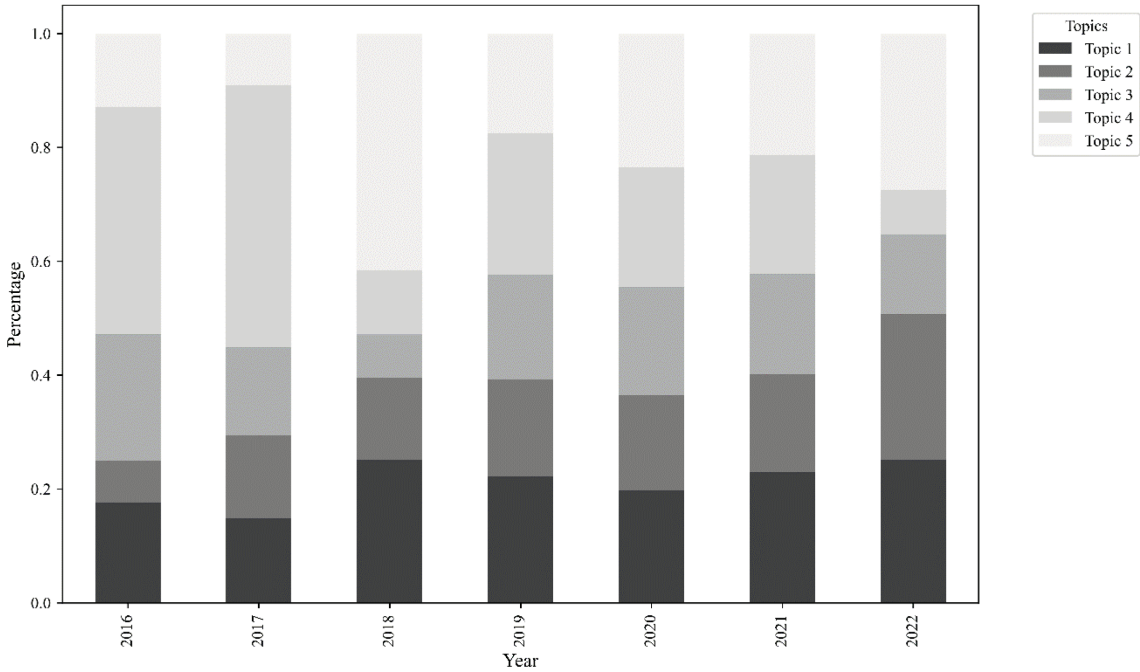

Figure 1. Topic distribution over the sample period

Topic labels: 1. Risk Factors, 2. General Financial, 3. Mining, 4. Trading and Investment, 5. Blockchain Technology Solutions

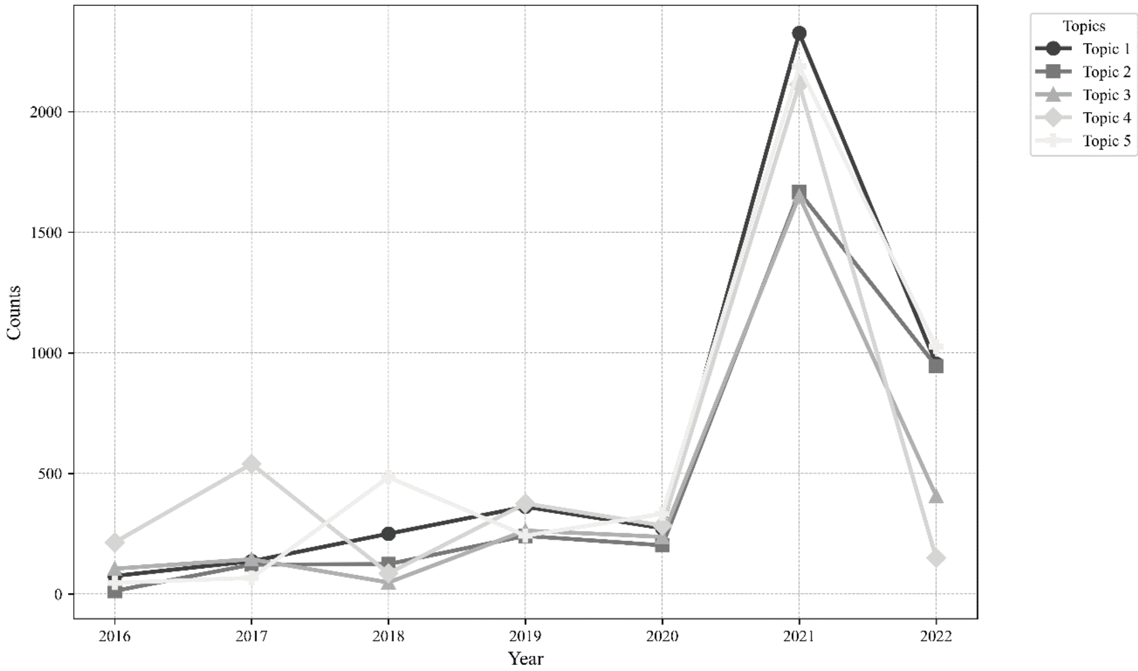

Figure 2. Time evolution of the blockchain disclosure topics

Topic labels: 1. Risk Factors, 2. General Financial, 3. Mining, 4. Trading and Investment, 5. Blockchain Technology Solutions

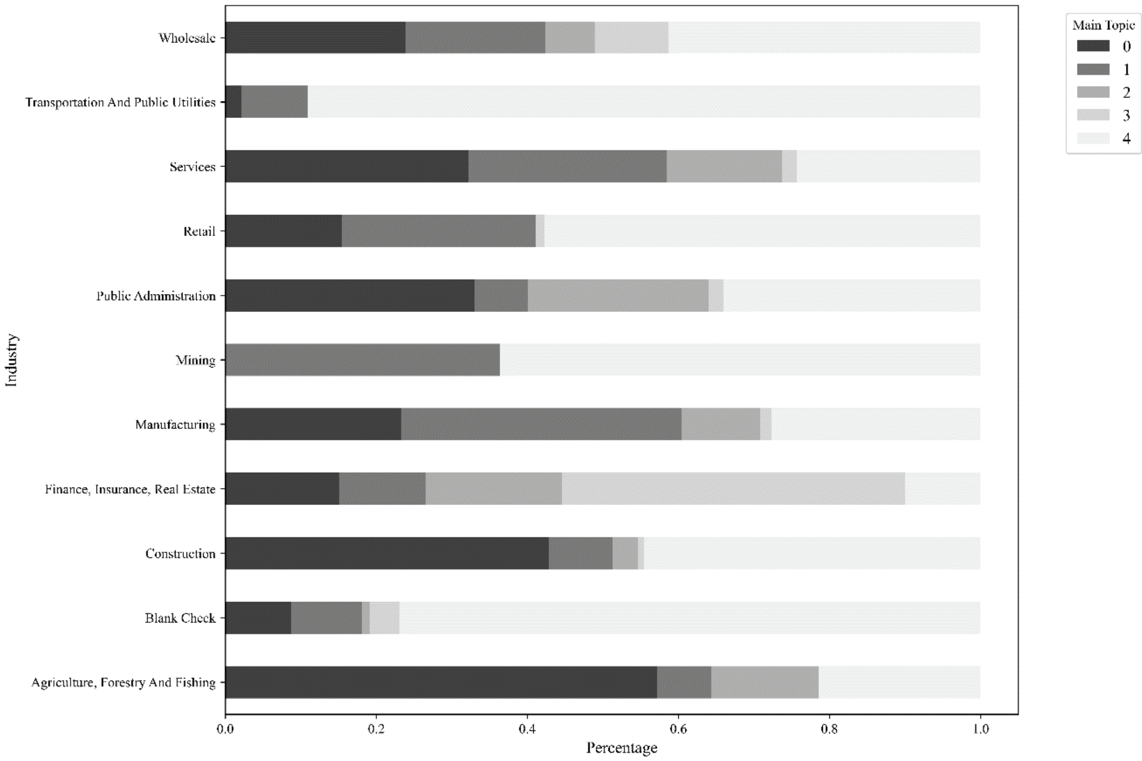

Figure 3. Industry topics

Topic labels: 1. Risk Factors, 2. General Financial, 3. Mining, 4. Trading and Investment, 5. Blockchain Technology Solutions

Appendix

Table A.1
Exemplary text extracts from the ten most common industries in the sample

6770 – Blank Check

(2) Queen’s Gambit Growth Capital II — February 22, 2021

Business Strategy

[…]

Based on the diverse set of relationships of our management team, board, Advisory Board and sponsor, we have identified several likely industries where we expect these companies exist. While the areas set forth below are representative of our primary areas of focus, this is by no means an exhaustive list.

[…]

Emerging Technology: There are new and rapidly developing technologies with wide applications and potential disruptions across a range of industries, such as sustainable AI, blockchain, deep learning and automation, agriculture and protein technology that we believe present tremendous opportunity.

(1) Iron Horse Acquisitions Corp. — November 1, 2022

Since February 2021, Mr. Caragol has also served on the Board of Directors and is Chairman of the Audit Committee of Greenbox POS (NASDAQ: GBOX) a financial technology company leveraging proprietary blockchain security to build customized payment solutions, and since July 2021 has served on the Board of Directors of Worksport Ltd. (NASDAQ: WKSP), an emerging electric vehicle company.

7372 – Prepackaged Software

(1) Vertex, Inc. — July 2, 2020

We have also established an innovation lab where we design, test and incubate next-generation tax solutions and adjacent market opportunities like blockchain, payment platforms and machine learning technologies. Over time, we expect this will bring additional value to existing customers and help us acquire new customers.

(2) APT Systems, Inc. — June 22, 2018

The Company became a member of the Enterprise Ethereum Alliance last November.

As a Fintech company developing its platforms, APT services and products can later extend to include:

- Financial Software and Analytical Software Development
- Algorithmic Applied Technology
- Trading Platforms and Exchanges, Linked to User Brokerage Accounts
- Explore Smart Contracts using our Enterprise Ethereum Alliance membership
- Use Blockchain to develop platforms for cryptoassets based on ERC-20 protocols

7389 – Business Services

(1) Altavoz Entertainment, Inc. — December 31, 2016

We implement Music Public Blockchain protocol (“Blockchain”). Blockchain is a computer-based content management system that allows artists to manage their music catalogs.

(2) Helbiz, Inc. — June 28, 2022

For example, we were recently a defendant in a putative class action suit in New York relating to an initial coin offering of a crypto currency, the HBZ coin, conducted by HBZ Systems PTE Ltd. (“HBZ Systems”) in early 2018. Although HBZ Systems has some common ownership with us, we consider it an unrelated party. Following the initial coin offering, HBZ Systems had entered into an arms’-length loan agreement pursuant to which we received a loan of $1,361,717 with a 9% interest rate per annum (as disclosed in our financial statements), fully repaid during 2021. Helbiz received no other funds from HBZ Systems.

7370 – Computer Programming, Data Processing and Other Services

(1) Applied Blockchain, Inc.	August 13, 2021
Our Business We are a participant in the dynamic Cryptoeconomy, actively mining cryptoassets and developing hosting services offered to other cryptoasset (crypto) miners. We are positioning ourself to play an active role in increasing acceptance of cryptoassets, recognizing cryptoassets as a currency and store of value. While crypto itself is not new, demand for it in the recent years has led to many financial institutions giving clients exposure through various investment vehicles, retail investors allocating portions of their portfolios to crypto, merchants accepting crypto as a form of payment, crypto ATM machines becoming more readily available around the world, and at least one government approving Bitcoin as legal tender requiring merchants to be able to accept Bitcoin as well as other forms of currency.	
(2) Core Scientific, Inc.	February 8, 2022
A slowdown in the demand for blockchain technology or blockchain hosting resources and other market and economic conditions could have a material adverse effect on our business, financial condition and results of operations. Adverse developments in the blockchain industry, and in the blockchain hosting market could lead to a decrease in the demand for hosting resources, which could have a material adverse effect on our business, financial condition and results of operations.	
7374 – Computer Processing and Data Preparation and Processing Services	
(1) Greenidge Generation Holdings Inc.	September 17, 2021
Risks Related to Our Business […] • Regulatory changes or actions may alter the nature of an investment in us or restrict the use of bitcoin in a manner that adversely affects our business, prospects or operations. • We are subject to risks related to Internet disruptions, which could have an adverse effect on our ability to mine bitcoin. • Our future success will depend significantly on the price of bitcoin, which is subject to risk and has historically been subject to wide swings and significant volatility. • The impact of geopolitical and economic events on the supply and demand for bitcoin is uncertain. • Bitcoin miners and other necessary hardware are subject to malfunction, technological obsolescence, the global supply chain and difficulty and cost in obtaining new hardware.	
(2) Phunware, Inc.	February 5, 2019
Gains and losses realized upon sale of cryptocurrencies are also recorded in other income (expense) in our consolidated statements of operations and comprehensive loss. The Company sold its remaining holdings in cryptocurrencies during the third quarter of 2018. Realized losses on sales of digital currencies were $31 and $21 for the three and nine months ended September 30, 2018, respectively.	
2834 – Pharmaceutical Preparations	
(1) Hoth Therapeutics, Inc.	August 30, 2019
From October 2017 to May 2018, Mr. Briones served as the Chief Financial Officer of Bitzumi, Inc., a Bitcoin exchange and marketplace.	
(2) Rapid Therapeutic Science Laboratories, Inc.	April 2, 2021
We were previously engaged in pursuing the business of bitcoin mining and digital currency and were not successful in that business. In November 2019, we adopted a new business strategy focused on developing potential commercial opportunities involving the rapid application of therapeutics using inhaler technology that the Company has licensed for its use.	
6221 – Commodity Contracts Brokers and Dealers	
(1) Etherindex Ether Trust	July 15, 2016

Regulatory changes or actions may alter the nature of an investment in the Shares or restrict the use of ether or the operation of the Ethereum Network in a manner that adversely affects an investment in the Shares.

Until recently, little or no regulatory attention has been directed toward Digital Assets by U.S. federal and state governments, foreign governments and self-regulatory agencies. As Digital Assets, including bitcoin and ether, have grown in popularity and in market size, the Federal Reserve Board, U.S. Congress and certain U.S. agencies (e.g., the CFTC, FinCEN and the Federal Bureau of Investigation) have begun to examine the nature of Digital Assets and the markets on which they are traded.

(2) United States Bitcoin and Treasury Investment Trust — January, 2019

The investment objective of the Trust is for the Shares to closely reflect the exposure of the Index to Bitcoin, less the Trust's liabilities and expenses. The Index is used to determine the Trust's allocation between Bitcoin and U.S. Treasuries, and the Trust adjusts its assets on a monthly basis to closely replicate the exposure of the Index to Bitcoin without the use of any derivatives and/or leverage or any similar or related products or strategies. The Shares are intended to provide investors with a cost-effective and convenient way to gain exposure to Bitcoin while hedging some of the risk by reducing the volatility typically associated with the purchase of a stand-alone Bitcoin. Historically, Bitcoin has been extremely volatile, which, for many investors, may make it a difficult investment. While the Shares are not intended to replicate a direct investment in Bitcoin, they seek to provide investors with exposure to Bitcoin with substantially lower volatility than a direct investment in Bitcoin and without the uncertain and often complex requirements relating to acquiring and/or holding Bitcoin.

8742 – Management Consulting Services

(1) GreenBox POS — January 29, 2021

GreenBox POS is a technology company that develops, markets and sells innovative blockchain-based payment solutions, which we believe will lead to major developments and advances in the payment solutions marketplace. Our core focus is to develop and monetize disruptive blockchain-based applications, integrated within an end-to-end suite of financial products, capable of supporting a multitude of industries. Our proprietary, blockchain-based ecosystem is designed to facilitate, record and store a virtually limitless volume of tokenized assets, representing cash or data, on a secured, immutable blockchain-based ledger.

(2) GBT Technologies Inc. — April 28, 2021

On September 14, 2018, the Company entered into an Exclusive Intellectual Property License and Royalty Agreement (the “GBT License Agreement”) with GBT, a fully compliant and regulated cryptocurrency exchange platform that currently operates in Costa Rica as a decentralized cryptocurrency platform, pursuant to which, among other things, the Company granted to GBT an exclusive, royalty-bearing right and license relating intellectual property relating to systems and methods of converting electronic transmissions into digital currency as reflected in that certain patent filed with the United Stated Patent and Trademark Office on or about June 14, 2018 […]. Pursuant to the GBT License Agreement, the Company granted GBT an exclusive worldwide license to use the Digital Currency Technology to make, use, sell, lease or otherwise commercialize and dispose of products and devices utilizing the Digital Currently Technology.

7371 – Computer Programming Services

(1) Turing Holding Corp. — August 20, 2021

Notable open source contributions from Thoughtworks include:

[…]

• TWallet. A next-generation digital wallet that supports general payment, digital currency transfer and duplex-offline payment;

(2) Sysorex, Inc. — November 17, 2022

TTM Digital has an evolving business model which is subject to various uncertainties. As digital assets and blockchain technology become more widely utilized on a mass scale, we anticipate

that the services and products associated with the technologies will continue to evolve. To successfully continue in the industry, our business model may need to evolve to reflect the trends of the industry. Over time, we may modify aspects of our business model relating to our strategy. We cannot offer any assurance that we will be successful or that the future industry or business operation changes will not result in harm to our business. We may not be able to manage growth effectively, which could damage our reputation, limit our growth and negatively affect our operating results. Management cannot provide any assurances that we will identify all emerging trends and growth opportunities in this business sector, and we may lose out on those opportunities to current or future competitors. As anticipated, any such circumstances could have a material adverse effect on our business, prospects, or operations.	
6199 – Publicly Traded Finance Services	
(1) Kryptoin Bitcoin ETF Trust Bitcoin and the Bitcoin Network Bitcoin is based on the decentralized, open source protocol of a peer-to-peer network (the “Bitcoin network”). No single entity owns or operates the Bitcoin network. Bitcoin is not issued by governments, banks or similar organizations. The infrastructure of the Bitcoin network is collectively maintained by a decentralized user base. The Bitcoin network is accessed through software, and software governs bitcoin’s creation, movement, and ownership. The value of bitcoin is determined, in part, by the supply of, and demand for, bitcoin in the global exchange markets for the trading of bitcoin, market expectations for the adoption of bitcoin as a decentralized store of value, the number of merchants and/or institutions that accept bitcoin as a form of payment and the volume of private end-user-to-end-user transactions.	October 15, 2019
(2) Kryptoin Ethereum ETF Trust Risks Associated with the Tax Treatment of ETH · Shareholders could incur a tax liability without an associated distribution from the Trust. · The tax treatment of ETH and transactions involving ETH for United States federal income tax purposes may change and the tax treatment of ETH and transactions involving ETH for state and local tax purposes is not settled. · A hard “fork” of the Ethereum Blockchain could result in Shareholders incurring a tax liability.	August 12, 2021