Titre: Title:	Development and Validation of Innovation Indicators to Help Companies in their Decision-Making Process Regarding the Introduction of new Technologies
Auteur: Author:	Pietro Cruciata
Date:	2025
Type:	Mémoire ou thèse / Dissertation or Thesis
Référence: Citation:	Cruciata, P. (2025). Development and Validation of Innovation Indicators to Help Companies in their Decision-Making Process Regarding the Introduction of new Technologies [Thèse de doctorat, Polytechnique Montréal]. PolyPublie. https://publications.polymtl.ca/67851/

URL de PolyPublie: PolyPublie URL:	https://publications.polymtl.ca/67851/
Directeurs de recherche: Advisors:	Catherine Beaudry, & Andrea Lodi
Programme: Program:	Doctorat en mathématiques de l'ingénieur

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Development and Validation of Innovation Indicators to Help Companies in Their Decision-Making Process Regarding the Introduction of New Technologies

PIETRO CRUCIATA

Département de mathématiques et de génie industriel

Thèse présentée en vue de l'obtention du diplôme de *Philosophiæ Doctor*

Mathématiques

Août 2025

POLYTECHNIQUE MONTRÉAL

affiliée à l'Université de Montréal

Cette thèse intitulée :

Development and Validation of Innovation Indicators to Help Companies in Their Decision-Making Process Regarding the Introduction of New Technologies

présentée par **Pietro CRUCIATA**

en vue de l'obtention du diplôme de *Philosophiæ Doctor*

a été dûment acceptée par le jury d'examen constitué de :

Michel DESMARAIS, président

Catherine BEAUDRY, membre et directrice de recherche

Andrea LODI, membre et codirecteur de recherche

Arash HAJIKHANI, membre

Josep DOMENECH, membre externe

DEDICATION

To my parents, my brother and my partner.

Thanks for your support and love during this journey.

ACKNOWLEDGEMENTS

I would first and foremost like to express my deepest gratitude to my supervisor, Catherine Beaudry. Your unwavering support, insightful guidance, and constant encouragement have been the bedrock of my doctoral journey. Your expertise and mentorship have not only shaped this thesis but have profoundly influenced my development as a researcher. I am incredibly grateful for the countless hours you dedicated to our discussions, for challenging my ideas, and for fostering an environment of intellectual curiosity and rigor. Thank you for believing in my potential and for guiding me through the complexities of academic research.

Thanks to my co-supervisor Andrea Lodi for his guidance during my doctoral journey and for his support in addressing major challenges along the way. His input during these key moments were valuable in helping me move forward and bring this thesis to completion.

I extend my sincere thanks to my thesis committee members, Arash Hajikhani and Josep Domenech and Michel Desmarais, for generously dedicating their time to review my work and for providing me with valuable feedback, critical insights, and constructive suggestions. Your perspectives have greatly enriched this thesis and further inspired me to continue working in this field.

My gratitude also goes to the faculty and staff at the department of Mathematics and Industrial Engineering. I am especially grateful to the administrative and technical staff, who were always ready to help with a smile and made the day-to-day logistics of this journey effortless.

To all my colleagues at the research chair, I offer my most sincere gratitude. Your guidance, support, and constant encouragement were invaluable throughout this journey. I will remember our daily lunches, which were always filled with laughter, camaraderie, and conversations on every imaginable topic. You transformed a professional endeavor into an unforgettable personal experience, making the chair feel like a second home. I look forward to the many wonderful moments we will surely share in the years to come.

I owe my deepest thanks to my family and friends for their unwavering love, patience, and encouragement. To my friends, thank you for the laughter, the diversions, and for reminding me that there is a world outside the lab/library. Your friendship has been a precious gift.

To my parents, thank you for your unconditional love, endless patience, and for instilling in me a passion for learning. You have been my constant cheerleaders, and your support has been my greatest strength.

Finally, to my partner, Kristel thank you for your incredible understanding, support, and for shouldering so much while I was immersed in my research. I could not have done it without you, this achievement is as much yours as it is mine.

RÉSUMÉ

Cette thèse aborde les limites critiques des indicateurs traditionnels pour mesurer l'innovation et la durabilité des entreprises. Alors que les mesures conventionnelles souffrent de décalages temporels, de coûts élevés, d'un manque de consensus et d'une couverture incomplète, les sites web d'entreprises offrent une source de données riche et évoluant potentiellement en temps réel. Cependant, les tentatives précédentes d'exploitation des données web ont été freinées par des lacunes méthodologiques, donnant des résultats mitigés. Cette recherche comble cette lacune en répondant à la question générale suivante : Les signaux des sites web représentent-ils des mesures valides et fiables de la qualité sous-jacente de l'innovation et de la durabilité que les organisations tentent de communiquer ?

Pour répondre à cette question, cette thèse développe et valide de nouvelles méthodologies basées sur le traitement du langage naturel (TAL ou *Natural language processing* – NLP – en anglais) à travers trois études distinctes.

La première étude analyse les sites web de 1 110 entreprises certifiées B-Corp, en utilisant un modèle de classification de texte sans apprentissage supervisé (*Zéros-Shot Text Classification* – ZSTC) pour créer des indicateurs environnementaux basés sur le web. Les résultats démontrent que ces indicateurs, combinés à des métadonnées classiques sur l'entreprise, expliquent 57 % de la variance de l'indice de performance environnementale officiel de B-Lab. Ceci confirme que les signaux de durabilité dans les sites web peuvent refléter de manière crédible des performances d'entreprise tangibles et validées par des tiers.

La deuxième étude examine les signaux liés à l'innovation en analysant les sites web de 5 696 entreprises canadiennes de la plateforme CrunchBase. En utilisant un nouveau pipeline combinant la génération augmentée par la recherche (*Retrieval-Augmented Generation* – RAG) et la modélisation de sujets, la recherche montre que les signaux des sites web, tels que l'expérience des fondateurs·rices et les annonces de financement antérieures, sont significativement corrélés avec la capacité d'une entreprise à obtenir des capitaux privés. Ceci constitue l'une des premières applications de la théorie du signal numérique (utilisant les sites web) aux résultats de financement privé.

La troisième étude compare les indicateurs dérivés du web avec les données de l'enquête sur l'innovation et les stratégies d'entreprise (EISE ou *Survey of Innovation and Business Strategy* –

SIBS – en anglais) officielle. Tout en révélant un écart quantitatif dans les déclarations, les résultats montrent un fort alignement thématique pour les signaux stratégiquement importants, en particulier ceux concernant les collaborations et les innovations ayant un bénéfice environnemental, renforçant le rôle des sites web en tant qu'outils de communication organisés.

Cette thèse contribue principalement à la théorie du signal, démontrant que la fiabilité d'un signal numérique dépend de son contexte stratégique. Dans les domaines fortement surveillés comme la durabilité, les sites web agissent comme un miroir, reflétant des performances vérifiables. Dans les domaines compétitifs comme la levée de fonds, ils fonctionnent comme une lentille, projetant des qualités qui attirent les capitaux. Sur le plan méthodologique, ce travail introduit des cadres de TAL généralisables et peu coûteux qui fournissent des données granulaires et opportunes pour compléter les enquêtes traditionnelles. Sur le plan pratique, ces outils ont le potentiel d'offrir aux décideurs politiques, aux investisseurs et aux gestionnaires un moyen de suivre les tendances des entreprises en temps quasi réel, permettant une prise de décision plus agile et fondée sur des données émises par les entreprises elles-mêmes.

ABSTRACT

This thesis addresses the critical limitations of traditional indicators for measuring corporate innovation and sustainability. While conventional metrics suffer from time lags, high costs, a lack of consensus, and incomplete coverage, corporate websites offer a rich, real-time data source. However, previous attempts to leverage web data have been hampered by methodological shortcomings, yielding mixed results. This research fills this gap by answering the following general question: Do website signals represent valid and reliable measures of the underlying innovation and sustainability quality that organizations are attempting to communicate?

To answer this question, this thesis develops and validates novel methodologies based on advanced Natural Language Processing (NLP) across three distinct studies.

The first study analyzes the websites of 1,110 B-Corp certified companies, using a Zero-Shot Text Classification (ZSTC) model to create web-based environmental indicators. The findings demonstrate that these indicators combined with basic company metadata explain 57% of the variance in the official B-Lab environmental performance index, confirming that sustainability signals on websites can credibly reflect tangible, third-party validated corporate performance.

The second study investigates innovation-related signals by examining the websites of 5,696 Canadian firms from the CrunchBase platform. Using a novel pipeline combining retrieval-augmented generation (RAG) and topic modeling, the research shows that website signals, such as founder experience and prior funding announcements, are significantly correlated with a firm's success in securing private capital. This marks one of the first applications of website signaling theory to private funding outcomes.

The third study compares web-derived indicators with data from the official Survey of Innovation and Business Strategy (SIBS). While revealing a quantitative discrepancy in reporting, the results show a strong thematic alignment for strategically important signals, particularly concerning environmental products and collaborations, reinforcing the role of websites as curated communication tools.

The primary contribution of this thesis is to Signalling Theory, demonstrating that the reliability of a digital signal is contingent on its strategic context. In high-scrutiny domains like sustainability, websites act as a mirror, reflecting verifiable performance. In competitive domains like fundraising,

they function as a lens, projecting qualities that attract capital. Methodologically, this work introduces generalizable and cost-effective NLP frameworks that provide timely, granular data to supplement traditional surveys. Practically, these tools have the potential to offer policymakers, investors, and managers a way to monitor corporate trends in near real-time, enabling more agile and evidence-based decision-making.

TABLE OF CONTENTS

DEDICATION III

ACKNOWLEDGEMENTS IV

RÉSUMÉ.......... VI

ABSTRACT VIII

TABLE OF CONTENTS X

LIST OF TABLES III

LIST OF FIGURES.......... V

LISTE OF SYMBOLS AND ABBREVIATIONS VI

LIST OF APPENDICES X

CHAPTER 1 INTRODUCTION.......... 1

CHAPTER 2 LITERATURE REVIEW.......... 8

2.1 Innovation: concept definition and evolution.......... 8

2.2 Sustainable Innovation 13

2.3 Financing innovation.......... 19

2.3.1 Private Equity.......... 21

2.4 Innovation Indicators.......... 23

2.5 New innovation indicators 28

2.6 How to study the web: Webometrics and Web mining.......... 30

2.7 Signalling Theory.......... 32

2.7.1 Signalling theory in economy and management 34

2.7.2 Signalling theory in sustainability.......... 39

2.8 New innovation indicators using websites.......... 43

2.8.1 Early-Heuristics phase.......... 43

2.8.2 Classical Machine Learning phase 45

2.8.3 Advanced NLP phase 48

2.9 Sustainability measure using website 50

CHAPTER 3 RESEARCH DESIGN 52

3.1 Research question and propositions 52

3.2 Objectives 55

3.3 Methodologies 55

3.3.1 Data 56

3.3.2 Web Crawler 58

3.3.3 Article 1 59

3.3.4 Article 2 65

CHAPTER 4 ARTICLE 1: FIRST IMPRESSIONS ON SUSTAINABLE INNOVATION MATTER: USING NLP TO REPLICATE B-LAB ENVIRONMENTAL INDEX BY ANALYZING COMPANIES' HOMEPAGES 81

4.1 Introduction 82

4.2 Literature review 84

4.3 Data and Methodology 89

4.3.1 Data 89

4.3.2 Methodology 91

4.4 Web-based variable construction 95

4.4.1 Zero-Shot Text Classification 95

4.4.2 Correlation Tests 97

4.4.3 Principal Component Analysis 100

4.5 Regression Results 103

4.5.1 Direct effects 103

4.5.2 Moderating effects 105

4.6 Discussion and conclusion 116

4.7 Limitation and future works 119

CHAPTER 5 ARTICLE 2: THE DIGITAL PITCH: A HYBRID AI APPROACH TO DECODING FUNDING SIGNALS ON CORPORATE WEBSITES 122

Abstract 122

5.1 Introduction 123

5.2 Literature review 125

5.3 Data 128

5.4 Methodology 129

5.5 Top-down approach 129

5.5.1 Website Analysis using LLM 129

Regression analysis 134

5.6 Bottom-up approach 135

5.6.1 BERTopic 135

5.6.2 Supervised methods 137

5.7 Results 139

5.7.1 Top-down approach 139

5.7.2 Bottom-up approach 147

5.8 Discussion 152

5.9 Conclusion 155

CHAPTER 6 THE DISCREPANCY BETWEEN PUBLIC DISCLOSURE AND PRIVATE REPORTING: VALIDATING WEB-BASED INNOVATION INDICATORS AGAINST SURVEY DATA 158

6.1 Introduction 158

6.2 Methodology 158

6.2.1 Data 158

6.2.2 Methods 159

6.3 Results 161

6.3.1 New product introduction..162

6.3.2 Introduction of environmental product 162

6.3.3 Government programs...163

6.3.4 Collaboration 164

6.4 Discussion and Conclusion 166

CHAPTER 7 DISCUSSION AND CONCLUSION 168

7.1 Discussion 168

7.2 Conclusion...170

7.3 Limitations 171

7.4 Future research 172

REFERENCES...174

APPENDICES..239

APPENDIX A SUPPLEMENTAL MATERIAL ARTICLE 1 239

APPENDIX B SUPPLEMENT MATERIAL ARTICLE 2 246

APPENDIX C SUPPLEMENT MATERIAL CHAPTER 6 254

LIST OF TABLES

Table 4.1 Labels used in the Zero-Shot Text Classification 93

Table 4.2 Zero-Shot Text Classification (ZSTC) results 96

Table 4.3 Pearson Correlation results 98

Table 4.4 PCA solution and factor loadings 101

Table 4.5 Moderating effect of the industry category for the EnergEff factor 106

Table 4.6 Moderating effect of the industry categories for the ManTransSaf 109

Table 4.7 Basic regression results and moderating effect of industry 110

Table 4.8 Regression results exploring the moderating effect of size 113

Table 4.9 Regression results exploring the moderating effect of country 115

Table 5.1: Hyperparameter used in the RAG. 132

Table 5.2: Prompts used to extract the information 133

Table 5.3: BERTopic parameters tested and their meaning 136

Table 5.4: Confusion Matrix 138

Table 5.5: Regression results (odds ratios) and moderating effect of industry. 144

Table 5.6: The settings of the two best BERTopic models. 148

Table 5.7: Comparison BERTopic 1 and BERTopic 2 using XGBoost. 149

Table 5.8:Comparison between XGBoost, Random Forest and a Neural Network on the BERTopic 1. 150

Table 6.1: Hyperparameters used in the RAG 161

Table 6.2: Announced product or service innovations on the websites versus the percentage that reported such activities in the SIBS survey, including a breakdown by the type of innovation. 162

Table 6.3: Rate of reporting on environmental innovations on company websites with data from the SIBS survey regarding product or process innovations with environmental benefits, and details whether the web-disclosed innovations were newly introduced by the company or adaptations of existing solutions 163

Table 6.4: Utilization of government programs reported in the SIBS survey and the acknowledgment of such support on company websites, with a breakdown of the types of funding mentioned.. 164

Table 6.5: Frequency and type of business cooperations mentioned on corporate websites against those reported in the SIBS survey. 165

Table A. 1 Variable description 239

Table A. 2 Descriptive statistics... 241

Table A. 3 Correlation matrix 242

Table A. 4 Correlation table (continued) 243

Table A. 5 Base regression and robustness check.. 244

Table B 1 Description variables. 246

Table B 2 Descriptive statistics dependent variables 248

Table B 3 LLM's Prompt to create collaboration score. 250

Table B 4 LLM's Prompt to create fundings and founder experience. 251

Table B 5 LLM prompt template for topic interpretation, showing example for Topic 2. 252

Table B 6 Retrievers' prompts. 253

Table C 1 Prompts used to create web-indicators. 254

LIST OF FIGURES

Figure 3.1 Methodological pipeline Paper 1 60

Figure 3.2 Methodological pipeline Paper 2 66

Figure 3.4 Retrieval Augmented Generation 72

Figure 4.1 Summary of the propositions to be tested. 89

Figure 4.2 Pre-processing steps 91

Figure 5.1 Dataset creation steps 129

Figure 5.2 Pipeline RAG 131

Figure 5.3 Top 10 Most Important Topics – Random Forest Results. 150

Figure 5.4 Partial Dependence plot of the 4 most important topics. 151

Figure 6.1 Dataset Construction Process 159

Figure 6.2 Pipeline RAG 161

LISTE OF SYMBOLS AND ABBREVIATIONS

ADASYN	Adaptive Synthetic Sampling Approach for Imbalanced Learning
AGI	Artificial General Intelligence
AI	Artificial Intelligence
AUC	Area Under the Curve
BART	Bidirectional and Auto-Regressive Transformer
BERT	Bidirectional Encoder Representations from Transformers
BM25	Best Matching 25
c-TF-IDF	Class-based TF–IDF
CFA	Confirmatory Factor Analysis
CIS	European Community Innovation Survey
CoT	Chain of Thought
CSR	Corporate Social Responsibility
CSS	Cascading Style Sheets
EIS	European Innovation Scoreboard
ERP	Emissions Reduction Plan
ESG	Environmental, Social, and Governance
FA	Factor Analysis
FPR	False Positive Rate
FPT	Federal, Provincial or Territorial
GDP	Gross Domestic Product
GPT	Generative Pre-Training
GRI	Global Reporting Initiative
GRPO	Group-Relative Policy Optimisation

HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
HTML	Hypertext Markup Language
ICT	Information and Communication Technologies
IP	Intellectual Property
IPO	Initial Public Offer
IR	Information Retrieval
ISO	International Organization for Standardization
IT	Information Technologies
K-means SMOTE	Oversampling for Imbalanced Learning Based on K-Means and SMOTE
KMO	Kaiser-Meyer-Olkin
LDA	Latent Dirichlet allocation
LLM	Large Language Model
MIP	Mannheimer Innovations panel
ML	Machine Learning
MLP	Multilayer Perceptron
MST	minimum spanning tree
MTMM	Multitrait-Multimethod
NAICS	North American Industry Classification System code
NLI	Natural Language Inference
NLP	Natural Language Processing
NLU	Natural Language Understanding
OECD	Organization for Economic Cooperation and Development
OLS	Ordinary Least Squares
OOB	Out-Of-Bag

PCA	Principal Component Analysis
PDP	Partial Dependence Plot
POS	Part of Speech
PR	Precision-Recall
R&D	Research and development
RAG	Retrieval-augmented generation
RF	Random forest
RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback
ROC	Receiver Operating Characteristic
ROSE	Random Over-Sampling Examples
SBICs	Small Business Investment Companies
SDG	Sustainable Development Goals
SI	Sustainable Innovation
SIBS	Survey of Innovation and Business Strategy
SME	Small Medium Enterprise
SMOTE	Synthetic Minority Over-sampling Technique
SNLI	Stanford Natural Language Inference
SVMSMOTE	Borderline Over-sampling for Imbalanced Data Classification
TBL	Triple Bottom Line
TF-IDF	Term Frequency – Inverse Document Frequency
TPR	True Positive Rate
UMAP	Uniform Manifold Approximation and Projection
URL	Uniform Resource Locator

VC	Venture Capital
VIF	Variance Inflation Factor
WCED	World commission on environment and development
XGBoost	Extreme Gradient Boosting
ZSL	Zero Shot Learning
ZSTC	Zero-Shot Text Classification
2SLS	two-stage least squares

LIST OF APPENDICES

APPENDIX A SUPPLEMENTAL MATERIAL ARTICLE 1 239

APPENDIX B SUPPLEMENT MATERIAL ARTICLE 2 246

CHAPTER 1 INTRODUCTION

Innovation plays a central role in driving high productivity, generating revenue streams, and fostering an environment of continuous learning and adaptation at both micro and macroeconomic levels. It helps firms overcome resource limitations and develop new markets, leading to economic growth and stronger competitive positions (OECD, 1997; Grübler et al., 1999). As technologies and markets evolve, innovation capabilities allow firms and nations to continuously innovate to ensure their products and services remain relevant and competitive (Bate et al., 2023; Pisano, 2015). The resulting competitive advantages enhance both firm survival and economic performance (Edeh et al., 2020). However, the view of innovation as solely economic has evolved significantly. Over the past 15 years, the academic discourse surrounding "sustainability" has intensified, bringing together economic, environmental, and social objectives. This evolution has prompted management practices to expand beyond traditional economic frameworks and address broader structural and systemic challenges, including human rights considerations, alternative liberalization approaches, and environmental impact mitigation (Cillo et al., 2019).

The integration of sustainability considerations into innovation frameworks has become particularly urgent given the scale of contemporary environmental challenges. Industries and societies face escalating environmental challenges, including climate change, resource consumption, and ecological degradation. Sustainable innovations are crucial to overcome these escalating environmental challenges, which threaten global stability. To face these emergencies, national and international agreements were enacted. At the international level, the 2030 UN Agenda[1] calls for every country, at every income level, to align national strategies, budgets, and reporting systems with the SDGs and to mobilize government, business, civil society, and citizens in a global partnership.

At the national level, for instance, Canada is implementing a comprehensive set of national regulations, aligned with international standards, to drive transparency, accountability, and capital flows into green investment. These measures are designed to support the country's net-zero goals, prevent greenwashing, and position Canada as a global leader in sustainable finance. Specifically, Canada's 2030 Emissions Reduction Plan (ERP), released on 29 March 2022, lays out the first

[1] https://sdgs.un.org/2030agenda

roadmap under the Net-Zero Emissions Accountability Act to cut national greenhouse-gas emissions 40–45 percent below 2005 levels by 2030 and reach net-zero by 2050. The plan demonstrates that sustainability is not just about risk mitigation, but unlocks significant economic opportunities. Backed by $9.1 billion in new federal spending, the ERP combines economy-wide tools (e.g., strengthened carbon pricing, clean-fuel standards and emissions modeling) with sector-specific measures intended to drive both climate action and economic growth[2].

This push towards sustainability, and the need for rapid scaling and technological adaptation, has created capital-intensive challenges requiring new funding approaches. The investment landscape has responded accordingly, with private equity assets having hit an all-time high of $10.8tn in 2024, reflecting an 11.6% increase[3]. Among private equity, venture capitals (VCs) are the primary source of early-stage funding for young companies with high growth potential, providing, in addition to funds, strategic guidance and specialized expertise essential for navigating these complex challenges (Metrick & Yasuda, 2021; Van den Heuvel & Popp, 2023). This evolution has made VC funding critically important for bridging the gap between nascent innovative ideas and commercially successful enterprises (Bellavitis et al., 2017). KPMG's Venture Pulse further corroborates this, noting that the US alone saw $209 billion in VC investment in 2024, the third-highest total in the past 20 years. Global VC investment reached $368 billion, with AI being a significant driver of renewed investor interest[4]. However, effectively directing these substantial investments, whether the $368 billion in global VC funding or the billions in government sustainability initiatives, requires robust measurement systems that can help policymakers and entrepreneurs monitor and validate the impact of their actions. Unfortunately, current innovation and environmental indicators face significant limitations that compromise their effectiveness in guiding these critical investment decisions.

On the sustainability side, a key insight from empirical studies is that, despite the proliferation of sustainability metrics, there remains a lack of consensus on standard measures and definitions. For example, in green supply chain management, researchers have documented the existence of over 2500 unique metrics, many addressing similar issues with slight variations. This heterogeneity

[2] https://www.canada.ca/en/environment-climate-change/news/2022/03/2030-emissions-reduction-plan--canadas-next-steps-for-clean-air-and-a-strong-economy.html

[3] https://www.ocorian.com/news-press-releases/global-private-equity-assets-hit-record-108-trillion-following-2022-dip

[4] https://kpmg.com/xx/en/media/press-releases/2025/01/2024-global-vc-investment-rises-to-368-billion-dollars.html

complicates cross-industry benchmarking and necessitates the development of integrated measurement frameworks that can consolidate and contextualize diverse metrics into coherent performance assessments (Ahi & Searcy, 2015). These frameworks are often criticized because they lack independent verification and rely on voluntary self-reported data, which can be subjective and inconsistent across companies (Antolín-López et al., 2016). This issue is partially addressed by third-party certification such as B Corporation which represents a more reliable and trustworthy signal for stakeholders (Blasi & Sedita, 2022; Cormier & Magnan, 2015).

On the innovation side, innovation indicators serve as fundamental instruments for both quantitative and qualitative assessment of innovation processes within organizations and economies, enabling critical measurement and evaluation capabilities that drive strategic decision-making (OECD, 2018). From a public policy perspective, these robust indicators are indispensable tools for designing, implementing, and evaluating programs that stimulate technological advancement and economic development, allowing policymakers to identify critical bottlenecks in national innovation systems, monitor the effectiveness of R&D funding schemes, and develop targeted initiatives that support both high-tech sectors and broader, non-R&D based innovation activities essential for inclusive growth (Bate et al., 2023; Janger et al., 2017). Beyond policy applications, innovation indicators enable policymakers to evaluate proposals from different applicants for innovation projects more effectively and assess the progress of subsidized initiatives, while also assisting investors in making informed decisions about funding new ventures.

For organizations specifically, these indicators provide managers with essential tools to promote policies that foster growth, support individual and team creativity, and enhance creative team building, making them indispensable for managing and controlling the numerous innovative ideas and concepts submitted to companies (Pirola-Merlo & Mann, 2004). Furthermore, well-defined selection criteria become equally important for efficient resource allocation and performance evaluation throughout each phase of the innovation process, offering managers a deeper understanding of significant determinants that contribute to superior company performance, even in turbulent market environments (Dewangan & Godse, 2014; Evanschitzky et al., 2012; Hult et al., 2004).

Despite these benefits, innovation measurement faces significant challenges. Early measurement approaches encountered significant limitations, as literature-based output indicators obtained from trade journals and expert-based lists of major innovations suffered from bias toward product innovation and economically significant innovations, potentially overlooking other important forms of innovative activity (Geroski et al., 1997; Kleinknecht et al., 1993). These measurement challenges have become increasingly critical as public policy focuses more on promoting innovation to stimulate economic growth, employment, and ecological sustainability, creating an urgent need to measure and assess innovation and technological change while expanding knowledge about the driving forces behind innovation and its socio-economic consequences.

These measurement challenges are compounded by fundamental issues with traditional data collection methods. Moreover, traditional data sources are often incomplete, with representative samples much smaller than the actual firm population or lack specificity. Questionnaire-based surveys, particularly large-scale ones like the biennial European Community Innovation Survey (CIS) or annual Mannheimer Innovations panel (MIP), suffer from poor regional granularity, limited coverage, timing delays, and high operational costs (Axenbeck & Breithaupt, 2021). The situation has worsened as the proliferation of low-cost web-based surveys has overwhelmed firms, causing response rates to fall to just 5-10% in most cases.

In response to these limitations of traditional measurement approaches, emerging big data analytics and composite indices have emerged as promising alternatives that offer increasingly nuanced insights into complex innovation dynamics that were previously beyond the reach of traditional metrics. New indicators leverage large, often unstructured data sources such as web content, providing more timely and comprehensive insights that enable real-time analysis and fill the gaps left by traditional surveys (Gök et al., 2015). Among the different big data sources, websites legitimized by Signalling Theory and Webometrics represent a promising source for building indicators.

However, despite this theoretical promise, early attempts to implement web-based innovation indicators have encountered significant practical challenges. While measuring innovation “signals” from corporate websites initially seemed challenging, researchers in innovation and technology management have achieved promising yet limited results by developing new indicators from large text datasets (e.g., Blazquez & Domenech, 2018; Gök et al., 2015). Indeed, previous attempts to develop web-based innovation indicators have yielded poor results, with studies finding either no significant correlation or only weak correlations between website-derived metrics and traditional survey-based measures of innovation activities (e.g., Gök et al., 2015; Héroux-Vaillancourt et al., 2020).

These disappointing results can be attributed to methodological constraints that hinder their effectiveness and, in turn, limit the development of web-based indicators. Many researchers analyze company websites using keyword searches and frequency counts, often employing weighting schemes like TF-IDF. These approaches face two significant challenges. First, polysemy, i.e., single words carry multiple meanings (for example, "bank" can refer to a riverbank or financial institution). Second, semantic ellipsis, i.e., concepts are implied rather than explicitly stated (for instance, "collaboration" might be conveyed through phrases like "joint venture" or "working together"). Natural Language Processing (NLP) offers advanced methodologies to address these issues, but successful implementation requires social scientists to bridge the methodological gap with computer scientists.

Recognizing these limitations and the opportunity presented by advanced NLP methodologies, this thesis addresses a key research question: whether website signals represent valid and reliable measures of the underlying innovation and sustainability qualities that companies are attempting to communicate.

This thesis answers this question by closing the methodological gap between social science and computer science by leveraging the latest NLP techniques. Specifically, this thesis pursues these objectives through two research papers.

The first published article investigates the potential for developing web-based environmental culture indicators by analyzing signals extracted from the homepages of 1110 Canadian and American companies that obtained the B-Corp certification. The goal of the paper is to assess the proposed method's ability to generate indicators that can serve as proxies for real environmental measures by leveraging the homepage content. In the paper, we performed a Zero-Shot Text Classification (ZSTC) using a BERT-type model, followed by a regression analysis to test the ability of these web- based indicators to replicate the B-Lab environmental index and to understand the factors driving the results. This pilot study explains 57 % of the variance of the B-Lab environmental index using the results of the ZSTC score and companies' characteristics. This research makes two significant contributions. The research makes two significant contributions, advancing both theory and methodology in Sustainability Research and Signal Theory.

Theoretically, it establishes a correlation between web-based indices and the B-Lab environmental index, demonstrating that website communication can reflect environmental performance. The regression analysis clarifies which factors correlate with higher B-Lab environment scores. The research also addresses Connelly et al. (2011) question about signal validity by confirming that meaningful sustainability signals can be extracted from company websites. Methodologically, it introduces a generalizable methodology for studying the performance of companies through their websites without the need for heavy pre-processing, significantly reducing the time and cost of research. Furthermore, the method could provide policymakers with a real-time landscape to create and fine-tune policies about specific topics, partially addressing the problems associated with questionnaire-based surveys.

The second article explores how signals present on companies' websites relate to their success in attracting private funding. Guided by signalling theory, we conducted a pilot study using a dataset of 5696 Canadian companies sourced from the CrunchBase business intelligence platform. We then applied a combined top-down pipeline based on retrieval-augmented generation (RAG) and bottom-up topic modeling to mine unstructured website text and reproduce, or even extend, insights drawn from structured datasets. Given that websites are updated more frequently than traditional data sources, the method provides timelier signals of funding, collaboration, founder background and other factors, giving scholars, investors and policymakers an earlier view of venture dynamics. To our knowledge, this is the first paper to deploy both approaches together and one of the first to use a RAG framework and large language models to create targeted web indicators. From a signalling-theory perspective, our findings confirm that the analyzed signals (funding announcements and founders' human capital) affect a firm's likelihood of securing private capital, marking, to our knowledge, the first application of website-based signalling theory to private-funding outcomes.

Finally, chapter 6 represents a third contribution. This chapter details a study that compares aggregated results from the Survey of Innovation and Business Strategy (SIBS) with indicators gathered from the websites of a subset of the surveyed Canadian companies. Building on a previously established methodology of paper 2, the research aims to evaluate the potential of using corporate websites as a valuable data source, for near real-time indicators related to innovation and environmental signals, to monitor industry trends. The validity of these web-derived indicators is tested against the representative data from the SIBS, which was chosen for its comprehensive sample and its direct questions on innovation and environmental aspects. The first results we get suggest general differences between SIBS and websites, with alignments on aspects that seems deemed important to communicate and highlights such as collaboration with clients and environmental products and services.

CHAPTER 2 LITERATURE REVIEW

This chapter systematically presents a comprehensive literature review to establish a solid theoretical foundation and contextual understanding necessary for exploring the core themes addressed in subsequent chapters.

The chapter begins by exploring the concept of innovation, outlining its historical evolution and various definitions. It then examines sustainable innovation, integrating Environmental, Social, and Governance (ESG) considerations and the Triple Bottom Line framework to explain how sustainability has become central to modern innovation. Building on these foundational concepts, the next section dives into financing innovation and private equity, particularly focusing on venture capital (VC). This inclusion logically connects innovation and sustainable innovation to the financial mechanisms that enable and drive these activities, emphasizing the critical role of VC in fostering innovation and entrepreneurial ventures. With the conceptual framework and financial enablers established, the chapter then examines innovation indicators. This section critically assesses traditional methodologies for measuring innovation, highlighting their strengths and limitations. The discussion then turns to the emergence of novel, data-driven methodologies, notably big data approaches, designed to overcome the constraints of traditional innovation surveys. Particular attention is given to new innovations and environmental indicators derived from company websites. The use of web-based data is justified through Webometrics and Signaling Theory, which provide theoretical frameworks for understanding how companies communicate their qualities online. The chapter concludes by synthesizing the methodological insights, emphasizing the complementary potential of traditional surveys and big data methods.

2.1 Innovation: concept definition and evolution

Innovation is a cornerstone of economic growth and competitiveness. At microeconomic level, it is vital for creating and sustaining competitive advantages, entering new markets, and adapting to global competition (OECD, 1997; Stock et al., 2002). At the macroeconomic level, it drives productivity, output, and employment (Asheim & Isaksen, 1997; Michie, 1998). Advancing technological knowledge through innovation has long been recognized as the most significant factor for long-term economic progress (Grübler et al., 1999). Beyond economics, innovation is increasingly essential for addressing global challenges like climate change, biodiversity loss, and social inequality (Hekkert & Negro, 2009; Mazzucato, 2022).

The modern study of innovation traces back to Joseph Schumpeter, who in the early 20th century identified it as a primary driver of economic development (Schumpeter, 1934). Early research emphasized technological innovation, particularly in manufacturing, reflecting the industrial focus of that era. This perspective laid the groundwork for understanding innovation's economic impact at both micro (firm-level) and macro (national) scales. Before standardized guidelines, studying innovation was hampered by inconsistent data and definitions (Becheikh et al., 2006). The Oslo Manual, first published in 1992 by the Organization for Economic Cooperation and Development (OECD) and Eurostat, addresses this challenge by providing a standardized framework for measuring innovation. The original 1992 edition focused exclusively on technological innovation, categorizing it into two distinct types: product innovation and process innovation, defined as follows:

"A product innovation is the commercialisation of a technologically changed product. Technological change occurs when the design characteristics of a product change in ways which deliver new or improved services to consumers of the product" (OECD, 1992, p. 10)

"A process innovation occurs when there is significant change in the technology of the production of an item. This may involve new equipment, new management and organisation methods, or both" (OECD, 1992, p. 10)

In other words, the 1992 edition defined innovation as new or significantly improved technological products or processes that are either introduced to the market or implemented in production. This definition distinguished between product innovation (such as microprocessors or incremental improvements like portable cassette players) and process innovation (such as new production methods), with a clear focus on manufacturing sectors. A significant expansion occurred in the 2005 third edition, which broadened the definition to include non-technological innovations, specifically marketing and organizational innovations. The revised definition reads as follows:

"Innovation is the implementation of a new or significantly improved product (good or service), or process, a new marketing method, or a new organizational method in business practices, workplace organization or external relations." (OECD, 2005, p.46).

The 2005 edition's revised definition explicitly incorporated marketing methods and organizational methods as distinct innovation categories, emphasizing that innovation extended beyond technological change. The inclusion of marketing innovation was particularly significant, as it acknowledged that market-driven innovations could provide substantial competitive advantages for firms. Additionally, the two key characteristics of novelty and implementation provide a good foundation to understand what is meant by innovation more generally. However, critics argued that this expanded definition was overly broad and consequently difficult for practitioners to apply in research (Héroux-Vaillancourt, 2023). These concerns led to the development of the latest edition of the Oslo Manual, the fifth edition, which defines innovation as:

"a new or improved product or process (or combination thereof) that differs significantly from the unit's previous products or processes and has been made available to potential users or brought into use" (OECD, 2018, p.33).

This broadened definition includes all sectors distinguishing innovation as an outcome (e.g., a new product) and a process (e.g., activities leading to innovation). This evolution reflects a broader, more inclusive understanding of innovation, supported by the European Community Innovation Survey (CIS) for cross-country data collection. One of the key changes in the fifth edition is the modification of the market introduction criterion for product innovations; instead of requiring innovations be introduced at economically significant prices, it allows for the broader notion of innovations simply "made available to potential users." By decoupling innovation from immediate performance metrics, the OECD framework thus avoids penalizing early-stage innovations that might yield delayed but substantial benefits (OECD, 2018).

The evolution of the innovation definition came alongside the conceptual evolution of the innovation process which has undergone significant transformation, reflecting the increasing complexity of technological, economic, and social systems. Initially, innovation was understood through the "linear model," a simplistic sequential progression from basic research to applied research, technological development, and eventual market diffusion (Bush, 1945). Central in this model was basic research, which could lead to discoveries and be translated into materialized inventions (Kuznets, 1962). R&D investments were often conducted in isolation and were considered a significant asset, particularly for larger companies. Joseph Schumpeter's distinction between invention, innovation, and diffusion laid foundational insights, highlighting innovation as a dynamic economic force through "creative destruction" that disrupts existing industries (Schumpeter, 1934). His initial focus on entrepreneurial individuals evolved into recognizing large firms with robust R&D capabilities. This capacity for R&D allowed large firms to achieve substantial economies of scale, securing a distinct competitive advantage (Chandler, 1990). The "technology-push" model posited scientific breakthroughs, suggesting R&D investment would inherently yield technological advancements (Burns & Stalker, 1961; Nemet & Kammen, 2007). By the mid-20th century, this theory was contested by the "demand-pull" theory, emphasizing market needs over scientific progress alone as a key innovation driver, leading to broader debates and the emergence of comprehensive frameworks (Greenacre et al., 2012).

Subsequently, the mid-20th century brought critical refinements by economists such as Solow, Nelson, and Arrow. Solow (1957) demonstrated that technological change is a key driver of long-term economic growth, beyond what can be explained by capital and labor inputs. Nelson (1959) and Arrow (1962) highlighted a critical market failure in innovation: because knowledge tends to spill over to competitors, firms cannot fully appropriate the returns from their R&D investments, leading them to chronically underinvest in innovation. This insight highlights the need for public policy intervention (e.g., subsidies or intellectual property protections) to encourage socially optimal levels of R&D investment.

From the 1970s to the 1990s, innovation theory diversified into induced innovation, evolutionary economics, and path dependency. Induced innovation suggested relative price changes or scarcity in other factors drive technology development (Hayami & Ruttan, 1971), while evolutionary economics (Nelson & Winter, 1982) emphasized incremental innovations influenced by bounded rationality and uncertainty. Path dependency theory, articulated by Arthur (1994), illustrates how early technological choices can become well established through increasing returns and self-reinforcing mechanisms, leading to lock-in effects that may perpetuate suboptimal solutions despite the availability of superior alternatives.

By the late 20th century, systemic and iterative models like Kline's "chain-linked model" (Kline & Rosenberg, 1986) and frameworks by (C. Freeman & Perez, 1988) emphasized feedback loops and interactions between various innovation actors. National Innovation Systems (C. Freeman & Perez, 1988) and Porter's National Innovative Capacity (Porter, 1990) further underscored institutional and networked innovation dimensions.

R&D's role evolved significantly during this period. In the early decades, R&D was characterized by closed, linear processes dominated by large firms that could leverage economies of scale and maintain extensive internal research operations (Bush, 1945; Chandler, 1990). Over time, this approach shifted toward more multidisciplinary and integrated management practices that enhanced interaction between R&D activities and strategic business decisions (Wheelwright & Clark, 1992). Collaborative and open innovation models emerged, highlighting the necessity of partnerships among firms, academia, and governmental bodies, significantly reducing risks and accelerating technology commercialization (Cropper, 2008; Etzkowitz & Leydesdorff, 2000; Liyanage et al., 1999). This transition was facilitated by socio-economic factors, including the advent of globalized communication technologies and the rise of venture capital, which increased innovation mobility and accessibility (H. W. Chesbrough, 2003).

Open Innovation (OI), conceptualized by (H. W. Chesbrough, 2003), proposed extending R&D beyond organizational boundaries to leverage external knowledge and technology. OI includes three types of activities: inbound flows (bringing outside knowledge in), outbound flows (sharing internal knowledge out), and coupled activities (bi-directional collaborations) (H. Chesbrough & Bogers, 2014; Dahlander & Gann, 2010). These activities can involve financial mechanisms like intellectual property licensing or non-financial approaches such as technology scouting and community knowledge sharing. By facilitating broader knowledge dissemination and creating new commercialization opportunities, open innovation helps firms better manage knowledge spillovers that were previously beyond their control (H. Chesbrough, 2006; C. M. Christensen, 1997).

Despite these advancements, measuring innovation remains complex. Much of the literature relies on ex-post indicators (assessing innovation after it occurs) rather than ex-ante indicators (predicting innovation capacity) (Dziallas & Blind, 2019). This limits policymakers' ability to allocate resources efficiently. Additionally, the lack of precise data forces reliance on indirect factors (e.g., R&D spending) that influence innovation but do not measure it directly. A missing logical step here is the need for new methodologies to bridge this gap and enhance innovation assessment. Beyond economic benefits, innovation is critical for tackling global crises. Issues like global warming and social injustice demand innovative solutions (Levidow & Papaioannou, 2018). The Triple Bottom Line framework, i.e. the idea of balancing economic, social, and environmental outcomes, has emerged as a guide for directing innovation toward sustainable development. This shift highlights a missing link: a clear explanation of how innovation policies must align with sustainability goals to address these challenges effectively.

2.2 Sustainable Innovation

The widely accepted view that innovation stems from scientific research, technological advancements, business implementation, and market distribution has evolved significantly. Innovation now extends beyond merely enhancing market competitiveness or advancing technology across industries. Instead, it is increasingly recognized as a tool for addressing social challenges, improving quality of life, and promoting social and environmental well-being. Policymakers actively define and promote sustainable innovation (SI) as part of environmental, social, and governance (ESG) considerations. The concept of SI originates from the Brundtland Report, which closely tied sustainability to innovation goals (Zhu & Hua, 2017). The report defines sustainability as “meeting the needs of the present without compromising the ability of future generations to meet their own needs” (WCED, 1987, p. 15). Historically, companies perceived environmental strategies as conflicting with business growth and profitability (Andersen, 2004; Porter & Linde, 1995). However, Weale’s (1992) seminal paper influenced the development of the Triple Bottom Line (TBL) framework, introduced by Elkington in 1996, which has become a cornerstone of sustainability discourse.

The TBL framework advocates a balanced approach to economic, environmental, and social, often summarized as the “3Ps”: profit, planet, and people (Cohen et al., 2008; DiVito & Bohnsack, 2017; Elkington, 1997). These dimensions align with sustainable development principles: economic prosperity supports quality of life and productivity, environmental integrity ensures ecosystem health, and social equity guarantees fair resource access for all stakeholders (Cohen et al., 2008; Elkington, 1997). TBL stresses that a comprehensive sustainability assessment requires equal attention to all three dimensions, shifting away from a sole focus on economic profitability (Bansal & DesJardine, 2014; WCED, 1987). This evolution of TBL and the broader sustainability concept reflects a growing acknowledgment of the interdependence among economic, social, and environmental factors.

Despite extensive research, defining sustainability remains complex due to its diverse interpretations. Over time, several terms have emerged to describe sustainable innovation. The OECD (2010) defines eco-innovation as the development or improvement of products, processes, marketing methods, or organizational structures that reduce environmental impacts compared to alternatives, regardless of intent. Kemp & Pearson (2007) describe it as any new product or process that mitigates environmental risks, pollution, and resource consumption. Green innovation focuses on products and processes incorporating environmental management, energy efficiency, pollution prevention, and waste recycling (Y.-S. Chen et al., 2006). Sustainability-oriented innovation, according to Adams et al. (2016), involves intentional changes in organizational practices, philosophies, and strategies to achieve economic, social, and environmental goals. However, these terms are often used interchangeably in the sustainable innovation literature (e.g., Ben Arfi et al., 2018; Forsman, 2013; Hojnik & Ruzzier, 2016; Iñigo & Albareda, 2016; Karakaya et al., 2014). Increased research expands from a focus on technological and economic aspects to encompass broader societal and ecological concerns (Cohen et al., 2008; DiVito & Bohnsack, 2017; Elkington, 1997; Garriga & Melé, 2004). Incorporating the social dimension alongside environmental and economic considerations enables a more integrative examination of SI. Building on this perspective, understanding the drivers and motivations behind sustainable innovation is crucial (Afeltra et al., 2023).

Sustainable entrepreneurship has emerged as a critical driver of economic development, innovation, and social progress, integrating traditional entrepreneurial practices with environmental and social priorities through frameworks such as the Triple Bottom Line (TBL) approach (Carree & Thurik, 2010; D. Fischer et al., 2020; Muñoz & Cohen, 2018; Terán-Yépez et al., 2020). This integration has gained significant momentum through the United Nations' 2030 Development Agenda, which encourages global collaboration on sustainable business practices (Anand et al., 2021; Schaltegger et al., 2018). Consequently, managers increasingly recognize that sustainability initiatives create value not only for the environment and society but also for their organizations' long-term performance. Small and medium-sized enterprises (SMEs) represent a particularly interesting case in this context, as their inherent agility and close stakeholder relationships present both unique opportunities and challenges in achieving sustainable performance outcomes (Borah et al., 2022). Despite having limited resources and capacity to adopt new practices, SMEs can see marked improvements in sustainable performance when they secure access to finance (Khattak, 2020), while ownership structure shapes CSR disclosure among publicly listed firms (Nekhili et al., 2017). Flexibility and high degree of local embeddedness can translate into quicker adoption of sustainability practices. For instance, many SMEs have closer relationships with local communities and customers; this naturally facilitates CSR engagement through face-to-face interactions and immediate feedback mechanisms (Cantele & Zardini, 2020). In contrast, larger firms, despite having formal strategies for CSR, may experience delays in execution due to the need to navigate complex internal approval processes and diverse stakeholder interests spread across multiple geographies (Dias et al., 2019)

Despite its promise, the adoption of sustainable innovation remains slow due to several interconnected challenges. Del Río et al. (2010) identified key barriers including external market pressures, internal financial and technological constraints, and broader techno-economic challenges. Additionally, the complexity of sustainable innovation creates various externalities, such as knowledge spillovers that discourage corporate investment (Dybvig & Spatt, 1983; Rennings, 2000). However, certain factors can accelerate sustainable innovation adoption. Government regulations and stakeholder pressures serve as significant external drivers, while regulatory push/pull mechanisms provide structured incentives for implementation. Eco-innovation, involving technological, organizational, social, or institutional changes, often requires targeted regulatory support that goes beyond standard market or technological incentives (Rennings, 2000). When these regulatory drivers are combined with adequate organizational resources and unique firm capabilities, they can enhance competitiveness by improving operational efficiency, reducing environmental impact, and strengthening financial performance. Green et al. (1994) found a strong link between commercial pressures and regulatory influences, particularly in anticipating regulations, competing products, and potential market share losses. Building on this foundation, Horbach (2008) showed that environmental management tools and regulations positively influence eco-innovation, supporting the Porter Hypothesis (Porter & Linde, 1995), which suggests regulations can reduce pollution while enhancing profitability. Moreover, Hojnik and Ruzzier (2016) found that government incentives such as grants and tax breaks are a significant driver of firms' process eco-innovation. This aligns with findings by Horbach et al. (2012) who emphasized the need for tailored policies for different environmental impacts, noting cost savings, such as in energy consumption, as significant motivators. Similarly, anticipating strict regulations can spur innovation, offering firms a competitive edge and reducing future compliance costs (Christmann, 2000; Khanna et al., 2009). These findings are reinforced by evidence that firms subject to state-imposed regulation are more likely to pursue sustainable innovations (Doran & Ryan, 2012; Horbach et al., 2012).

Beyond regulatory compliance, stakeholder pressure also significantly shapes corporate SI. Freeman, (1984) introduced stakeholder theory to highlight the role of key groups in organizational studies, aiding in opportunity identification, responses to external pressures, and demand management. This theory aligns closely with TBL, emphasizing stakeholders' importance in sustainability initiatives. Stakeholders increasingly pressure firms to adopt TBL practices, valuing their ability to generate benefits across economic, social, and environmental dimensions (Clarkson, 1995; R. E. Freeman, 1984; Hörisch et al., 2014). Stakeholders are categorized as primary (e.g., employees, customers, the environment, and society) or secondary (e.g., media, interest groups), with primary stakeholders being essential for organizational sustainability (Evans et al., 2017; Mitchell et al., 1997). Stakeholders significantly influence corporate social responsibility (CSR[5]) and environmental management practices, underscoring the need to address their demands (Buysse & Verbeke, 2003; S. Sharma & Henriques, 2005). D. Li et al. (2017) demonstrated that the pressures to maintain legitimacy positively affect product and process innovations. Stakeholders, including customers and investors, demand that businesses adopt sustainable practices and assess performance using the TBL framework (D. Fischer et al., 2020). Du et al. (2018) showed that integrating green practices with customers and suppliers, supported by internal processes, enhances green innovation performance. The literature highlights stakeholder pressures for legitimacy and the pursuit of profitability as key drivers of corporate green innovation, leading to practices like green technologies, supply chain management, and corporate culture. Thus, sustainable entrepreneurs benefit greatly from strong relationships with stakeholders who share sustainability values and can impact economic, ecological, and social systems (Schlange, 2006). Aligning values, beliefs, and strategic priorities is crucial for legitimizing sustainable ventures, engaging stakeholders, and meeting sustainability goals (O'Neil & Ucbasaran, 2016; Reynolds et al., 2018). However, how entrepreneurs translate sustainability goals into action through stakeholder collaboration remains unclear and merits further study (D. Fischer et al., 2020; Savage et al., 2010).

[5] Following Christensen et al. (2021) the term CSR is employed interchangeably with sustainability initiatives and practices to represent organizational measures that evaluate, oversee, and regulate a company's obligations for and effects on society and the environment.

Other than external motivations like governments and stakeholders, internal motivations like cost reduction and market demands as seen as driver of sustainable innovation (Díaz-García et al., 2015). Market and financial factors, such as customer demands, brand reputation, and cost savings, further drive SI adoption (Clark & Charter, 2007). Jansson et al. (2011) found that consumer perceptions and market demand outweigh demographic characteristics in influencing eco-innovation adoption. Consumers' willingness to pay more for eco-friendly products encourages voluntary innovation (Khanna et al., 2009; Manget et al., 2009), although Kammerer (2009) notes that regulatory interventions may be needed when eco-products' benefits are unclear. Yalabik & Fairchild (2011) suggest that market pressures often outweigh regulatory influences in driving sustainable innovation.

The increased emphasis on corporate sustainability is driven by multiple factors including stringent regulations, stakeholder pressures, heightened environmental risks, and the need to adopt a long-term perspective in decision-making (Beske-Janssen et al., 2015). Corporate sustainability frameworks typically incorporate standards such as the Global Reporting Initiative (GRI), ISO 14000 series, ISO 26000, and more recently, sustainability reporting standards from entities such as the Sustainability Accounting Standards Board (SASB) and the International Sustainability Standards Board (ISSB) under IFRS. Each of these frameworks is designed to guide organizations toward more transparent, accountable, and effective sustainability performance management (Maas et al., 2016). These internationally recognized standards provide structured approaches that enable organizations to systematically measure, report, and improve their sustainability performance. They offer guidelines for both internal management systems and external reporting to ensure consistency, credibility, and comparability across industries.

Within these frameworks, sustainability performance is typically evaluated across three dimensions: environmental, economic, and social. The application of performance measures extends beyond environmental consideration alone. Economic measures such as operational cost savings, profit margins, market share, and return on investment are integrated to capture the financial impacts of sustainability initiatives. Although historically less quantifiable, social performance now involves indicators for employee health and safety, diversity, human rights, and community engagement. While the quantification of social performance has traditionally lagged behind environmental metrics, there is growing recognition of the need for standardized social measures. Such standardization is essential to provide a holistic view of corporate sustainability (Rezaee, 2016).

A key insight from empirical studies is that, despite the proliferation of sustainability metrics, there remains a lack of consensus on standard measures and definitions. For example, in green supply chain management, researchers have documented the existence of over 2500 unique metrics, many addressing similar issues with slight variations. This heterogeneity complicates cross-industry benchmarking and necessitates the development of integrated measurement frameworks that can consolidate and contextualize diverse metrics into coherent performance assessments (Ahi & Searcy, 2015). The challenges inherent in sustainability measurement range from a lack of standardized indicators to difficulties in capturing qualitative social impacts (Boiral & Henri, 2017). This underscores the need for ongoing research and industry collaboration (Rezaee, 2016)

In response to this complexity and lack of standardization, some companies pursue alternative forms of validation to signal their commitment. In this vein, some organizations are moving beyond metric-by-metric reporting to adopt holistic certification models, where certifying bodies position themselves as guarantors of the certified companies' genuine commitment to stakeholder governance. One prominent example is the B[6] Corp certification, which has attracted considerable academic attention (e.g., Liute & De Giacomo, 2022; Paelman et al., 2020), awarded by B Lab to companies effectively balancing social objectives and economic performance[7]. The independent non-profit B Lab, established in the United States (US) in 2006, issues B Corp certifications. B Lab pursues four main objectives: (i) building a community of Certified B Corporations; (ii) advocating legislation to establish a new corporate form adhering to higher standards of purpose, accountability, and transparency; (iii) accelerating the growth of impact investing through its rating system; (iv) and promoting the movement by sharing Certified B Corporations' success stories (Cao et al., 2017). B Corps are for-profit entities incorporating social and environmental goals into their core strategies and business models. To qualify, firms must complete a B Impact Assessment (BIA), a self-assessment evaluating impacts on workers, customers, society, and the environment. Specifically, the BIA produces a B Impact Report addressing Governance (Mission, Engagement, Ethics, Transparency), Workers (Financial Security, Health, Wellness, Safety, Career

[6] B stands for Benefit. B-Lab posits that benefit goes to all stakeholders of the company once received the certification. https://bcorporation.com.au/blog/blog-what-does-the-b-corp-logo-mean/

[7] https://www.bcorporation.net/en-us/

Development, Engagement, Satisfaction), Environment (Environmental Management, Air and Climate, Water, Land, Life), and Customers (Customer Stewardship)[8].

2.3 Financing innovation

Before delving into the core topic of innovation indicators, it is important to acknowledge the significant role that private and public funding plays in driving both innovation and, more broadly, sustainable innovation. The importance of financing innovation is recognized and highlighted in the OECD (2018). The OECD (2018) describes 11 sources of funding which are pre-existing internal funds, transfers from affiliations, customer orders which include public procurement, shareholder loans, public loans, loans from multinationals, private equity, grants or subsidies, bonds and obligations and other sources such as crowdfunding. This section will first provide a broad overview of the theory behind financing innovation, examining the roles and rationales for both public and private investment. Subsequently, it will narrow its focus to private funding mechanisms, particularly private equity and venture capital, which are central to the empirical investigation of this thesis.

The acknowledgment that research and development (R&D) could promote science and engineering began post-World War II. After World War II, governments increased investments in R&D and introduced government agencies and research programs to encourage economic welfare. Subsequently, during the Cold War, governments also started investing in R&D in the private sector (David et al., 2000). Since then, understanding how to support technological entrepreneurship has been a central question for governments and regional authorities to improve innovation at national and regional levels (Venkataraman, 2004). The diverse range of government funds, including tax incentives and subsidies, intensified corporate interest in initiating R&D projects with their own capital (David et al., 2000). This surge in investment has sparked extensive research, which led to two primary reasons for government financing of private R&D projects. First, economists justified government investment to overcome knowledge spillover (Griliches, 1991) (i.e., the fact that another firm could be able to replicate or apprehend without investment in R&D), the main cause of market failure that hinders the investment of private companies in radically new R&D project (Bloom et al., 2019). Second, government investment allows firms to cross the valley of death. The Valley of Death represents a critical bottleneck in the early stages of

[8] https://www.bcorporation.net/en-us/find-a-b-corp/company/linus/

science-based innovation (Ellwood et al., 2022). During this phase, promising technologies face significant development challenges due to insufficient funding and inadequate support systems that bridge the gap between initial scientific research and commercial application (Auerswald & Branscomb, 2003; Frank et al., 1996; Markham, 2002). With the increasing issue of climate change, government investments have become essential for sustainable innovation due to two primary factors. First, clean technologies demonstrate exceptionally strong knowledge spillovers, with empirical research showing that spillovers from clean technology patents (measured through forward citations in energy production and transportation) exceed those from conventional alternatives (Dechezleprêtre et al., 2019). Second, the investment landscape reveals additional challenges. Clean energy startups and funding experienced a notable surge around 2008 in the United States, followed by a subsequent decline (Ghosh & Nanda, 2010). This proved that without robust policy frameworks and incentives, investors have increasingly perceived early-stage clean technologies as carrying higher risks compared to established conventional alternatives, creating significant financing barriers (Gaddy et al., 2017; Nanda et al., 2015). This risk perception has contributed to the volatile funding patterns observed in the clean energy sector over the past decades.

Financing of firms by governments or private investors is a way to encourage firm innovation (Yigitcanlar et al., 2018) and sustainability (Stern & Valero, 2021). However, the relationship between private and public funding depends on a lot of variables and for this reason, research results do not have a unified vision. Some studies support the view that public funding produces additionality in R&D expenditures and innovation by private firms (Aerts & Schmidt, 2008; Czarnitzki & Lopes-Bento, 2013; González & Pazó, 2008; Mateut, 2018), while others provide evidence that public intervention crowds out private R&D investments (Busom ∗, 2000; Hall & Rosenberg, 2010; Marino et al., 2016) or that the effectiveness of public subsidies varies with the context (Heijs et al., 2022; Hud & Hussinger, 2015; Klette et al., 2000; Zúñiga-Vicente et al., 2014) and the design of the subsidy program (Bellucci et al., 2019).

As a result, raising external finance is typically more expensive and problematic for innovative enterprises compared to non-innovative ones (Aghion et al., 2004; Hall & Lerner, 2010), and especially among SMEs, for which information asymmetries and the financial gap are larger, as well as for long-term debt maturities (Alessandrini et al., 2010; J. R. Brown et al., 2012; Neville & Lucey, 2022; Wellalage & Fernandez, 2019). In addition, there are two reasons that justify the choice between public or private funding: the resource effect and the certification effect (Bellucci et al., 2023). Resource effects reflect the grant's role as an immediate injection of low-cost capital that closes a funding gap (Guo et al., 2022; Howell, 2017). Grant can replace short-term bank borrowing and lower the cost of debt (debt-substitution), crowd-in extra short-term credit when co-financing is required (debt-additionality)(Czarnitzki, 2006), or finance proof-of-concept work whose success later opens the door to new long-term loans (prototyping)(Howell, 2017). Certification effects arise from the rigorous, public evaluation process inherent in competitive subsidy programs. When firms receive these awards, they certify a certain quality that helps reduce information gaps between companies and their potential lenders or investors. Consistent with signaling theory total debt remains unchanged after the award, but firms systematically rebalance liabilities: short-term borrowing decreases as long-term debt increases. Simultaneously, companies benefit from lower average interest rates and reduced dependence on trade credit arrangements. The effects are particularly strong among young companies, high-tech firms, and other businesses where information transparency is typically limited. Importantly, these benefits occur regardless of the actual grant amount, providing compelling evidence that the certification value, rather than the direct financial injection, serves as the primary driver of improved financing conditions.

2.3.1 Private Equity

The word "private equity" encompasses a broad range of investment activities, extending beyond early-stage funding of new firms. It includes development and replacement capital, management buy-outs, and management buy-ins, among others. Venture capital (VC) is specifically positioned as a sub-sector within this broader private equity landscape, and it often receives special mention due to its significant role in total private funding investment. This prominence, particularly of VC, stems from its crucial role in addressing specific challenges faced by young, high-growth potential firms.

Gompers & Lerner (2001) provided an overview of the VC history. VC industry emerged in 1946 with American Research & Development, a closed-end fund created to commercialize WWII technologies. This closed-end fund traded shares like stocks on an exchange, allowing it to invest in illiquid assets without needing to redeem investor capital. Although the shares were liquid and exempt from SEC regulation, institutional investors considered them too risky, so most were purchased by individual investors (Liles, 1977). Additionally, brokers began selling the funds to inappropriate investors who failed to realize the promised gains, further eroding trust in the market. A decisive shift came in 1958 when Draper, Gaither & Anderson pioneered the limited-partnership model, which insulated funds from stringent securities disclosure rules while aligning incentives through finite lifetimes and profit-sharing with sophisticated institutional investors. During the 1960s, early venture organizations primarily raised funds via closed-end funds or government-backed Small Business Investment Companies (SBICs). Launched post-Sputnik to boost tech competitiveness, the SBIC program offered government funds to those with private capital. However, poor design, excessive regulations, not strict supervision allowed fraudulent operators to lead most SBICs to collapse by the late 1960s and 70s, highlighting the importance of professional fund governance. Venture activity accelerated in the late 1970s after the U.S. Department of Labor's 1979 clarification of ERISA's "prudent-man" rule unlocked pension-fund commitments (P. Gompers & Lerner, 2001a). By the mid-1980s, limited partnerships dominated fundraising and invested heavily in emerging information-technology and life-science firms, centered in California and Massachusetts. At the same time, a small contraction of the investment in the late 80s due to returns on venture capital funds declined in the mid-1980s, apparently because of overinvestment in various industries and the entry of inexperienced venture capitalists. As investors became disappointed with returns, they committed less capital to the industry. Robust IPO markets in the 1990s reignited commitments, growing twenty times between 1991 and 2000, and attracted new actors such as corporate venture arms and publicly traded hybrid incubators. By the century's end, pension funds supplied over half of all commitments, venture capitalists backed more than half of U.S. IPOs, and the model began spreading internationally, marking a transition from an experimental financing niche to a central engine of high-technology innovation.

Nowadays, venture capitalists (VCs) are a primary source of early-stage funding for young companies with high growth potential. They are particularly well-suited to help these growth-oriented firms develop and implement radical innovations due to their expertise and strategic guidance capabilities (Metrick & Yasuda, 2011; Van den Heuvel & Popp, 2023).

The need for VC investment frequently arises because young ventures often lack the resources required for comprehensive market analysis and expertise in managing later-stage value-chain operations (P. A. Gompers et al., 2020; Leiponen & Helfat, 2010; Rothaermel & Deeds, 2004).

Additionally, these companies struggle with successfully commercializing their R&D efforts (Shin et al., 2025).

This resource gap exists because product development is inherently resource-intensive. Due to the resource-intensive nature of product development, new ventures typically require considerable time to build experience and acquire the technological and market knowledge needed for new product development (Atuahene-Gima, 2004; S. L. Brown & Eisenhardt, 1995; Kiss & Barr, 2017). A typical VC fund structure involves an initial investment period of around five years, after which the fund manager has five to seven years to realize returns for investors (S. N. Kaplan & Schoar, 2005) exiting their investments, commonly through an acquisition or an initial public offering (IPO), within the lifespan of the VC fund. Researchers found that investee firms experience improvement of the organization structure and process, including the planning and development of information flow which helps increase their chances of a successful exit (Cailou & DeHai, 2022; Caselli et al., 2009a; Caselli & Negri, 2021; Hellmann & Puri, 2002; Savaneviciene et al., 2015). Thus, Venture capital-backed firms grow faster, are more innovative, and perform better than firms without venture capital support (Alemany & Marti, 2005; M. G. Colombo et al., 2016; Dushnitsky & Lenox, 2006; Guo & Jiang, 2013; Hellmann & Puri, 2002). Specifically, VCs contribute to increased professionalization (Hellmann & Puri, 2002), enhance product market and strategy formation (Peneder, 2010), and support internationalization efforts (Fernhaber et al., 2009; Peneder, 2010). In addition, there are scientific findings that enterprises backed by VCs are more innovative and obtain more patents than those without VC (Bertoni & Tykvová, 2012; Caselli et al., 2009b; Hirukawa & Ueda, 2011). For this reason, venture capital is recognized as an important component of regional innovation systems (D. F. Smith & Florida, 2013). Research so far provides evidences that venture capital financing reveals in positive effect not only by providing financial capital but increasing portfolio companies value in the following perspectives.

2.4 Innovation Indicators

Studies unveiled that innovation is a complex phenomenon with many processes and stakeholders involved under many different contexts has always been very difficult to conceptually grasp and even more challenging to objectively measure (Archibugi & Sirilli, 2000; Héroux-Vaillancourt, 2023; Makkonen & Van Der Have, 2013).

> An innovation indicator is a statistical summary measure of an innovation phenomenon (activity, output, expenditure, etc.) observed in a population or a sample thereof for a specified time or place. Indicators are usually corrected (or standardised) to permit comparisons across units that differ in size or other characteristics. For example, an aggregate indicator for national innovation expenditures as a percentage of gross domestic product (GDP) corrects for the size of different economies (OECD, 2018, p.247)

Where “unit” indicates the actor responsible for innovations (Gault et al., 2023). The history of innovation measurement begins with research and development (R&D) expenditure as the first widely recognized innovation indicator. The acknowledgment that R&D could promote scientific and engineering advancement emerged after World War II, when governments increased investments in R&D and established agencies and research programs to encourage economic welfare. This focus intensified during the Cold War period, when governments also began investing in private sector R&D (David et al., 2000). As understanding of innovation’s importance grew alongside R&D’s central role in early innovation models, the OECD introduced the Oslo Manual in 1992 to provide standardized guidelines for the research community. This manual significantly advanced the field by enabling country comparisons through the Community Innovation Survey (CIS) (OECD, 2018).

Motivated by the release of the Oslo Manual, Becheikh et al. (2006) conducted a systematic review of studies on technological innovation indicators in the manufacturing sector published between 1993 and 2003. These indicators enable managers to promote policies facilitating organizational growth and support individual and team creativity (Pirola-Merlo & Mann, 2004). They also help organizations build creative teams and achieve superior performance even in turbulent market environments (Hult et al., 2004). Additionally, innovation indicators assist policymakers in framing policies and strategies for developing innovation programs (Vladimirov & Williams, 2018).

Despite these advances, measuring complex and multidimensional phenomena like innovation has remained difficult. The challenges span multiple dimensions, from establishing precise and widely accepted definitions (Gault, 2018) and developing broad conceptual models (Björk et al., 2023) to determining individual indicators and their responses (Cirera & Muzi, 2020). These persistent challenges have stimulated numerous measurement approaches, ranging from simple indicators for patents (Griliches, 1991; Ponta et al., 2021) or citations (Coombs et al., 1996) to elaborate systems like the European Innovation Scoreboard (EIS), which covers several areas with many indicators (European Commission, 2024).

In response to the growing complexity of innovation measurement, Todorov et al. (2024) developed a valuable framework by separating innovation indicators into two main groups. The first group comprises indicators concerned with and interpreted at the micro-level, i.e., individual enterprises. These indicators reflect the specific conditions for the emergence of innovations (product, process, organizational, and marketing), as well as the innovation activities within companies that lead to innovations (Aimiuwu & Bapna, 2011; Bellavitis et al., 2017; Gama et al., 2007; OECD, 2018). What makes these micro-level indicators particularly valuable is their direct connection to the origin of innovation. They provide detailed information about the innovation itself, the people involved in the process, and the specific conditions and constraints that shape development.

Regarding data usefulness, these micro-level indicators offer the greatest cognitive value because they provide in-depth information about innovation activities, associated challenges, and measures that could accelerate innovation development and knowledge transfer (Dziallas & Blind, 2019). They are typically measured through large-scale questionnaire surveys, such as the biennial European Community Innovation Survey (CIS) and Germany's annual Mannheim Innovation Panel (MIP), both based on the Oslo Manual guidelines (OECD, 2018). These surveys collect firm-level data from innovative and non-innovative enterprises, including R&D expenditures and innovation characteristics categorized by novelty (new to firm, market, industry, or world) and type (product, process, marketing, and organizational). In contrast to micro-level indicators, Todorov's second group consists of indicators at a high level of aggregation that provide a generalized picture of innovation processes in a particular region, institutional sector, or country. These macro-level indicators are particularly suitable for planning and reporting policies and strategies and can be extensively used for statistical and econometric analysis (OECD, 2018). They sacrifice granularity for breadth and comparability across larger units. Between these two extreme positions exists a middle ground of indicators that can be evaluated at both micro and macroeconomic levels. These include indicators for patents, trademarks, and citations, some financial indicators of companies, R&D expenditures, employed persons and their qualifications, and high-tech production (Todorov et al., 2024). These intermediate indicators provide valuable information on general state, structure, and trends, allowing disaggregation to enrich analysis and showing domain-specific regularities (Erdin & Çağlar, 2023; Virkkala & Mariussen, 2021).

Dziallas & Blind (2019) conducted the most comprehensive summary of classical innovation indicators to date. Their work identified 800 different dimensions that can be summarized using Becheikh et al. (2006) classification framework. This classification organizes innovation indicators into seven distinct dimensions, each representing a different aspect of the innovation process with varying degrees of emphasis in the literature. The most prominent dimension, Innovation Culture and Organizational Structure (accounting for 20% of emphasis in the literature) includes indicators such as management's commitment to innovation, leaders trained in creativity techniques, and the allocation of management time towards innovation relative to routine tasks. This is followed by the Market dimension (13%), which covers external market-oriented indicators like market demand, customer satisfaction, competitor analysis, and overall market share. The R&D Activities and Input dimension (11%) emphasizes financial investments in research and development, human resources dedicated to innovation, and R&D intensity, while the Organizational Structure dimension (10%) highlights firm characteristics such as company size, complexity, and internal organizational arrangements that facilitate or hinder innovation. Completing the framework is the Competence and Knowledge dimension (9%), which focuses on employee competencies, knowledge management, and continuous learning practices that drive innovation processes; the Environment dimension (5%), addressing both internal and external environmental contexts influencing an organization's innovation activities; and the Network dimension (4%), emphasizing indicators related to external collaborations such as partnerships, alliances, and interactions with external stakeholders crucial for innovation performance. Together, these dimensions form a comprehensive framework capturing the multifaceted nature of innovation processes and the diverse factors influencing innovation effectiveness and outcomes within firms.

The abundance of innovation metrics resulting from these complex frameworks has created both opportunities and challenges. On one hand, it enables measurement of different aspects of innovation phenomena (Erdin & Çağlar, 2023; Taques et al., 2021), and the actors in these processes (Virkkala & Mariussen, 2021). On the other hand, limitations in data comparability, coverage, and quality restrict the efficient use of information in scientific research and government policy implementation (Brenner & Broekel, 2011; Reeb & Zhao, 2020).

As reported by Kinne & Axenbeck (2020), these limitations are well illustrated by the German MIP, which covers 10,000 firms annually but represents only 0.3% of Germany’s total firms. Consequently, the total number of innovative firms remains unknown and can only be estimated through statistical extrapolation. This sampling limitation means that rare but potentially important innovation activities in unobserved sectors or technological fields may not be captured in the data. This gap also affects the analysis of geospatial innovation processes operating on micro geographical scale (Arzaghi & Henderson, 2008; Carlino & Kerr, 2015; Catalini, 2012; Jang et al., 2017; Kerr & Kominers, 2015). As a result of these sampling issues, established innovation indicators from questionnaire-based surveys lack sectoral, technological, and geographical granularity. Additionally, these surveys are costly and time-intensive, with significant delays between data collection and results. Furthermore, they require firm participation, and voluntary surveys like the MIP suffer from incomplete responses and inaccessible information (Kleinknecht et al., 2002).

Beyond survey limitations, micro-level indicators face their own challenges. Their highly specific nature makes it difficult to establish general patterns and regularities. They can produce inaccurate estimates and, due to lack of standardization, limit comparative analysis, planning, and targeted innovation policies (Ponta et al., 2021; Taques et al., 2021). Additionally, firm-level indicators cannot capture innovation effects that extend beyond the organization to broader geographic, economic, or institutional spheres (Hoelscher, 2016).

These persistent limitations in traditional innovation measurement have created opportunities for new methodologies, particularly those leveraging big data. Realizing this potential requires advanced statistical techniques for data collection and enhanced information technologies for processing unstructured data. These developments could improve firm-level innovation measurement and generate new indicators of knowledge sharing and diffusion. Big data approaches differ fundamentally from traditional data collection methods, as they rely on data generated for other purposes and can produce results quickly. However, this data is typically unstructured, and its processing requires advanced methods and computational power that weren’t previously available. Recent developments in analytical methods and web-based techniques have facilitated big data’s use as a source of information on firms’ innovation output, creating new possibilities for innovation measurement.

Innovations are undeniably a major factor influencing modern economic development across nations worldwide. Promoting innovation-based economic growth and creating adequate conditions for developing new products, processes, and organizational forms has become a key policy objective for many governments seeking to build advanced, sustainable knowledge economies. For example, this approach has been explicitly established as a central goal in European policy (European Commission: Directorate-General for Communication, 2019), with the European Union working toward creating an advanced and sustainable knowledge-based economy (European Commission: Directorate-General for Communication, 2023). Similarly, Canada's Innovation and Skills Plan commits the country to becoming "a world-leading centre for innovation" and to "building a knowledge-based economy in all regions of Canada" (Government of Canada, 2017; Innovation, Science and Economic Development Canada, 2019).Achieving such an ambitious goal requires effective implementation of government-introduced policies, connected with accurate and timely measurement, assessment, and monitoring of innovation processes, results, and consequences for individuals, enterprises, society, and the environment (El Bassiti & Ajhoun, 2016; Taques et al., 2021).

As we have seen throughout this analysis, traditional methods face significant limitations in capturing the full complexity of innovation ecosystems. Therefore, emerging approaches like big data analysis offer promising complementary tools for more comprehensive innovation measurement. By combining traditional survey-based approaches with newer big data methodologies, researchers and policymakers can develop a more complete picture of innovation activities and their impacts, enabling more effective support for innovation-driven growth and development. The future of innovation measurement likely lies in this hybrid approach, leveraging the strengths of multiple measurement methodologies while mitigating their individual weaknesses.

2.5 New innovation indicators

Although surveys were considered the first resource when it comes to study innovation on a firm level, they come with several shortcomings that have been highlighted in the literature. According to Rammer & Es-Sadki, (2023) there are mainly 6 drawback of using surveys. First, most surveys are not comparable due to the differences in questionnaire design, sampling, and survey methods across countries (Archibugi & Planta, 1996; Kleinknecht et al., 2002). Second, surveys are often targeting specific sectors or region which limit the granularity (Archibugi & Planta, 1996; Cirera & Muzi, 2020; Tether, 2002). Third, surveys are biased by the respondents' interpretation of the definition included in the survey (Arundel et al., 2013; Cirera & Muzi, 2020). Fourth, response rates for innovation surveys have been steadily decreasing. Fifth, conducting innovation surveys usually takes a significant amount of time (Kinne & Lenz, 2021a), leading to time lag between the collection of the data and the actual publication. Finally, surveys target specific respondents rather than examining innovation processes comprehensively, which limits research into emerging technologies or market trends.

Given these significant limitations of survey-based approaches, researchers have increasingly turned to alternative methods. This proliferation of innovation measurement has enabled the assessment of diverse aspects of the innovation phenomena (Erdin & Çağlar, 2023; Taques et al., 2021), and the various actors involved (Virkkala & Mariussen, 2021). However, while methodological diversity offers new possibilities, it has also created new challenges. The absence of high-quality data can lead to biased results and inaccurate conclusions in innovation studies, complicating and even hindering research in the field. Specifically, a lack of comparable data prevents longitudinal or international comparisons, while differing concepts and methodologies impede the combination of various data sources. These data quality issues also affect policymakers and businesses. For policymakers, poor data quality impacts their ability to be proactive or react adequately to innovation developments. This can result in incorrect decisions, contradictory policies, and ultimately hamper innovation management at national or transnational levels. Policy interventions may be misdirected or fail to address crucial areas needing support. Additionally, objectively assessing policy effectiveness becomes difficult when data coverage is limited or unrepresentative. For businesses, diverse and unstructured data can negatively impact decision-making, making investment choices costlier, impeding proper knowledge diffusion, and increasing the risks of missed opportunities. Collectively, these issues underscore the critical importance of high-quality data concerning innovations and their impact on economic and social development.

Big data has the potential to overcome some of these shortcomings of innovation surveys and may offer a more complete picture of innovation in firms (Kinne & Axenbeck, 2018, 2020b). In this thesis, we refer to "big data" as digitized data sources, often unstructured, that originate from various sources containing information about different aspects of firms' innovation activities and can cover a wide range of firms, potentially encompassing the entire business sector (Agarwal &

Dhar, 2014; Rammer & Es-Sadki, 2023). For instance, typical sources include online communities (Füller et al., 2008), patent databases (S. Lee et al., 2009), crowdsourcing contests (Terwiesch & Xu, 2008), crowdfunding platforms (Kleinert et al., 2022), or social media (Testa et al., 2020). Key characteristics that distinguish big data from traditional data are variety, velocity, veracity, and volume (Gök et al., 2015; Rammer & Es-Sadki, 2023). The variety is given by the different types of data (unstructured or semi-structured, text, numbers and/or images, pre-coded or uncoded). Velocity refers to the speed with which data is generated, such as the high frequency of social media or online communities. Veracity refers to the accuracy and reliability of the data. High veracity data contains accurate and trustworthy records that contribute meaningfully to research results. Data volume refers to the amount of information available (Ghasemaghaei & Calic, 2020).

For the purpose of this thesis, we will focus on the innovation indicators created using websites. Two research areas legitimate the use of website to create innovation indicators: webometrics and signaling theory.

2.6 How to study the web: Webometrics and Web mining

The following text describes the history of two intertwined disciplines that created the tools to analyze the web. In the 1990s the advent of the World Wide Web remarkably impacted people’s life. The advantages of the Internet motivated both companies and people to embrace this new technology. Companies began using the Internet for diverse purposes: reporting achievements, presenting innovative products, selling directly to consumers, expanding increasing their customer base, and in particular for exporting their products (Reuber & Fischer, 2011). Company websites also became repositories of valuable information, including location data, strategic plans, and relationships with other organizations. Meanwhile, as users navigate online, they inadvertently provide information about themselves, making the Internet the greatest source of data about people’s behaviors, choices, and habits. This user-generated information has become central to companies’ critical strategic decision-making.

As users increasingly relied on search engines and browsers, the growing need for precise and reliable information led to the rise of Web data mining. Defined as the exploration and analysis of purposeful information from World Wide Web data (Madria et al., 1999), Web mining (also known as “web data mining”) is more broadly described as “the discovery and analysis of useful information from the World Wide Web” (Cooley et al., 1997, p. 558). Social scientists were among the first to recognize the potential of the internet as a data source for research, adapting "informetric" methods for content analysis of the World Wide Web. This led to the introduction of "webometrics." Webometrics was first mentioned by (Almind & Ingwersen, 1997, p. 404): "the approach taken here will be called webometrics, which covers research of all network-based communication using informetric or other quantitative measures". They demonstrated a method for webometric analysis through a case study comparing Denmark's web usage with other Nordic countries. They found that applying informetric methods to the web was useful for a range of tasks, including issue management, gathering business intelligence, and research evaluation. Later on, Björneborn & Ingwersen (2004) identified the webometrics resemblances to informetric and scientometrics methods, comparing web-links with citations. They argue that the webometric analyses of the nature, structures and content-properties of the websites and web-pages are important for understanding the web and its connections. They also acknowledged four main areas for webometrics which are (1) web page content analysis, (2) web link structure analysis, (3) web usage analysis and (4) web technology analysis (Batista-Navarro et al., 2013; Björneborn & Ingwersen, 2004; Hassan & Haddawy, 2015; Jahangir et al., 2017; Shardlow et al., 2018; Thompson et al., 2013, 2017; Waheed et al., 2020). Finally, Thelwall, a prominent figure who had earlier developed a web link mining crawler, later redefined webometrics as "the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study." (Thelwall, 2009, p. 6). This definition attempted to distinguish webometrics from purely mathematical analyses and emphasize its cross-disciplinary nature.

On the other hand, Etzioni (1996) coined the term "Web mining" as the use of data mining techniques to automatically discover web documents and services, extract information, and uncover general patterns (H. Chen & Chau, 2004; Etzioni, 1996). Web mining is the area of data mining dealing with the extraction of interesting knowledge from the web (Etzioni 1996). It is commonly classified into content, structure, and usage mining (Facca & Lanzi, 2005; K. Sharma & Bhatnagar, 2011). Subtasks across these areas include resource finding, information selection and preprocessing, and analysis or generalization (Kosala & Blockeel, 2000; Q. Zhang & Segall, 2008). Srivastava et al. (2000) described usage mining in three phases: preprocessing, pattern discovery and pattern analysis. They identified four main types of data that can be mined on the web: content, structure, usage and user profiles and privacy as a key issue due to the web's nature and users' anonymity desires (Srivastava et al., 2000). According to Miner et al. (2012), web structure mining facilitates link analysis, while web usage mining and web content mining are employed to evaluate web impact and conduct blog searches. Some examples of early work in web mining are: Nasraoui et al. (2006) studied the task of tracking emerging topics and clusters in noisy and evolving text data sets and in mining evolving user profiles from clickstream data in a single pass and under different trend sequencing scenarios, where a trend sequencing scenario corresponds to a specific way of ordering web sessions or text documents. (Ou et al., 2008) introduced a flexible model of mining with dynamic thresholds for analyzing paths. They applied the Markov chain model and used dynamic thresholds to reduce the number of unnecessary rules produced. Kundu (2012) proposed hybrid approach for web traffic analysis based on pattern discovery and pattern analysis, where the most recently accessed data was prioritized and used alongside clustering. Finally, Arbelaitz et al. (2013) developed a method for generating semantically enriched usage profiles for link prediction, web design, and marketing purposes.

In the literature review of Lorentzen (2014), the author investigates the difference between webometrics and web mining. Although at the beginning the differences were more evident with their origin and goals, through the years the evolution of methodologies used made the differences more nuanced, and more researchers are using web mining as set of tools used in webometrics. Indeed, the influence from the computer science methods and the evolution of the field of NLP have positively affected the social science research. This can be contrasted with the definition of web mining as the application of data mining techniques on web structure or usage data (Kumar & Gosul, 2011). According to Lorentzen (2014), programming skills are needed to take advantage of the web's vast data resources, webometricians are likely to need to collaborate with web miners if they do not already possess the skills needed. Hence, web mining will, in all likelihood, play an important part in future webometrics.

2.7 Signalling Theory

Signaling theory was initially developed in the context of evolutionary biology with the aim to study successful communication within and between species (Carpenter, 1969; Grafen, 1990; Héroux-Vaillancourt, 2023). Communication is a fundamental aspect of all biological systems, ranging from molecular signals within cells to exchanges between individuals of the same or different species (Schaefer & Ruxton, 2011). Effective communication requires a signaller, a receiver and a signal. Signals are actions or characteristics of signallers that stimulate the sensory systems of receivers and potentially change their behavior (Schaefer & Ruxton, 2011; Sun et al., 2020). For example, the peacock's elaborate tail imposes a substantial survival cost yet reliably indicates male genetic quality, as only high-quality males can maintain this costly characteristic (Zahavi, 1975). Although it may seem advantageous for a weaker individual to mimic the signals corresponding to the characteristics of the fittest individuals (Krebs & Dawkins, 1984), there are two major theories that explain why signalling is usually honest, (i.e., the signal corresponds to the true quality of the signaller): (1) costly signalling theory and (2) the index hypothesis (Sun et al., 2020). Zahavi (1975, 1977) argued that for a signal to be honest, it must be costly or "handicapping" to the signaler. Costly signals achieve honesty by imposing a differential cost: the handicap ensures that the signal is unaffordable by the weak. Grafen (1990) later proved this theory by publishing a rigorous game-theoretic model demonstrating that costly signals can indeed form an evolutionarily stable communication system. For the index hypothesis, M. J. Smith & Harper (1995) showed that honest signals are not mimicable only when the causal link between the sender's quality and signal intensity is created by intrinsic characteristics of the sender related to physiological or development traits. In addition to these two theories, there is another class of signals, called conventional signals, which are not recognized for intrinsic characteristics of the signal itself but are kept honest through the outside intervention of laws and social norms. Conventional signals are low-cost signals often linked to costs or consequences that already exist. These signals are intended to be more detectable and easily interpreted (Guilford & Dawkins, 1995).

Scholars characterize signals of quality based on their detectability by receivers, often classifying them as "strong" or "weak" (Gulati & Higgins, 2003), similar to how movie studios signal film quality (Lampel & Shamsie, 2000). Two core dimensions work together to determine a signal's effectiveness: strength and observability. First, observability reflects the quality of the signal without considering the distortion or deception (Connelly et al., 2011; Ramaswami et al., 2010). It is also described using terms like signal clarity (Warner et al., 2006), intensity (H. Gao et al., 2008), and quality (Kao & Wu, 1994). Second, signal strength (also called signal fit) measures the correlation between the signal and the signaler's unobservable quality, representing how well the public signal align with private information about actual quality. This statistical relationship between public and private information differs from honesty (a characteristic of the signaler). When signals fail to align with the underlying quality, signaling becomes poor or ineffective (Busenitz et al., 2005; Y. Zhang & Wiersema, 2009). Enhancing signaling effectiveness can be achieved by increasing signal observability or frequency (Janney & Folta, 2003). In dynamic environments with constantly changing information (Davila et al., 2003), repetitive signaling is vital for continuously reducing information asymmetry (Janney & Folta, 2003, 2006; Park & Mezias, 2005). Repeating signals, especially using diverse methods to convey the same message (Balboa & Martí, 2007), improves effectiveness. Likewise, signal consistency, i.e. the agreement among multiple signals from one source (H. Gao et al., 2008), is important for mitigating receiver confusion from conflicting signals (Chung & Kalnins, 2001; E. Fischer & Reuber, 2007).

2.7.1 Signalling theory in economy and management

One of the first applications of signaling theory addressed situations of information asymmetry which could lead to adverse selection (Akerlof, 1970) or moral hazard (Arrow, 1962). Spence (1973) applied signalling theory to labor economics, providing a framework for understanding interactions characterized by information asymmetry. Spence examined the labor market, specifically situations where employers cannot assess potential employees' productivity levels, creating asymmetric information about worker quality. In this context, education level serves as a signal of quality, though obtaining higher education is costly for workers with low productivity. This dynamic illustrates the core principle of signalling theory, which in organizational and economic contexts provides several key concepts are central to understanding communication processes under information asymmetry. The signaler is the party possessing private information about an unobservable quality or intention. In an organizational context, this typically includes executives or managers (Spence, 1973). The signal is an observable attribute or action transmitted by the signaler that conveys information about their hidden qualities or intentions. For example, signals might originate from organizational insiders (internal stakeholders) who possess relevant information about the organization's activities or future plans. The receiver is the party who lacks the private information held by the signaler and must interpret the signal to make decisions. Receivers are typically external to the immediate source of the private information, such as individuals outside the organization (Basuroy et al., 2006; Lester et al., 2006; Rao et al., 1999; Y. Zhang & Wiersema, 2009).

Referring to the characteristics of an honest signal in biology, three conditions are typically required for an effective signal (Connelly et al., 2011). First, receivers must know what information is being sought (the signal). Second, the signal must be clear and easily observable. Third, using the signal must be costly for other parties, especially for those of low quality. These conditions ensure that the signal serves as a pertinent piece of information, allowing receivers operating under asymmetric information to distinguish "high-quality parties" from others before making decisions (Lanchimba et al., 2021). Signalling theory provides a valuable framework for understanding how high-quality entrepreneurs navigate environments characterized by asymmetrically distributed information, particularly when seeking funding (Bafera & Kleinert, 2023; O. Colombo, 2021). By employing such signals, entrepreneurs enable resource holders to infer the high quality of the related attribute, as low-quality counterparts would be unable or unwilling to incur the necessary cost. Examples of signals commonly studied in this domain include previous funds raised (Vanacker et al., 2020) or the development of a prototype (Steigenberger & Wilhelm, 2018).

Although signaling theory is a powerful tool for predicting which entrepreneurs secure funding, its traditional assumption are frequently challenged in entrepreneurial settings (Bafera & Kleinert, 2023). Classical signaling theory assumes that entrepreneurs are fully aware of their attributes' quality while resource holders do not. However, the Knightian uncertainty often faced by entrepreneurs challenges this assumption, as it means entrepreneurs themselves may not fully acknowledge the quality of their own characteristics (Packard et al., 2017). Moreover, the entrepreneurs' signals may only be imprecisely linked to their actual potential for success. Furthermore, resource holders, frequently overloaded with information, may not perceive every signal or interpret them in a rational way, as psychological processes and biases can significantly influence their evaluation (Steigenberger & Wilhelm, 2018). Although signalling theory has been a powerful tool for understanding entrepreneur-investor interactions (O. Colombo, 2021) and fundraising outcomes (Clough et al., 2019), entrepreneurial contexts often involve communications that do not fit classical signaling assumptions. Traditional signaling theory requires signals to be costly to send (Spence, 1973), but many entrepreneurial communications lack this characteristic. In response, scholars have introduced terms such as "costless" signals (Anglin et al., 2018), "rhetorical" signals (Steigenberger & Wilhelm, 2018), and "informational" signals (Wesley II et al., 2022) to better capture these communication dynamics. These concepts describe forms of communication that, although not costly in the classic sense, can still reduce resource-holder uncertainty and attract funding. Nevertheless, the gap between theory and practice creates ambiguity, making it difficult to definitively determine when a phenomenon is appropriately analyzed from a signaling perspective. Finally, Steigenberger et al. (2024) suggested an empirical way to mitigate this ambiguity by focusing on a minimum characteristic of the signaling process: signals reduce resource holders' level of uncertainty.

The literature on entrepreneurship found several signals that have an impact on the capacity of the firms to signal their quality to venture capitalist, customers, stakeholders, IPO investors (E. Fischer & Reuber, 2007; Hsu, 2007; Reuer et al., 2012; Wehnert et al., 2019). I group these research in five primary categories of signals employed by entrepreneurs: human capital signals, intellectual property signals, strategic alliance formation and networks, financial and governance signals, and other contextual factors.

Human capital signals consistently emerge as the most influential, with prior founding experience, successful venture backgrounds, and elite educational credentials serving as particularly powerful indicators in early-stage funding decisions (Certo et al., 2001; Ko & McKelvie, 2018). These signals are critical for reducing information asymmetries and mitigating the risk of adverse selection (Bellavitis et al., 2019). The overall prestige of the top management team, for example, serves as a powerful heuristic for quality, often leading to higher firm valuations from IPO investors (Certo et al., 2001).

Within this category, an entrepreneur's direct experience is one of the most persuasive signals. Serial founders, especially those with demonstrably successful exits, send the strongest economic signals, influencing not only which investors they attract but also the valuation they receive (Hsu, 2007). Similarly, the experience of a founder or CEO serves as a direct pricing signal for investment bankers (Daily et al., 2005). The impact of this prior founding experience is particularly pronounced in a venture's initial financing round, where it has a major effect on the amount of external funding secured (Ko & McKelvie, 2018).

Alongside experience, the educational credentials of founders provide a distinct and durable signal. While the influence of founding experience is strongest in the earliest stages, the signaling effect of a founder's education remains significant even in later funding rounds (Ko & McKelvie, 2018). The value of elite educational credentials, and the networks they provide, is often amplified in contexts of high technological or market uncertainty Hsu (2007). This effect is also visible in specific fundraising environments, such as equity crowdfunding, where an entrepreneur's business education has been shown to significantly contribute to their success (Piva & Rossi-Lamastra, 2018).

Moving beyond human capital, intellectual property signals, particularly patents, present a more nuanced signaling mechanism. While patents serve as quality indicators in high-technology sectors and regions with robust IP protection (S. Chen et al., 2018; Solé Udina et al., 2022), their effectiveness varies significantly across venture stages and contexts. Patent portfolio size, quality, and complexity positively influence venture capital funding decisions (Caviggioli et al., 2020; L. Zhang et al., 2019), though their signaling value diminishes as firms mature and additional information becomes available (Hoenig & Henkel, 2015). Notably, research reveals the dual nature of patent signals, particularly for radical innovations, which simultaneously communicate high potential and elevated risk (M. G. Colombo et al., 2023).

Building upon these individual signals, strategic alliances represent another crucial signaling mechanism. Stuart et al. (1999) found that privately held biotech firms with prominent strategic alliance partners and organizational equity investors go to IPO faster and earn greater valuations at IPO. Gulati & Higgins (2003) posited that the impact of strategic alliances with major pharmaceutical firms on a young biotechnology firm's IPO success is contingent on market conditions suggesting that the signaling value of strategic alliances is context-dependent and more impactful when general market enthusiasm is high. Baum & Silverman (2004) find that alliance capital, which includes partnerships and strategic alliances, plays a significant role as a selection criterion for VCs when deciding to finance startups given that VCs are more likely to invest in startups that have established alliances. Ozmel et al. (2013) found a mixed impact of strategic alliances on a startup's future financing trajectory and exit. While engaging in a strategic alliance makes the likelihood of securing future alliances higher, it makes the likelihood of securing future VC activity lower. Both increase the probability of a firm going public (an IPO), but only increased VC activity is associated with a higher likelihood of exit through acquisition. Hoehn-Weiss & Karim, (2014) find that alliance portfolios increase a firm's likelihood of an IPO or acquisition. Additionally, strategic alliances are important for young firms seeking liquidity events, and the specific mix and nature of these alliances play a crucial role in signaling viability and influencing the type of exit. Doblinger et al. (2019) highlighted the significant role of alliances with government entities for cleantech startups, providing access to unique resources that can accelerate innovation in this sector.

Beyond strategic alliances, a firm’s financial and governance history provides a rich set of signals that act as additional validation mechanisms for investors. A straightforward signal is the performance of past funds. For private equity firms, stronger performance in a first fund, whether measured by realized cash exits or unrealized portfolio valuations, increases the likelihood of successfully raising a follow-on fund (Vanacker et al., 2020).

The specific history of a firm’s private placements also sends nuanced signals. The timing of these placements has an impact: a longer period since the last funding round can signal higher returns on the current placement, as it suggests the new investment is resolving a greater degree of accumulated information asymmetry (Janney & Folta, 2003). Conversely, the number of prior placements can have an inverse effect. Firms with more previous placements may see lower returns on a new placement, as their established reputation diminishes the need for the current signal to be exceptionally strong (Janney & Folta, 2003).

The reputation of previous investors acts as a powerful endorsement signal. Funding from an experienced investor enhances the perceived value of a firm’s subsequent financing events (Janney & Folta, 2006), with the prominence of prior investors having a significant influence on the amount of external funding a new venture receives (Ko & McKelvie, 2018). The backing of reputable venture capitalists can even act as a substitute for a startup’s positive cash flows, reassuring lenders and unlocking further debt financing (De Rassenfosse & Fischer, 2016)

Finally, third-party validation in the form of public grants offers another distinct signal. Grants from entities like university technology-transfer offices can effectively signal a venture’s quality and attract subsequent private venture capital (Gubitta et al., 2016). While securing an initial grant has been shown to increase the rate at which ventures acquire private investment, it is a signal of potential rather than a direct predictor of future revenue (Stevenson et al., 2021).

Finally, other contextual factors influence the effectiveness of these signals. Successful crowdfunding campaigns signal product-market validation and consumer acceptance to venture capitalists (Roma et al., 2021; Thies et al., 2018), while social media presence and engagement serve as quality indicators for digital ventures (Lo Mele et al., 2024; Nigam et al., 2020). Industry characteristics, development stages, and geographic markets all influence which signals prove most valuable (Solé Udina et al., 2022; Topaler & Adar, 2025; Zhou et al., 2023). Intermediary institutions, including venture capital firms, crowdfunding platforms, and university technology transfer offices, play crucial validation roles by amplifying or authenticating entrepreneurial signals (Kleinert et al., 2022; Ragozzino & Blevins, 2023).

Given that we want to develop indicators using corporate websites for companies created from 2020-2024, we focus our study on signals related to founder background experience, strategic alliances (partnerships), grants or private funding received, and consider the moderating effect of contextual factors. These signals are chosen because they are commonly disclosed on corporate websites and are most aligned with our research objectives. Furthermore, patent-related signals are excluded due to the characteristics of our sample, which consists of young companies that typically have limited patent portfolios during their early stages of development.

2.7.2 Signalling theory in sustainability

Another interesting application of the signaling theory framework is related to the sustainability of the companies. The impact of signaling theory on sustainable innovation is multifaceted and profound, serving as a theoretical framework that explains how firms communicate unobservable qualities such as true environmental commitment and innovation capability to diverse stakeholders, thereby reducing information asymmetry and enhancing legitimacy in the marketplace (Brito-Ramos et al., 2024).

The essence of such signaling is that observable actions (sustainability reports, third-party certifications, or high-quality ESG ratings) act as proxies for underlying sustainable innovation efforts, reducing the information asymmetry that exists between firms and their stakeholders (Jolink & Niesten, 2021). These signals allow stakeholders to differentiate between genuine sustainability leaders and firms engaging in superficial "greenwashing" (Zerbini, 2017). Thus, greenwashing represents the misalignment between signal honesty and signal fit in sustainability (Moratis, 2018). In general, organizations can implement sustainability governance mechanisms strategically in two ways (J. Gupta & Das, 2022): the first strategy is to implement the governance mechanism rigorously by consuming significant resources to generate positive Corporate Social Responsibility (CSR) outcomes. The second strategy is to symbolically implement the governance system or engage in "greenwashing" to improve the corporate image (E.-H. Kim & Lyon, 2015). Consequently, greenwashing may lead to diminished financial payoffs due to the legitimacy gap

(Sethi, 1975). The legitimacy gap emerges when there is a contradiction between an organization's activities and the expectations of society, leading to long-term inferior financial performance for such organizations (Majumder et al., 2017). Research has also demonstrated that greenwashing has a spillover effect, meaning that greenwashing practices of one brand negatively affect consumers' intentions to purchase other brands in the same industry (H. Wang et al., 2020). Consumers make an overall judgement about other firms, even if they produce genuine natural products (Kahraman & Kazançoğlu, 2019). Extensive research indicates that greenwashing significantly influences consumer purchase intentions and behavior (Ahmad & Zhang, 2020; Akturan, 2018; Bulut et al., 2021; Y.-S. Chen et al., 2020; Guerreiro & Pacheco, 2021; Guyader et al., 2017; Hameed et al., 2021; Jog & Singhal, 2024; Urbański & Ul Haque, 2020; H. Wang et al., 2020; L. Zhang et al., 2018). Multiple studies have recognized the negative consequences of these practices, particularly their impact on consumers (Y.-S. Chen & Chang, 2013; Nyilasy et al., 2014).

Despite these benefits, the literature also cautions that costly and effective signaling is not without its challenges. Ambiguities in the signal content and the risk of signal dilution through non-credible or misleading claims may all undermine the intended impact of a firm's sustainability messaging (Moratis, 2018) On the one hand, certification under standards such as ISO 26000, although flexible and guidance-oriented, functions as a signal of intent rather than a measurable performance indicator, highlighting the need for more observable and costly signals that better align with actual performance outcomes (Moratis, 2018). On the other hand, environmental patents require firms to invest significant time, effort, and money to conduct the underlying research, and the impacts on the firm's environmental performance take considerable time to materialize (Berrone et al., 2013). Patent information is publicly available and often appears in companies' annual reports. Furthermore, patents are granted only after a careful examination to ensure the novelty, and usefulness of the patented product or process. Thus, innovators bear considerable costs to conduct the underlying research, and these costs are nonrecoverable for firms that do not intend to implement the patented innovations. Such environmental actions are strong signals. In contrast, firms can participate in environmental programs sponsored by government agencies (e.g., Darnall & Sides, 2008), for a recent meta- analysis on the issue), with negligible costs (Delmas & Keller, 2005). No fines punish firms that fail to achieve environmental improvements, and participants can publicize their membership, regardless of their environmental record, thereby reducing stakeholder confidence in the overall sustainability reporting environment (Moratis, 2018). In contrast to noncertifiable standards (e.g., ISO 26000), certifiable systems and mandatory disclosures offer a more robust signaling mechanism because they reduce the possibility of misrepresentation by requiring firms to undergo periodic audits and provide transparent, comparable data about their sustainability initiatives (Hummel & Schlick, 2016). This increase in data quality and accountability directly translates into higher stakeholder trust, which is essential for fostering long-term investment in green innovation (Eccles et al., 2014).

CSR initiatives demonstrate the central role signaling theory plays in sustainable innovation. CSR disclosures (CSRD), whether in the form of comprehensive sustainability reports or through participation in ethical indices such as FTSE4Good, are designed to send reliable signals regarding a firm's ethical and sustainable practices (Zerbini, 2017) as well as third party certification such as B-Corporation (see section 2.2). Such signals not only mitigate concerns about greenwashing but also serve to build a strong, trust-based relationship with stakeholders that is essential for long-term competitive advantage (Zerbini, 2017). Empirical studies have repeatedly shown that firms with high-quality CSRD are often rewarded with better market valuations because investors interpret such signals as indicative of superior future performance and reduced risk (Ching & Gerab, 2017). Research indicates that consumers assess organizations in terms of CSR, with negative CSR linkages being more influential and having a greater negative impact than positive ones. Conversely, positive linkages improve the organizations' reputation and product assessments (Sen & Bhattacharya, 2001). Organizations can improve their performance levels through CSRD by implementing socially responsive investments that add economic value to their products by satisfying customer expectations (Skudiene & Auruskeviciene, 2012) and resulting in a positive and significant relationship between CSRD and financial performance of organizations, (Allouche & Laroche, 2005; Gallardo-Vázquez et al., 2019; J. Gupta & Das, 2022). Consequently, when an organization engages in CSR activities, it may expect to be seen as a good corporate citizen by society and stakeholders (J. Zhang et al., 2020). Thus, organizations can consider CSRD as a value-enhancing tool.

Furthermore, the importance of signaling theory becomes even more pronounced in emerging economies where institutional voids and information asymmetries are more severe. In these contexts, CSR activities and other sustainable initiatives become particularly valuable signals for overcoming the limitations of underdeveloped capital markets (Montiel et al., 2012; Su et al., 2016). Firms operating in such environments that commit to visible and costly CSR or green innovation projects often benefit from enhanced reputational advantages and improved access to external capital, which further incentivizes ongoing sustainable practices. These signals may act as insurance-like assets, mitigating information asymmetries and providing competitive differentiation in environments where traditional information channels fail to function effectively (Su et al., 2016).

Looking at the broader strategic impacts, the effective use of signaling not only enhances stakeholder trust and reduces information asymmetry but also reinforces the firm's position in competitive markets by creating a virtuous cycle: credible signals attract supportive investors, which in turn supply the resources necessary for further green innovation, thereby enhancing the firm's reputation and market performance (Brito-Ramos et al., 2024). This self-reinforcing mechanism underscores how signaling theory is central to both promoting sustainable initiatives and facilitating the diffusion of green innovations across industries (Flammer, 2021).

In conclusion, signaling theory profoundly impacts the way sustainable initiatives and green innovation are communicated and perceived by various stakeholders. Its core principle of credible, costly, and observable signals reduce information asymmetry providing, a robust framework for understanding why firms invest in high-quality sustainability disclosures, strive for recognized certifications, and design strategic alliances focused on green innovation (Brito-Ramos et al., 2024; Moratis, 2018). These signals not only enhance investor confidence and attract supportive external resources but also promote greater competitive differentiation in increasingly sustainability-driven markets (Zerbini, 2017). The authenticity of environmental commitments is particularly difficult to fabricate, especially for companies operating in environmentally sensitive sectors, where misrepresentation carries significant reputational risks. Environmental patents are typically costly and difficult to fake. Therefore, if firms have a deteriorated environmental footprint, it may be safer for them to remain silent rather than engage in deceptive green positioning (Berrone et al., 2017). This dynamic suggests that the prominence and visibility of environmental disclosures on corporate websites may serve as particularly strong signals compared to detailed sustainability reports. Website prominence represents a more visible and risky commitment that demands greater authenticity, as companies must maintain consistency between their public commitments and actual performance to avoid reputational damage that could outweigh any short-term benefits of greenwashing.

2.8 New innovation indicators using websites

The use of website to build innovation indicators has evolved through distinct phases, each marked by different methodologies and capabilities. Thus, we divide the literature review of the papers that used website to build innovation indicators for companies in three different subparagraphs according to the methodology used.

2.8.1 Early-Heuristics phase

The Early-Heuristics phase utilized rule-based approaches, relying on manually curated keyword lists, TF-IDF weighting, and hand-crafted Boolean rules to identify innovation-related phrases in text. The typical workflow involved researchers defining dictionaries, counting term occurrences or ratios, and using these counts in basic statistical tests. This phase was significant because it was cheap to implement and fully transparent, enabling large-scale text mining before sophisticated Natural Language Processing (NLP) techniques were widely available. However, its main drawbacks included missing synonyms and context, being language-specific. Below, we present the papers that match the methodologies described above.

Early work focused on using website content to map the commercialization strategies of firms in emerging high-tech sectors. The pilot application by Youtie et al. (2012) was a pioneering effort, notable for being one of the first to use the Internet Archive’s Wayback Machine for longitudinal analysis of firm websites. Applying keyword coding and hierarchical cluster analysis to a small sample of 30 nanotechnology SMEs, they discovered that commercialization pathways were complex and often nonlinear, challenging simplistic linear models of innovation. While foundational, the study’s key limitations were its small sample size and a basic keyword-matching approach that could not capture deeper semantic meaning. Building on this, Arora et al. (2013) employed a similar methodology to explore the entry strategies of 20 graphene firms, adding a cross-country comparative dimension. Their work introduced firm-specific metrics derived from keyword frequencies, such as “graphene-ness,” to quantify strategic focus. However, their approach shared the limitation of a small sample and introduced potential data integrity issues by using Google Translate for Chinese websites. Together, these foundational studies established websites as rich data sources but also highlighted the analytical challenges posed by small-scale studies and simplistic text-mining techniques.

A crucial development in this phase was the effort to validate these new web-based indicators against traditional measures. Gök et al. (2015) provided powerful evidence of their value, showing that simple keyword searches for R&D on SME websites identified a vastly greater number of innovative firms (nearly 70%) than traditional metrics such as patents or R&D surveys (5-20%). This demonstrated that websites capture downstream, commercially-oriented innovation. However, the authors noted their method's sensitivity to specific keyword choices and its inability to verify the quality or extent of the claimed R&D. Addressing these validation concerns, Héroux-Vaillancourt et al. (2020) performed a more rigorous validation by comparing web indicators to detailed survey data using standard psychometric tools, namely Multitrait-Multimethod (MTMM) matrices and Confirmatory Factor Analysis (CFA). They concluded that while web indicators are poor substitutes for specific survey questions, they function as effective proxies for a firm's strategic emphasis on innovation. As a critical counterpoint, Kinne & Axenbeck (2020) moved beyond validation to map systemic data gaps. Their large-scale study in Germany used a custom scraper (ARGUS) to retrieve companies' websites. Then, using a keyword-based approach and hyperlink analysis on scraped website data, they show that website data varies significantly, as URL availability is significantly lower for very small, very young, or rural firms. This finding on inherent sampling bias is a fundamental challenge for the field, questioning the representativeness of any study relying solely on firms with an established web presence.

The methodology also evolved to capture more dynamic aspects of firm behavior. A straightforward approach by Blazquez et al. (2018) demonstrated this dynamism's predictive power. Using longitudinal data and `survival analysis` models, they found that the mere act of making major website updates was strongly correlated with firm survival, acting as a simple but effective real-time indicator of business activity. Shifting the focus to external relationships, Li et al. (2018) used historical website data to quantify the influence of external relationships on firm growth. Their use of a `Hausman-Taylor panel data model` was methodologically significant, allowing them to control for unobserved firm characteristics while analyzing keyword-derived Triple Helix indicators. Their key finding was that it was not individual relationships but specific combinations (especially government-industry) that drove sales growth. A limitation, however, is that their keyword-based proxies for complex concepts like university relationships may be too simplistic to capture the true nature of these collaborations.

Moving beyond relationship analysis, Arora et al. (2020) attempted to measure abstract dynamic capabilities. Their novel contribution was to use Latent Dirichlet Allocation (LDA) to treat changes in website topics as a proxy for strategic seizing. To address the obvious endogeneity between website changes and firm performance, they employed a two-stage least squares (2SLS) regression. The result was a nuanced inverse U-shaped relationship between strategic change and sales growth, suggesting an optimal level of change exists. However, a critical limitation of their study, which they acknowledge, was that their measure for sensing capabilities, changes in R&D keyword mentions, was not found to be significant, highlighting the ongoing difficulty in using simple heuristic methods to capture complex strategic functions.

The maturity of the heuristics-based approach is evident in recent studies that integrate website analysis into a broader, multi-source data framework. Calvino et al. (2022) exemplified this by creating a composite picture of AI-adopting firms through the triangulation of website text with patent records, online job postings, and financial data. This integrated approach provides a more robust characterization of innovative firms than any single source could, though it also introduces challenges related to the complexity and potential for conflicting signals between disparate data types. Similarly, Antonelli et al. (2022) employed `semantic analysis` on web sources and patent data to explore the Open COVID Pledge. Their key finding was a disconnect between broad corporate statements and the narrow technical focus of the actual pledged patents, demonstrating the power of text mining to critically evaluate corporate discourse. A limitation of their approach is its reliance on a proprietary semantic tool, which hinders the replicability of their specific methodology. These studies illustrate a clear trend: website data is no longer treated as a standalone novelty but as a core component within a sophisticated, multi-source analytical toolkit.

2.8.2 Classical Machine Learning phase

The Classical-ML phase introduced supervised or unsupervised machine-learning models trained on engineered features extracted from text. Common methods involved representing text as high-dimensional feature vectors using techniques like bag-of-words, n-gram Term Frequency-Inverse Document Frequency (TF-IDF), or topic distributions from Latent Dirichlet Allocation (LDA). These vectors were then combined with algorithms such as Support-Vector Machines, Random Forests, Gradient-Boosting, or k-means. The typical workflow involved embedding text data as feature vectors, training classifiers or regressors against ground-truth labels (like a “product innovator” label from a survey), and evaluating performance using metrics like accuracy, F1-score, or Area Under the Curve (AUC). This phase represented a significant step-change, delivering improved recall and precision over heuristic methods while remaining computationally tractable. Feature importances and topic terms still offered some degree of interpretability. Nevertheless, its main drawbacks included requiring manual feature design, struggling with linguistic complexities like polysemy and word order, and losing fine-grained semantic nuances such as sarcasm or negation.

A central goal during this phase was to train classifiers that could automatically identify innovative firms from website text. This offered a scalable and timely alternative to traditional innovation surveys like the Community Innovation Survey (CIS). Kinne & Lenz, (2021) exemplified this by training a deep neural network on TF-IDF vectors of website texts, using German CIS data as ground-truth labels. Achieving an F1-score of 0.80, and its predictions correlated well with patent data and regional innovation indicators, demonstrating the viability of this approach for large-scale analysis.

Similarly, a logistic regression model with L1 regularization was used on Dutch firms, achieving an impressive accuracy of 93%. Their study showed that their estimates for innovative companies closely matched official CIS figures. Taking a different approach to model architecture, Mirończuk & Protasiewicz (2020) developed a stacked generalization method for Polish companies. They combined multiple Naïve Bayes classifiers trained on different parts of a website (main page text, link labels, etc.) into a single, more robust meta-classifier. While these studies successfully demonstrated high predictive accuracy, a critical dependency of these supervised methods is their reliance on survey data for ground-truth labels, which are time-lagged, potentially biased, and may not fully capture the spectrum of innovation activity.

The success of these models hinged on extensive and creative feature engineering. Axenbeck & Breithaupt (2021) provided a comprehensive overview of the feature landscape, cataloging indicators derived from website text (TF-IDF, topic models, readability scores), meta-information (website size, domain age), and network structures (hyperlink centrality). This work highlights the core activity of the Classical-ML phase: manually designing proxies for innovation from raw web data. Pushing this further, some research moved beyond visible text to analyze a website’s underlying structure. Crosato et al. (2024) enhance credit risk modeling by extracting features directly from website HTML code using Multiple Correspondence Analysis to create quantitative factors. Similarly, Bottai et al. (2024) used the frequency and co-occurrence of specific HTML tags to distinguish innovative from non-innovative Italian SMEs, finding that innovators used more structured and modern coding practices. While innovative, these approaches underscore a key limitation of the Classical-ML era: the reliance on specific tags, keywords, or hyperlinks that limit the full understanding of the context.

Researchers also leveraged website data to move beyond firm-level classification into the analysis of inter-firm relationships and new applications. The hyperlink network became a key object of study. Abbasiharofteh et al. (2023) constructed a "Digital Layer" of German firms, using logistic regression to classify hyperlink types and finding that firms' innovation capabilities were positively associated with the quality and quantity of their business-related links. In a related study, Arifi et al. (2023) compared these hyperlink networks to Twitter follower networks for IT companies. They revealed that while structurally different, connection decisions in both were driven by similar geographic and cognitive proximity factors. However, a key limitation of hyperlink analysis is the ambiguous nature of a link, which can represent anything from a business partnership to a simple citation, requiring significant effort to classify correctly.

The versatility of this phase is evident in how researchers applied these methods to generate novel indicators for diverse economic phenomena, moving beyond a simple innovative/non-innovative binary. For instance, Nathan & Rosso (2022) focused on identifying specific innovation events. They applied structural topic modeling to a vast corpus of online news and website content to distill discrete "product/service launch" events from raw textual mentions. This created a new, high-frequency indicator of innovation output. However, a methodological weakness of this unsupervised approach is its potential for ambiguity. The model might identify marketing campaigns or minor updates as launches.

Moving further afield from innovation, Mirtsch et al. (2020) demonstrated the methodology's power to track the adoption of management standards. They combined large-scale web scraping with a probit model guided by the Technology-Organization-Environment framework to analyze the drivers of ISO/IEC 27001 adoption in Germany, uncovering a significant landscape of "indirect certification" invisible to official data. This study's scalability was constrained by its reliance on keyword search, which misses conceptual mentions of security standards. Additionally, it required a highly labor-intensive manual categorization process to interpret why the standard was mentioned.

The complementary role of web data was also highlighted in the financial domain. Crosato et al. (2024) explored how website indicators could enhance SME credit risk modeling. They combined historical accounting data with web indicators scraped from the Wayback Machine, using Multiple Correspondence Analysis to create factors from thousands of text and HTML features. While the web features provided valuable complementary information by correctly re-classifying over 8% of defaulted firms that the traditional model missed, a key trade-off of their complex feature engineering was a loss of interpretability. This made it difficult to pinpoint which specific website elements were driving the predictions.

The most advanced work in this phase sought to enrich website data by linking it to external knowledge bases, thereby creating more granular and dynamic firm classifications. Hajikhani et al. (2022) presented a novel method to connect company website text to the scientific literature. They calculated the cosine similarity between a firm’s TF-IDF vector and the vectors of scientific Fields of Study from the Microsoft Academic Graph. This approach offers a much richer classification of a firm’s knowledge base than traditional industry codes like NACE. In a complementary approach, the OECD working paper by Dernis et al. (2023) used web mining to identify actors in the AI ecosystem and then matched these firms to patent and trademark databases. This integration revealed an interesting finding: while many firms claim AI expertise on their websites, fewer than 10% protect these innovations with formal IP. This highlights a disconnect between public positioning and traditional innovation metrics.

In summary, the Classical Machine Learning phase marked a significant leap forward, establishing that engineered features from website data could be used to build high-performing predictive models for a variety of economic and innovation-related tasks. However, its reliance on labor-intensive feature engineering and its inherent inability to grasp deep linguistic context and semantic nuance ultimately paved the way for the next evolution in methodology: the adoption of end-to-end deep learning models based on language embeddings.

2.8.3 Advanced NLP phase

The most recent development is the Large-Language-Model (LLM) phase, which leverages large, pretrained transformer-based models (such as BERT, RoBERTa, and GPT) to obtain dense semantic representations or perform classification tasks directly. These transformer models are neural networks that can understand context and meaning of a text. The typical workflow involves passing raw text, through a pretrained LLM to obtain contextual embeddings. Contextual embeddings are numerical representations that capture the meaning of words based on their surrounding context. Researchers can then fine-tune the model or use prompting for specific labels, or cluster the embeddings to create unsupervised indicators. This phase is impactful because it captures abstract, context-dependent cues, supports multilingual data, and enables rapid prototyping with minimal or no labeled training data. However, it presents new drawbacks, including higher computational costs, challenges in reproducibility due to model updates and random elements in text generation, potential bias inherited from the pretrained models, and a lack of transparency regarding how specific signals contribute to predictions. Two recent studies illustrate both the potential and practical applications of this LLM-based approach in economic monitoring.

The power of this approach for real-time economic monitoring was demonstrated by Dörr et al. (2022), who developed an integrated framework to provide timely policy information during the COVID-19 pandemic. They scraped 1.18 million German corporate websites using the ARGUS web scraping tool. Next, they employed a fine-tuned XLM-ROBERTa model to classify text passages into categories like “Problem” or “Adaption”. Their key finding, validated through Probit regressions against subsequent credit rating data from Creditreform, was that the web-based indicators that served as significant leading indicators for financial distress, providing insights much faster than traditional business surveys. The study showcases how fine-tuned LLMs can extract nuanced, policy-relevant signals from unstructured web text during a crisis.

Shifting from classification to understanding economic phenomena, Dahlke et al. (2024) investigated the epidemic-like diffusion of AI among firms in the DACH region (Germany, Austria, and Switzerland). They used a transformer language model on the webAI database, a proprietary dataset provided by the company ISTARI.AI, to identify firm-level AI adoption from website text, differentiating between deep and superficial knowledge. By constructing inter-firm networks and applying a binomial generalized linear model, they found that AI adoption was significantly driven by three epidemic mechanisms: co-location in knowledge hot-spots, direct exposure to firms with deep AI knowledge, and centrality within the AI knowledge network. Their work illustrates how LLM-derived indicators can be used to analyze complex systemic processes like technology diffusion, revealing that knowledge spread is highly clustered and faces barriers to broader dissemination.

2.9 Sustainability measure using website

Building on the evidence presented before, the proliferation of corporate sustainability frameworks reveals a critical challenge: reporting lacks standardization, with researchers identifying over 2,500 unique metrics across different frameworks. This inconsistency raises questions about how companies effectively communicate their environmental commitments. However, companies in environmentally sensitive sectors face high reputational risks if they misrepresent their sustainability efforts. This suggests that prominent environmental claims on corporate websites may be more authentic signals than detailed sustainability reports.

The examination of corporate website sustainability communication reveals a critical implementation-communication disconnect that spans multiple industries and methodological approaches. For instance, Siano et al. (2016) established the foundation for understanding this disconnect. Their OSEC model assessment of 37 energy and utilities companies' websites identifies that while companies demonstrate strong content delivery, they consistently underperform in orientation, structure, and ergonomics dimensions. This suggests that effective online sustainability communication requires more than just information dissemination.

Building on this framework, Centobelli et al. (2020) exposes a fundamental gap in corporate practices. Their comprehensive analysis of 1,275 logistics service providers' web-based materials across Germany, Italy, and the UK finds that companies maintain green practices claims without corresponding technological infrastructure for measurement and reporting.

This disconnect becomes even more pronounced in subsequent research. Calabrese et al. (2021) analyzed 23 fitness equipment manufacturers' websites using their SOSI Matrix framework. They demonstrate that firms integrating high levels of service innovation with sustainability practices via their online presence achieve the greatest SDG contributions. Yet many fail to effectively communicate this integration.

The pattern reaches its most striking manifestation in the fashion industry. SanMiguel et al. (2021) conducted a systematic evaluation of major fashion groups' corporate websites and e-commerce platforms using an adapted OSEC framework. Corporate sites achieve strong sustainability communication scores (76-80/100), yet individual brand sites perform significantly worse. Luxury brands particularly struggle, with some prestigious labels receiving zero scores due to complete absence of sustainability content

In summary, this chapter has established a clear theoretical through-line connecting innovation, sustainability, and venture capital financing with the critical challenge of measurement. The literature review has highlighted a significant shift from traditional innovation surveys towards more dynamic, big data-driven indicators, with web-based data emerging as a particularly rich source.

Despite these advancements, significant gaps persist in the literature. The powerful financial influence of venture capital on innovation is well-documented, as is the theoretical potential of web-based metrics. However, these two streams of research have yet to be systematically integrated. Moreover, the literature that use novel, web-derived indicators to assess and compare the environmental and innovation activities of firms yield mixed results, hampering the use of web indicators for real applications. Furthermore, many studies that develop web-indicators are not grounded in Signalling Theory. Consequently, they lack the theoretical framework to fully explain their results or to connect the literature on web-metrics with that of innovation and financing.

This thesis is positioned within this gap. Its primary objective is to investigate whether big data from corporate websites can serve as a reliable indicator of the underlying innovation and sustainability quality of firms and venture-backed enterprises. By reaching this goal, this study will develop a novel framework to address this question, the methodology for which is detailed in the following Chapter 3.

CHAPTER 3 RESEARCH DESIGN

3.1 Research question and propositions

Chapter 2 identified two major limitations of traditional innovation indicators. First, they rarely address environmental and social dimensions, which are now critical to the sustainable development agenda. Second, relying on data from patents and surveys, they fail to offer real-time feedback for policymakers or timely insights for firms. These gaps underscore the need for alternative approaches to measuring innovation.

In response to these limitations, social science researchers have turned to web data over the past decade, particularly company websites, due to their ability to reflect a firm's strategic priorities and public image. Signalling Theory, i.e., the study of signals that represent the underlying quality of a subject to reduce information asymmetry (Spence, 1973), is a well-suited framework to study the indicators using companies' website. This approach is legitimate because companies intentionally choose what information to display on their websites, and because any information that reveals company quality can be considered a signal (Steigenberger et al., 2024). Although it seems difficult to leverage corporate websites or other web sources to create reliable indicators, researchers achieved mixed results. For instance, Gök et al. (2015) compared a web-based R&D indicator created through keywords frequency analysis with traditional questionnaire-based R&D indicators, finding no significant correlation, while, Héroux-Vaillancourt et al. (2020) built innovation indicators related to R&D, IP protection, collaboration and external financing, from the complete texts of 79 corporate websites obtaining a weak correlation with the real value from the surveys. Concerning the ESG aspect, several researchers tried to evaluate ESG performance disclosure using websites, but the evaluation is for the majority based on indicators that represent performances that the company itself disclosed on the website, creating heavy biases (Centobelli et al., 2020; Lopez, 2020; SanMiguel et al., 2021; Siano et al., 2016).

These mixed results may stem from methodological limitations, as the majority of web indicators for innovation and for the sustainable website have been created using keywords search and classical ML. Most web-based innovation indicators have been created using keyword searches and traditional machine learning approaches. Keyword searches have limitations including missing synonyms and lack of context, while traditional ML methods require manual feature design and struggle with linguistic complexities like polysemy and word order. Advanced NLP methods have been used only rarely and not directly for creating innovation indicators. For instance, Dörr et al. (2022) used advanced NLP to develop survival indicators for companies during the COVID-19 pandemic, though their focus was not on innovation, while Dahlke et al. (2024) studied the AI adoption among companies. Additionally, the LLMs offer new perspectives and perhaps better results than previous studies when it comes to the creation and validation of web indicators. If not to substitute them, the integration of website data with traditional measures enriches our understanding of innovation dynamics and ESG performance by providing timely, granular, and scalable indicators. Given these limitations and opportunities and inspired by the paper of Connelly et al. (2011), we aim to answer the following research question:

Do website signals represent valid and reliable measures of the underlying innovation and sustainability quality that organizations are attempting to communicate?

To address this main research question, we conducted two studies. Paper 1 and Paper 2 together test the validity of website signals for both sustainability and innovation, addressing our core research objective from complementary perspectives.

In paper 1, we look at the environmental signal. This study aims to determine if web-based environmental culture indicators, derived from analyzing company homepage content, can act as proxies for established environmental performance measures, specifically the B-Lab environmental index**.** To achieve this, we test several propositions divided in 2 groups. The first group includes Proposition 1. This is directly connected to the main goal of the research stating a positive correlation between environmental compliance indices (B-Lab index) and web-based indicators, suggesting that companies with stronger website signals tend to have better environmental performance. The second group encloses 8 different propositions. Proposition 2 states that the country in which a company is located influences its environmental compliance index (Aguilera-Caracuel & Ortiz-de-Mandojana, 2013; de Azevedo Rezende et al., 2019; Doran & Ryan, 2012); Proposition 2m posits that the relationship between a company’s web-based environmental culture indicators and its environmental compliance index is moderated by the country in which it is located (Cowan & Guzman, 2020; Magnusson et al., 2011); Proposition 3 claims that the size of a company influences its environmental compliance index (Aguilar-Fernández & Otegi-Olaso, 2018; de Azevedo Rezende et al., 2019); Proposition 3m argues that the relationship between a company’s web-based environmental culture indicators and its environmental compliance index is moderated by its size (Hoehn-Weiss & Karim, 2014); Proposition 4 asserts that the industrial sector in which a company operates influences its environmental compliance index (Hermundsdottir & Aspelund, 2021a; Tariq et al., 2017); Proposition 4m posits that the relationship between a company's web-based environmental culture indicators and its environmental compliance index is moderated by the industrial sector in which it operates (Yildiz et al., 2023).

While Paper 1 explores environmental signals, Paper 2 extends the framework to fundings related signals, focusing on private funding dynamics as a key driver of innovative firms. The objective of paper 2 is to explore the feasibility of using company's website for real-time monitoring of private funding dynamics. Private fundings as saw in the literature review is of crucial importance nowadays for firms and governments to develop innovative companies (see section 2.3). Grounded in signaling theory, as paper1, the paper proposes four key propositions: Proposition 1 asserts that companies with private funding are more likely to mention additional funding on their websites, reflecting prior research on government grants as signals for follow-on VC funding (Islam et al., 2018; Wu et al., 2020) Proposition 2 suggests these companies emphasize founder experience and background, consistent with the role of human capital in securing investment (Bhattacharyya & Subrahmanya, 2024; Ko & McKelvie, 2018); Proposition 3 posits that funded companies highlight collaborations, aligning with studies on alliances as indicators of startup potential (Baum & Silverman, 2004; Caviggioli et al., 2020; Hoenig & Henkel, 2015; Miloud et al., 2012); and Proposition 4 states that topics communicated by funded companies differ from those without funding, influenced by market dynamics and economic cycles (J. Chen & Ewens, 2021; Corea et al., 2021; P. Gompers et al., 2021; P. Gompers & Lerner, 2001b; Howell et al., 2020). Together, these approaches and propositions enable a robust system for policymakers and managers to track investment and entrepreneurial trends efficiently. This contributes to the thesis objective in two ways. Signals of funding, founder background, and collaborations potentially reflect innovation, and the paper aims to assess the validity and reliability of website signals, suggesting the paper's approach could be adapted or serve as a starting point for evaluating innovation and sustainability signals.

3.2 Objectives

The objectives of this research are:

Objective 1: Determine if web-based environmental culture indicators, derived from the text on company homepages, can act as reliable proxies for actual environmental performance measures.

- Identify the corporate websites of the B Corp's firms;
- Analyse the content of these websites to create web-indicators relate to environmental performances;
- Compare the web-based indicators with those in B Corp data;
- Assess the feasibility of using web-based signals for near-real time monitoring of environmental performances.

Objective 2: Explore the potential of using company websites for the real-time monitoring of private funding dynamics.

- Identify the corporate websites of the CrunchBase's firms;
- Analyse the content of these websites to create web-indicators relate to private funding;
- Compare the web-based indicators with those in CrunchBase;
- Assess the feasibility of using web-based signals for near-real time monitoring of private fundings.

Objective 3: Determine whether it is possible extract reliable, innovation and environmental related signals from companies' websites to monitor industry trends.

- Identify the corporate websites of the SIBS's firms;
- Analyse the content of these websites to create web-indicators relate to environmental and innovation performances;
- Compare the web-based indicators with those in SIBS;
- Assess the representative capacity of web-based signals compared to the SIBS' indicators.

3.3 Methodologies

This chapter outlines the methodological framework employed to address the research aims of this thesis, which is to verify whether signals represent valid and reliable measures. The following paragraphs are organized in three different parts: Data, Paper 1 and Paper 2. The Data section introduces the datasets used explaining the reason behind their choice, and explains the crawler program used to retrieve the website content. Then, the detailed methodologies used in paper 1 followed by paper 2 and study 3 are presented. Concerning study 3 presented in chapter 6, the methodology is the same as Paper 2, thus we only discuss the data used for this latter contribution.

3.3.1 Data

For Paper 1, we employ the B-Corp dataset as our primary data source based on its demonstrated theoretical value and practical accessibility. The dataset has established substantial scholarly credibility through its application in peer-reviewed research examining legitimacy signaling, employee productivity, and growth effects (S. Kim & Schifeling, 2016; Liute & De Giacomo, 2022; Paelman et al., 2020; Romi et al., 2018). From a practical standpoint, the dataset maintains open accessibility through established data platforms[9], facilitating reproducible research. Our analysis focuses specifically on the United States and Canada due to their similar regulatory frameworks regarding ESG policy implementation. Moreover, this geographic focus is particularly relevant given that B-Corp certification originated in the United States in 2006, and these markets represent a significant proportion of certified entities among the program’s expansion to over 160 countries encompassing nearly 10’000 certified companies. The dataset provides comprehensive organizational identifiers including legal names, unique identification codes, and website information, alongside critical certification variables such as initial certification dates, expiration date. The scoring framework represents a particularly valuable component, providing each organization with an overall score and detailed evaluations across five primary categories: governance, environment, customers, community, and workers, calculated as averages of more granular assessment criteria to deliver both aggregate and detailed performance metrics. This combination of established academic credibility, comprehensive data structure, detailed scoring mechanisms, and strong coverage within our target geographic markets positions the B-Corp dataset as both theoretically sound and empirically robust, enabling systematic analysis of how corporate sustainability practices and disclosures translate into measurable outcomes across multiple performance dimensions.

[9] https://data.world

In paper 2, we use CrunchBase, a business platform that CrunchBase dataset, created by CrunchBase[10]. CrunchBase is one of the most widely used, publicly available platforms for startup, investment, and entrepreneurial data. CrunchBase has established itself as one of the most comprehensive and widely adopted platforms for accessing detailed information about emerging companies and their financing activities. The dataset encompasses quantifiable attributes that systematically describe and categorize companies across multiple dimensions, including industry classification, organizational size, revenue metrics, geographic location, legal structure, and developmental stage. Additionally, the dataset captures detailed information about equity transactions, debt arrangements, and grant funding that companies have publicly disclosed. This comprehensive approach enables researchers to analyze both structural company characteristics and financing trajectories within a unified analytical framework. The dataset also incorporates summary metrics displayed on organizational profiles, such as total funding raised and most recent funding type, providing accessible overview information for each entity. These summary fields facilitate rapid assessment of company funding status and investment activity levels. CrunchBase has served as a foundational dataset for numerous academic studies over the past decade. Researchers have leveraged its extensive repository of company, founder, funding round, and investor information for diverse analytical purposes, including venture capital investment outcome prediction, early-stage company classification, and enhanced founder profiling through integration with complementary data sources. The dataset's quality is demonstrated by its use in several academic studies (Arroyo et al., 2019; Razaghzadeh Bidgoli et al., 2024; Żbikowski & Antosiuk, 2021). For our analysis, CrunchBase represented the optimal data source due to its established credibility within academic research communities and its alignment with our research goal. The platform's focus on private funding transactions and comprehensive firm-level information, including website URLs, made it particularly well-suited for our analytical objectives.

For the third contribution, we leverage the Survey of Innovation and Business Strategy (SIBS). The SIBS is a Statistics Canada's key instrument for understanding how Canadian enterprises compete and evolve, conducted roughly every three years to provide a panoramic view of their strategic decisions, innovation activities, and operational tactics. Designed to inform federal and provincial policy-making and academic research, SIBS links firms' management choices to

[10] https://www.crunchbase.com

outcomes like productivity and growth by asking detailed questions that fill gaps left by simpler metrics like R&D spending. It targets enterprises with 20 or more employees and at least $250,000 in revenue. The survey delves into a wide array of topics that can be divided in three groups: Innovation, with questions related to types, novelty, expenditures, obstacles, environmental benefits. Business strategy & advanced technology with questions related to markets, competition, skills, tech adoption, clean tech. Global value chain with questions related to imports, exports, affiliate sales, barriers. All the main concepts related to innovation and their definition are aligned to the Oslo Manual. SIBS provides an indispensable, evidence-based snapshot for benchmarking Canada's business ecosystem, with its anonymized microdata accessible to researchers for in-depth analysis.

3.3.2 Web Crawler

For both paper 1 and 2, the research used a crawler to retrieve the website content from the Wayback Machine, i.e., an online archive that contains the snapshot of websites. To achieve this goal, I used the the Wayback Machine Crawler developed by the 4POINT0 Chair Innovation–Poly MTL team. The program automates the bulk retrieval of historical website snapshots from the Internet Archive, extracting clean text and metadata for storage in MongoDB databases. The crawler targets researchers who require large, time-stamped corpora for academic studies, particularly those examining firm-level innovation patterns over time, and operates on a single machine to process approximately 10,000 pages per hour while respecting the Internet Archive's rate limits through an intelligent "gentle multi-retry" system. Users simply provide a comma-separated list of targets in the format "example.com, 2014". Wayback_Machine_Crawler processes this file by downloading up to five language pages per snapshot (or selecting the shortest available URLs), parsing the HTML content, and storing successful records in a designated data collection while logging any irrecoverable errors to data_err. The crawler employs BeautifulSoup to parse each archived page with a sophisticated content filtering approach that removes non-content elements before extracting text, specifically discarding script tags (<script>), style sheets (<style>), non-script content (<noscript>), vector graphics (<svg>), embedded frames (<iframe>), headers and footers (<header>, <footer>), and HTML comment nodes that would otherwise pollute the corpus. From the cleaned raw web content, the system preserves visible text from content-bearing elements including structural headings (<h1>–<h6>), paragraphs (<p>), list items (<li>, including <ul>/<ol> markers when bullets carry semantic meaning), table cells (<th>, <td>), and anchor text within <a> elements, as link labels frequently identify products, brands, or partners. The crawler also captures document metadata, including the <title> content and descriptive strings from <meta name= “description”> and <meta name=“keywords”> tags when present, with all extracted plain text undergoing normalization where HTML entities are decoded before storage. The final records include the cleaned text alongside Wayback timestamps, original URLs, and basic snapshot metadata in the specified MongoDB data collection, employing a tag-whitelisting strategy that preserves narrative and label content essential for downstream NLP pipelines.

3.3.3 Article 1

In the following section, I will explain in detail the methodology, displayed in Figure 3.1, used to reach the objectives of article 1. Article 1 aims to assess whether web-based indicators from company homepages can replicate the B-Lab environmental index by first collecting textual data from company homepages (via the Wayback Machine, using the URL in B-Corp data) and their corresponding B-Lab environmental scores. This textual data provides raw “signals” of environmental culture and a recognized environmental performance benchmark. Next, we proceed to the construction indicators that involve two steps: fist, a BERT-type Natural Language Processing (NLP) model (Bidirectional and Auto-Regressive Transformer – BART) performs Zero-Shot Text Classification (ZSTC) on the homepage text, classifying it against 31 B-Corp environmental assessment labels to generate initial “web-based environmental culture indicators” that quantify environmental themes without specific pre-training. Then, to refine these numerous indicators, Principal Component Analysis (PCA) is applied, reducing them to a smaller set of six aggregated, orthogonal factors representing broader dimensions of web-based environmental signals. Finally, in the validation step, these PCA-derived factors, along with company characteristics (like size, industry, and country) and control variables, are used in Ordinary Least Squares (OLS) regression model. The regression models estimate the factors that are associated with the B-Lab environmental index, testing the study’s propositions about these relationships. The coefficient of determination (R^2) of the regression determines the proportion of variance in the B-Lab environmental index that can be explained by the homepage signals. The text below explains every method used in article 1 highlighting the advantages and motivating their choice.

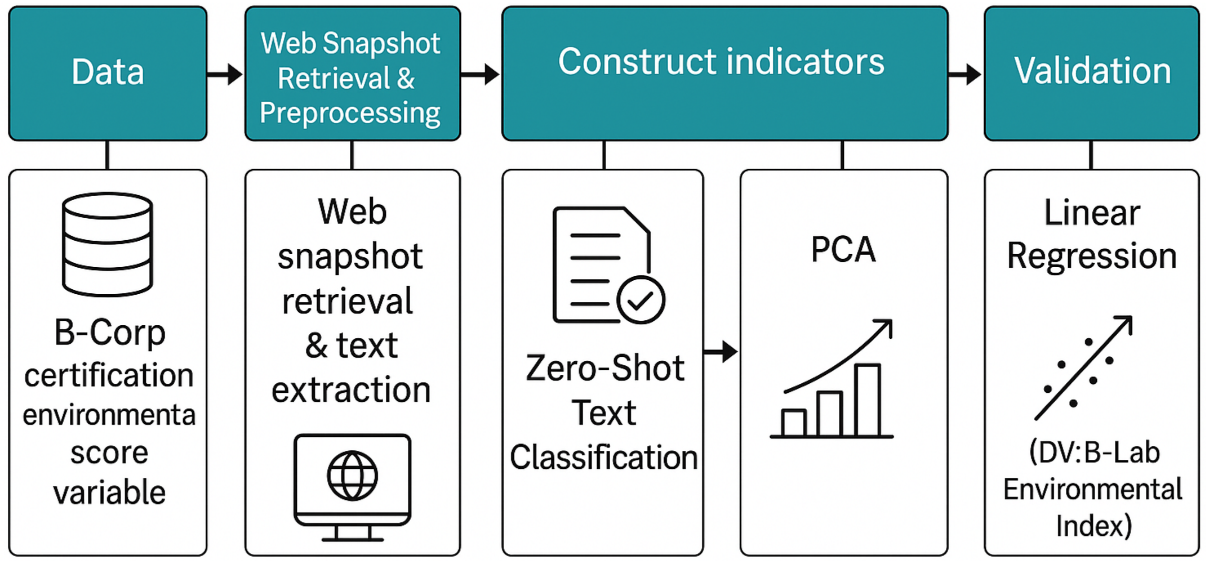
Data
Web Snapshot Retrieval & Preprocessing
Construct indicators
Validation
B-Corp certification environmenta score variable
Web snapshot retrieval & text extraction
Zero-Shot Text Classification
PCA
Linear Regression
(DV:B-Lab Environmental Index)

Figure 3.1 Methodological pipeline Article 1

Zero-shot text Classification

One of the biggest challenges in AI is the lack of large, properly labeled datasets for many real-world problems. Without these datasets, it is difficult or impossible to train supervised machine learning models effectively. In Article 1, we use zero-shot text classification (ZSTC) to determine whether company websites contain content related to environmental topics overcoming the missing annotated sample.

Transfer learning, zero-shot learning (ZSL), and zero-shot text classification (ZSTC) are interrelated paradigms that address the challenge of generalizing from known data to entirely novel or underrepresented target classes. Transfer learning broadly emphasizes the reuse of previously learned knowledge from a source domain to improve learning in a related target domain (Zhuang et al., 2020), whereas ZSL is a specialized branch of transfer learning that specifically focuses on enabling the recognition and classification of unseen classes without any labeled training examples for those classes (Xian et al., 2017). ZSTC further adapts these principles to the natural language processing (NLP) domain by leveraging semantic relationships, auxiliary knowledge, and pretrained representations to assign labels to text data that were not encountered during training (Yin et al., 2019).

Although the concept of transfer learning was introduced decades before deep learning, most early NLP systems were trained from scratch for each task using task-specific features and statistical models. The paper of Collobert et al. (2011) represented major milestone in the rise of transferable representation. They demonstrated that neural networks could learn features from large amounts of unlabeled training data, yielding internal representations that support multiple NLP tasks. They discovered that word embeddings, i.e., dense vectors learned from unlabeled corpora, capture semantic and syntactic properties of words. Soon after, Mikolov et al. (2013) created the Word2Vec algorithms, Pennington et al. (2014) introduced GloVe, and Bojanowski et al. (2017) created FastText. Although these models differ algorithmically, they efficiently train word vectors on billions of words, providing general semantic knowledge that encodes linguistic regularities and groups similar words together. By late 2018, NLP saw the advent of large-scale transformer-based pretraining, which took transfer learning to new heights. The Transformer architecture (Vaswani et al., 2017) enabled much deeper and more expressive models. At the same time, Devlin et al. (2018) released BERT (Bidirectional Encoder Representations from Transformers), which became a cornerstone of NLP transfer learning. BERT introduced a bidirectional pretraining objective (Masked Language Modeling + Next Sentence Prediction) to learn deep representations that incorporate both left and right context.

As mentioned above, ZSL tasks focus on transferring knowledge from seen classes to unseen classes. This is challenging because shared patterns among classes are rare. In zero-shot image classification, researchers typically establish connections between novel and familiar classes by exploiting various forms of semantic knowledge, including visual attribute descriptions (Lampert et al., 2009), distributional word representations derived from class labels (Norouzi et al., 2014), and hierarchical taxonomic relationships between categories (Socher et al., 2013). For zero-shot text classification, similar methods have been adopted (Ye et al., 2020). Although ZSL was attempted even before deep learning explosion to classify the text, a significant development was leveraging entailment. Natural Language Inference (NLI), i.e. determining whether a premise sentence entails or contradicts a hypothesis sentence, was recognized as a proxy for content understanding. In this vein, Bowman et al. (2015) released the large Stanford Natural Language Inference (SNLI) dataset, enabling robust NLI models. This dataset opened new possibilities for transfer learning and demonstrated the broader applicability of NLI frameworks beyond traditional entailment classification tasks.

The NLI approach was explicitly proposed by Yin et al. (2019) who unified various zero-shot classification tasks under a textual entailment framework. They trained the models by giving a candidate label in natural language (e.g. *"sports"* or even a full description like *"the text is about sports"*) and judging entailment versus contradiction. If the label statement is entailed, the model assigns that label to the text. Through this research they demonstrated that an NLI model trained on general-purpose entailment data can be repurposed, with no additional training, to solve a wide range of classification problems it has never seen before.

In article 1, we employ BART fine-tuned on the Multi-Genre Natural Language Inference (MNLI) dataset to predict entailment probabilities for input-hypothesis pairs. This approach enables zero-shot text classification (ZSTC) by allowing users to incorporate new labels written in plain language without additional training[11]. BART (Bidirectional and AutoRegressive Transformers) is a transformer-based deep learning model developed by Facebook AI for NLP tasks (M. Lewis et al., 2019). The model's key innovation was the combination of BERT's bidirectional contextual understanding with GPT's autoregressive generation capabilities within a unified denoising autoencoder architecture (Radford et al., 2018). This design allows BART to both comprehend text bidirectionally and generate coherent sequences from left to right.

Principal Component Analysis

Principal Component Analysis (PCA) is a multivariate method used to reduce the dimensionality of a data set with a large number of interrelated variables. The central idea is to retain as much of the original variation as possible during this reduction. This reduction is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that the first few retain most of the variation present in all of the original variables (Jolliffe, 2002). Originally introduced by Pearson (1901) and later formalized by Hotelling (1933), Principal Component Analysis (PCA) is a powerful technique that transforms a set of possibly correlated variables into a new set of linearly uncorrelated variables, known as principal components. Derived as linear combinations of the original data, these components are ranked according to the proportion of the total variance they explain, making PCA an effective tool for reducing

[11] https://huggingface.co/facebook/bart-large-mnli

dimensionality. To perform Principal Component Analysis (PCA), several assumptions should ideally be met. Firstly, it is assumed that there are multiple variables that have been measured at a continuous level, although in practice, ordinal variables are very frequently utilized. A second assumption is the presence of a linear relationship between all variables. This is crucial because PCA is based on Pearson correlation coefficients, which quantify linear associations. While this assumption can be somewhat relaxed in practice, particularly with ordinal data, it is important to consider. Non-linear relationships may necessitate data transformations. Thirdly, a sufficiently large sample size is necessary is required for PCA to yield reliable results. Sampling adequacy can be formally assessed using measures like the Kaiser-Meyer-Olkin (KMO) Measure of Sampling Adequacy for the overall dataset and for individual variables, as well as Bartlett's test of sphericity. The fourth assumption is that the data must be suitable for data reduction. This implies that there are adequate correlations between the variables, which is necessary for them to be effectively combined into a smaller number of components. Finally, the fifth assumption for PCA is that the data should not contain significant outliers, as these can disproportionately influence the results. It is therefore important to check for and potentially address these outliers before proceeding with the analysis[12].

While PCA is often associated with Factor Analysis, important differences exist, as highlighted in Jolliffe, (2002). Principal Component Analysis (PCA) and Factor Analysis (FA) differ primarily in their objectives and underlying statistical models. PCA aims at data reduction by finding linear combinations of variables, called components, that capture the maximum total variance, treating all variance (common, unique, and error) equally without an explicit generative model. The components are ordered by the total variance explained, and the solution is derived from the eigenvectors of the full covariance or correlation matrix. PCA does not distinguish between common and unique variance, and while rotation is optional for interpretation, components remain orthogonal if unrotated. The principal components are considered purely mathematical constructs, and any number of components up to the original number of variables can be retained without a model-based test for selection. PCA is typically used for data compression, visualization, managing multicollinearity, and preprocessing for supervised learning. In contrast, FA is a latent-variable

[12] https://statistics.laerd.com/spss-tutorials/principal-components-analysis-pca-using-spss-statistics.php (17-05-2025)

modeling technique focused on explaining observed pattern of correlations among variables through a smaller number of unobserved common factors. FA operates under a common-factor model, where each observed variables is represented as a function of these latent factors along with unique variance terms. It specifically addresses the common variance (covariances) shared among variables, separately estimating the unique variance associated with each observed variable. Solutions are typically derived using methods like maximum likelihood estimation applied to the reduced covariance matrix. Additionally, FA involves rotation, orthogonal or oblique, as a crucial step in obtaining an interpretable factor structure, with oblique rotations allowing for correlations among factors. In FA, these latent factors are interpreted as meaningful underlying constructs that drive the observed relationships. The determination of the optimal number of factors typically relies on model-based criteria. FA is frequently employed in theory-driven latent-feature measurement and scale development, especially in disciplines such as psychology and the social sciences.

In Article 1, we performed Principal Component Analysis (PCA) with the twofold goals of "1) reduce the correlation among the labels; and 2) groups them into a smaller set of components maintaining trends and characteristics." (Cruciata et al., 2024, p.7). After carefully checking for outliers and linearity and performing the PCA, we conducted tests to assess the construct and sample adequacy. Considered only items that loaded at least 0.5. The KMO yielded a value above the recommended minimum of 0.6 (Kaiser, 1974). Furthermore, the reliability of each dimension was assessed using Cronbach's alpha, with values exceeding 0.7 at exception of one component that was kept because we were in exploratory analysis (Hair et al., 1998)

Linear regression

In Article 1, we perform OLS regression as validation step to verify the contribution and level of significance of the web-based indicators in explaining the B-Corp environmental index, validating whether these "signals" accurately represent the real value of companies. A hierarchical regression analysis allows to assess the explanatory power of these web indicators regarding the variability of the B-Corp environmental index. To ensure valid regression results, we conducted comprehensive testing before and after the analysis to verify that key assumptions were met. Prior to regression, we applied several transformations to ensure variable normality and employed Principal Component Analysis (PCA) to eliminate potential correlation between dependent variables.

Following the regression, we calculated the Variance Inflation Factor (VIF) to detect multicollinearity among our predictors. The VIF quantifies how much the variance of a regression coefficient increases due to correlation with other predictor variables. A VIF of 1 indicates perfect independence between predictors, while higher values reveal increasing collinearity that reduces coefficient precision and stability. Standard practice considers VIF values below 5 as acceptable, whereas values exceeding 10 signal problematic multicollinearity that violates core regression assumptions. We also performed the Breusch-Pagan test, the standard diagnostic for heteroskedasticity, which examines whether error-term variance remains constant or varies systematically with predictors, where a significant p-value (< 0.05) indicates heteroskedasticity and necessitates robust error correction or model re-specification. Our analysis confirmed no autocorrelation and value of VIF less than 5 in the model.

3.3.4 Article 2

In the following section, I will describe in detail the methodology, depicted in Figure 3.2, used to reach the objectives of article 2.

Article 2 investigates whether company websites can serve as real-time indicators of private funding dynamics by employing a dual analytical approach with data from Canadian companies founded between 2020-2024. The research begins by collecting CrunchBase data to establish funding status and retrieving corresponding website content through the Wayback Machine using the URLs on CrunchBase. Thus, we created a comprehensive dataset that links funding outcomes of the companies with their web content. We used a double approach in the construct indicators steps. First a top-down methodology, employing a Retrieval Augmented Generation (RAG) framework with the large language model (LLM) to systematically extract predefined signals related to funding indicators, founder experience, and collaboration markers from website content. These web indicators are then combined with CrunchBase control variables in logistic regression models to assess their correlation with actual private funding status and validate established literature findings. Then, a bottom-up approach employs BERTopic to identify emergent linguistic patterns and topics from website text without prior assumptions, generating topic probabilities that serve as features for supervised machine learning models including Random Forest, XGBoost, and Neural Networks, with oversampling techniques applied to address class imbalance between funded and non-funded companies. Finally, the study interprets the most influential topics from the best-performing model using LLM analysis and Partial Dependence Plots to uncover novel, unanticipated signals that correlate with funding likelihood, potentially revealing new indicators for investment prediction.

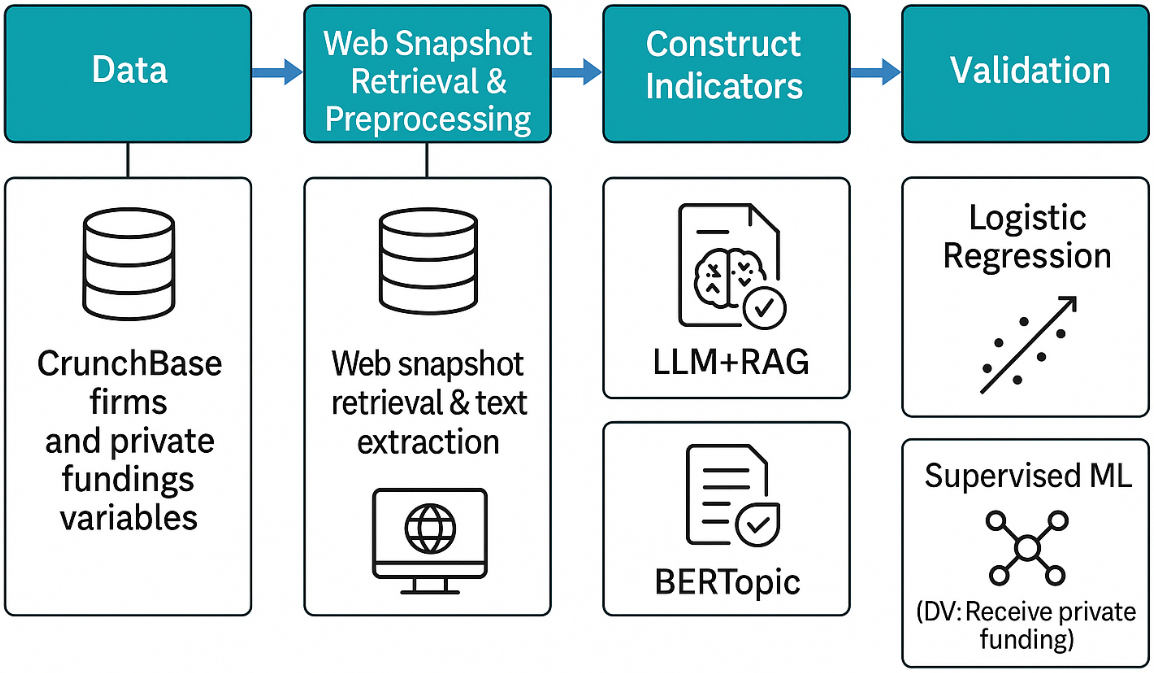
Data
Web Snapshot Retrieval & Preprocessing
Construct Indicators
Validation
CrunchBase firms and private fundings variables
Web snapshot retrieval & text extraction
LLM+RAG
BERTopic
Logistic Regression
Supervised ML
(DV: Receive private funding)

Figure 3.2 Methodological pipeline Article 2

Large Language Model

For Article 2, we leverage the capacity of large language models (LLMs), the most remarkable breakthrough achieved in Natural Language Processing (NLP), to build indicators related to the capacity of the firm to receive private fundings. LLMs have gained significant attention, especially since the release of ChatGPT in November 2022. These models excel at generative making them highly capable of complex question-answering requiring synthesis, summarization, translation, creative text generation, and synthetic data generation (Singh et al., 2025). Their strength lies in few-shot or zero-shot learning via sophisticated prompt engineering (Liu et al., 2023), allowing for flexible application without task-specific fine-tuning.

The enhancements in these models' capabilities have led to significant progress in text synthesis and a variety of downstream NLP applications (T. Brown et al., 2020) as confirmed by the increasing interest across both academic and industrial domains (Bommasani et al., 2022; Wei, Tay, et al., 2022). As demonstrated by existing work, the great performance of LLMs has raised promise that they could reach Artificial General Intelligence (AGI) shortly (Bubeck et al., 2023). Unlike previous models that were confined to specific tasks, LLMs demonstrate the capability to solve diverse problems across multiple domains. Their exceptional performance in handling applications ranging from general to specialized domain tasks has led to widespread adoption across various sectors and field (Chang et al., 2024). While the term LLM encompasses a broad range of large neural network architectures, the most prominent current developments (such as GPT-4, Claude 3, and Llama 3) are predominantly autoregressive, decoder-based models (Radford et al., 2018). In these autoregressive language models, the fundamental task involves predicting the next token based on a given context sequence.

At the core of the LLMs that allow these exceptional performances there are three main components: the self-attention module in Transformer, the feature of in-context learning (T. Brown et al., 2020), and Reinforcement Learning from Human Feedback (RLHF). Self-attention module in Transformer (Vaswani et al., 2017) serves as the fundamental building block for language modeling tasks. handling sequential data efficiently, allowing for parallelization and capturing long-range dependencies in text. In-context learning (T. Brown et al., 2020) allow the model to be trained to generate text based on a given context or prompt. This enables LLMs to generate more coherent and contextually relevant responses, making them suitable for interactive and conversational applications. Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2020) is another crucial aspect of LLMs. This technique involves fine-tuning the model using human-generated responses as rewards, allowing the model to learn from its mistakes and improve its performance over time.

Although scaling up the size of language models has been shown to confer a range of benefits, such as improved performance and sample efficiency (T. Brown et al., 2020; J. Kaplan et al., 2020; Wei, Wang, et al., 2022), this strategy alone has not proved sufficient for achieving high performance on challenging tasks such as arithmetic, commonsense, and symbolic reasoning (Rae et al., 2022). However, for some challenging tasks involving math or general reasoning, a direct generation often struggles to yield a correct answer. To address this shortcoming, researchers proposed to Chain-of-Thought prompting that induces LLMs to generate intermediate reasoning steps before reaching the answer (Kojima et al., 2022; Wei, Wang, et al., 2022). Researchers from DeepSeek AI

(DeepSeek-AI et al., 2025) brought this way to think about a problem to their model allowing it to "think out loud". DeepSeek-R1 teaches a large language model to "think out loud" through a four-steps process that involve fine-tuning the base model with a few thousand high-quality CoT demonstrations, a reasoning-oriented reinforcement learning (RL) with Group-Relative Policy Optimisation (GRPO) then optimises for answer correctness and legible reasoning, and finally, a second RL stage blends rule-based accuracy rewards with preference rewards to polish helpfulness and safety, locking in a policy that reliably emits readable CoT before the answer.

In article 2, we employ the model DeepSeek-r1[13] because of the capabilities and performances and the results that we were able to get compared to the others models according the constrained of the hardware capacity. Moreover, Deepseek-r1 model is open access, ensuring wide availability and ease of use for researchers and practitioners alike. Secondly, Deepseek-r1 is known for its outstanding performance across a broad spectrum of tasks, demonstrating exceptional reasoning capabilities.

Retrieval Augmented Generation

As aforementioned, LLMs showcase impressive abilities in natural language understanding and generation. However, they are affected by several notable limitations, including outdated knowledge, a tendency to "hallucinate" information, and difficulties in addressing domain-specific queries (Y. Zhang et al., 2023) .

The issue of AI hallucination, specifically, has stimulated significant research since its initial coinage by Maynez et al. (2020). This phenomenon describes the generation of text by LLMs that, while seemingly coherent and logical, is factually incorrect or unreal, particularly when it comes to common sense. In other words, hallucination results in output that sounds credible yet is factually incorrect. Hallucinations are a considerable concern because they not only impede the overall capabilities of LLMs but also introduce security risks, potentially leading to privacy violations. For example, Carlini et al. (2021) demonstrated that LLMs could be prompted to reveal sensitive personal details, such as emails, phone numbers, and addresses, which they had inadvertently

[13] The DeepSeek-R1 model was employed for this research before the ban put in place by **Polytechnique Montréal**. Moreover, for the purpose of the research, only the downloaded model's weights (available at https://huggingface.co/deepseek-ai/DeepSeek-R1) were used on the Chair's server for inference.

retained from their training data. Ultimately, hallucination can arise from an excessive volume of data, insufficient contextual relevance, or a combination of both, thereby undermining the reliability of LLMs in practical applications (Y. Gao et al., 2023).

Retrieval-Augmented Generation (RAG) (Figure 3.3) emerged as a solution to this issue, effectively integrating external knowledge directly into the language generation process (P. Lewis et al., 2020). A comprehensive survey by Gao et al. (2023) examines various RAG architectures developed by researchers. Based on their review, the fundamental components (indexer, retriever, generator) are common across all RAG paradigms, essentially representing the core steps of Naive RAG, the earliest framework created. The indexer is responsible for preparing the external knowledge by cleaning and extracting data from various document formats, segmenting it into manageable chunks, and then transforming these chunks into numerical vector embeddings to create an efficient search index. Once indexed, the retriever takes a user's query, converts it into a vector representation. It then efficiently searches the index to identify and fetch the most relevant "top-K" document chunk based on their similarity to the query. Retrieval methods are broadly divided into sparse and dense approaches. Sparse retrievers, such as BM25, rely on lexical matching and term-frequency statistics to score and rank documents, which can be efficient but often fail in capturing nuanced semantic relationships. Dense retrieval, in contrast, leverages transformer-based embeddings (e.g., using BERT or Sentence-BERT) to represent both queries and documents in a continuous vector space, thereby enabling semantic similarity search that goes beyond keyword overlap (S. Gupta et al., 2024). However, recent work has also explored hybrid approaches that combine the strengths of both sparse and dense retrieval. By leveraging complementary relevance signals from traditional IR metrics alongside dense semantic similarity, hybrid methods could improve overall retrieval precision and recall. Finally, the generator, which is an LLM, synthesizes the user's original query with the retrieved relevant information, creating an enriched prompt to produce a comprehensive and accurate response, with the option to constrain the LLM to rely solely on the provided context to enhance verifiability and reduce the likelihood of hallucinations. However, Naive RAG faces inherent challenges, including low retrieval precision and recall, persistent hallucinations, and the risk of generating irrelevant, toxic, or biased responses. It also struggles with coherent context integration, managing redundancy, and effectively balancing the importance of various retrieved passages.

To overcome these limitations, Advanced RAG refines Naive RAG by introducing sophisticated pre-retrieval and post-retrieval methods. The pre-retrieval process is dedicated to optimizing data indexing through enhancing granularity, optimizing index, adding metadata, aligning optimization, implementing mixed retrieval strategies and fine-tuning. Enhancing granularity involves techniques such as removing irrelevant information and special characters to boost the retriever's efficiency and reduce ambiguity. It also includes eliminating duplicate or redundant information to sharpen the retriever's focus. The overall objective when optimizing indexed data is to prioritize clarity, context, and correctness, thereby ensuring the system's efficiency and reliability. Optimizing index structures primarily entails adjusting the chunk size of the data. Adding pertinent metadata means embedding references like dates or specific purposes directly into the chunks for easier filtering. Aligning optimization focuses on ensuring that each document chunk is capable of adequately answering a posed query. Furthermore, implementing mixed retrieval strategies, such as combining keywords with semantic search, improves the overall retrieval performance. Finally, fine-tuning embedding models helps to increase the relevance between the retrieved content and the user's queries. Conversely, the post-retrieval process focuses on refining the retrieved content before it is given to the LLM, addressing issues like context window limitations and noise. This is achieved through methods such as re-ranking relevant information and prompt compression to reduce irrelevant context while highlighting pivotal paragraphs. Furthermore, RAG pipeline optimization involves the exploration of hybrid search techniques, optimizing retrieval steps, and the integration of cognitive backtracking and flexible query strategies.

A more performing technique is the modular RAG. It represents a more adaptable paradigm, allowing for the addition or replacement of various modules within the RAG process to suit specific problem contexts, and can even integrate other techniques like fine-tuning. At a high level, Modular RAG places an emphasis on the "Lego-like" assembly of different modules, each of which performs a particular function in the overall pipeline (Y. Gao et al., 2024). This modular design allows developers to upgrade, exchange, or fine-tune individual components, such as retrievers, query reformulators, or generators, without affecting the remaining parts of the system. For instance, a retriever may use vector-based similarity search techniques to extract relevant document chunks and then pass these chunks sequentially to a generation module that synthesizes a response based on both the user query and the retrieved information. (S. Gupta et al., 2024) Such clear division of labor simplifies both implementation and troubleshooting because each module can be independently tested, benchmarked, and optimized (Jin et al., 2025).

The last development in terms of RAG lead to the creation of Agentic RAG. Agentic RAG systems excel in scenarios where the retrieval requirements are dynamic, the domain context may change rapidly, or the complexity of queries demands multi-step reasoning and continuous adaptation (Singh et al., 2025). These systems are frequently applied in open-domain question answering, research report generation, and interactive dialogue systems, where the ability to autonomously decide when to seek new information is critical for ensuring accuracy and adaptability in real time. The autonomy inherent in Agentic RAG provides a significant performance boost in contexts that require iterative reasoning and integration of multiple sources of data, albeit at the cost of increased system complexity and computational overhead (J. Chen et al., 2024).

In article 2, we implemented advanced RAG. We implement the advanced RAG ensuring granularity pre-processing text to remove noise and increase consistency. For instance, we create a query for the retriever that was smaller and ad hoc to increase the chances to retrieve the majority of the chunks semantically closer, while filtering with a similarity threshold to make the retriever focus on the chunks that are more relevant. Given that we need to process the website one at the time, we do not have problems related to the post-retrieval process such as re-ranking the information. Important aspect of our pipeline is the evaluation process. We choose to do manual evaluation for two reasons: the data that are positively recuperated after the RAG are few hundred, and because we do not have extremely large document that would be need model assistance to be evaluated.

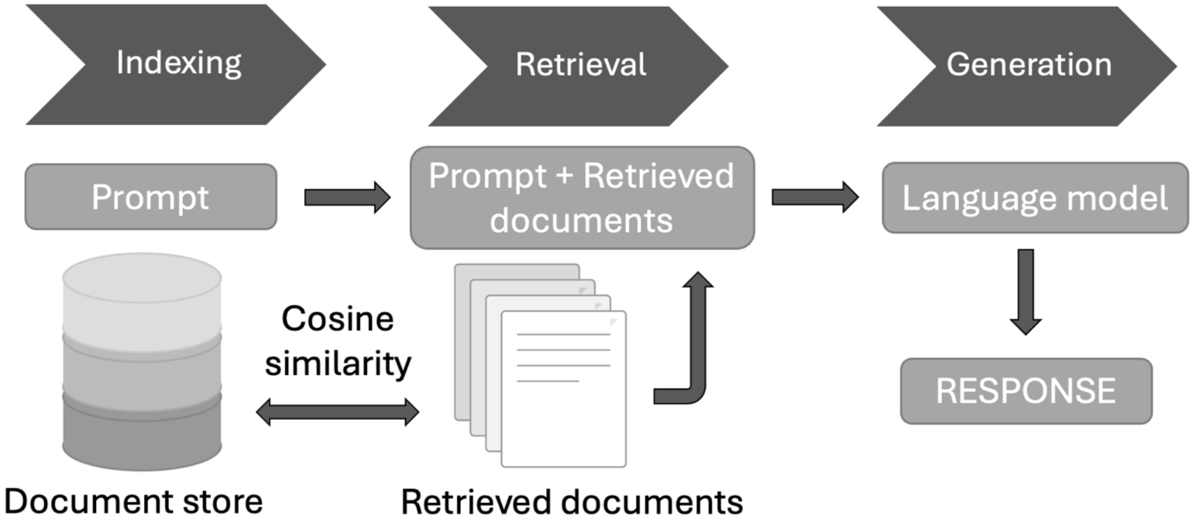

Figure 3.3 Retrieval Augmented Generation

Logistic Regression

In Article 2, the validation step for the RAG+LLM approach, and therefore of Propositions 1, 2, and 3, uses a logistic regression. The logistic Regression is employed because the dependent variable *dFunded* (having received VC funding or not) is binary. A logit model allows to (i) estimate how each signal changes the log-odds of receiving funding while (ii) controlling for well-known covariates such as firm size, growth stage, confidence scores, region, and NAICS industry. Odds-ratio outputs are readily interpretable for practitioners and align with prior empirical VC-signalling work. The logistic regression allowed to answer to Proposition 1, 2, and 3.

BERTopic

For Article 2, we also employ BERTopic, a topic model, to analyze and extract key themes from company webpages. Topic modeling falls under the category of unsupervised machine learning algorithms. Unsupervised methods operate on unlabeled data, identifying underlying patterns or structures through techniques such as clustering. In the context of topic modeling, this involves grouping words and documents based on semantic similarities. Prior to detailing the BERTopic methodology, we first present a brief overview of the historical development of topic models, leading to contemporary techniques that explain our choice.

Topic modeling represents a fundamental approach for automatically identifying thematic structures within large document collections (Blei, 2012). Traditional approaches progressed from basic term frequency methods like TF-IDF (Salton, 1983) through dimensionality reduction techniques such as Latent Semantic Analysis (Deerwester et al., 1990) to probabilistic models culminating in Latent Dirichlet Allocation (LDA) (Blei et al., 2003). While LDA established the foundation for modern topic modeling through its Bayesian framework, it suffered from limitations including topic independence assumptions and challenges with large vocabularies (Vayansky & Kumar, 2020).

The evolution of topic modeling entered a new phase driven by breakthroughs in natural language processing, particularly the advent of transformer architectures. A significant development was the release of BERT in 2018 (see section 3.2), which revolutionized language modeling through the generation of context-sensitive embeddings. These embeddings provided a more nuanced representation of words offering a promising solution to many limitations of earlier topic models that relied solely on bag-of-words or shallow embedding techniques.

Leveraging the power of these transformer-based contextual representations, researchers began developing models that integrated advanced clustering methods. Among these innovations, BERTopic (Grootendorst, 2022) emerged as a notable state-of-the-art topic modeling approach. BERTopic operates by first generating dense document embeddings using transformer models, frequently employing variants such as Sentence-BERT (Reimers & Gurevych, 2019) or MiniLM (W. Wang et al., 2020), to capture rich contextual information. Subsequently, dimensionality reduction techniques like UMAP are applied to these high-dimensional embeddings to project them into a space more suitable for clustering. Clustering is then performed using density-based algorithms, typically HDBSCAN, which excels at discovering clusters of varying densities and effectively managing noise. A key innovation to BERTopic is its novel application of a class-based term frequency-inverse document frequency (c-TF-IDF) procedure. Unlike traditional TF-IDF that measures word importance at the individual document level, c-TF-IDF aggregates word importance across entire clusters or classes, significantly enhancing the interpretability and coherence of the extracted topics. This methodological advancement allows BERTopic to produce topic-word distributions that are both interpretable and accurately reflect the underlying semantics embedded within the transformer representations. By addressing issues like topic granularity and offering mechanisms for dynamic topic reduction, BERTopic represents a significant evolution beyond previous count-based, probabilistic, and neural topic models. Thus, to fully understand how BERTopic works we need to understand the three main components: UMAP, HDBSCAN and c-TF-IDF.

Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) (McInnes et al., 2017) extends the DBSCAN algorithm by discovering clusters at every density level and then selecting those that persist most stably, using only one user-defined parameter, *min_cluster_size*, which specifies the smallest allowable cluster. For each data point x in the dataset, HDBSCAN first computes the core distance $core_k(x)$, where k (often set equal to *min_cluster_size*) denotes the number of nearest neighbors. The $core_k(x)$ distance represents between x and its k-th nearest neighbor under the chosen metric $d(a, b)$ (e.g., Euclidean distance). The algorithm then defines the mutual reachability distance between any two points *a* and *b* as:

$$d_m reach, k(a, b) = maxcore_k\{(a), core_k(b), d(a, b)\}. \tag{3.1}$$

This formulation ensures that points in sparse regions cannot artificially connect dense clusters. At this point, HDBSCAN constructs a complete graph with all data points as vertices and mutual reachability distances as edge weights, then extracts the minimum spanning tree (MST) to form a hierarchical clustering tree (dendrogram) across every possible density threshold. This hierarchy is condensed by removing any cluster whose size falls below *min_cluster_size*, and for each remaining cluster C, one records its birth density $\lambda_{b\mathrm{irth}}$ (the threshold at which C first appears) and, for each member point p, its death density λ_p (the threshold at which p leaves C). The stability of C is then quantified as:

$$Stability(C) = \Sigma_{p\in C}\left(\lambda_{birth} - \lambda_p\right), \tag{3.2}$$

which measures how long each point remains in C as density changes. Finally, HDBSCAN selects a flat clustering by choosing non-overlapping clusters in descending order of stability, labeling any points not assigned to a selected cluster as noise (cluster -1) (Campello et al., 2013; McInnes et al., 2017).

Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2020) is a nonlinear dimensionality-reduction algorithm that models data as a fuzzy topological structure in high dimensions and then optimizes a low-dimensional embedding to best preserve that structure. Given a dataset x_i, $i = 1, \dots, N$, UMAP first computes, for each point x_i, a local connectivity distance ρ_i, defined as the distance from x_i to its closest neighbor. This ensure that every point has at least one "nearest neighbor" connection. The algorithm then selects a smooth kernel with σ_i such that the sum of high-dimensional membership strengths (fuzzy connection):

$$\mu_{i \to j} = exp\left(\frac{-\, max\, \{\, 0, d\left(x_i, x_j\right) - \rho_i \,\}}{\sigma_i} \right), \#(3.3)$$

over the k nearest neighbors of $x_i$equals log$_2$(k), where $d\left(x_i, x_j\right)$ is typically the Euclidean distance and k is the user-specified *n_neighbors* parameter. These directed memberships $\mu_{i \to j}$ are then symmetrized into an undirected fuzzy simplicial set through:

$$\mu_{ij} = \mu_{i \to j} + \mu_{j \to i} - \mu_{i \to j} \cdot \mu_{j \to i}, \tag{3.4}$$

defining the edge weight between x_i and x_j in the high-dimensional graph. In the low-dimensional embedding y_i, UMAP defines a parametric membership function

$$\nu_{ij} = \frac{1}{\left(1 + a \cdot \left|\left|y_i - y_j\right|\right|^{2b}\right)}, \tag{3.5}$$

where a and b are constants chosen to match the desired *min_dist setting*. Finally, UMAP optimizes the embedding by minimizing the cross-entropy loss:

$$C = \sum_{i \neq j} \left[\mu_{ij} log\left(\frac{\mu_{ij}}{\nu_{ij}}\right) + \left(1 - \mu_{ij}\right) log\left(\frac{\left(1 - \mu_{ij}\right)}{\left(1 - \nu_{ij}\right)}\right) \right], \tag{3.6}$$

via stochastic gradient descent. The resulting two- or three-dimensional representation balances preservation of local neighbor relations (through *n_neighbors* and ρ_i) and global manifold structure (through fuzzy set symmetrization and loss), with *min_dist* controlling the maximum tightness of point clusters in the embedding.

Class-based TF–IDF (c-TF-IDF) is an adaptation of the traditional TF–IDF weighting scheme that operates at the level of document clusters (classes) instead of individual documents, enabling clear identification of distinguishing terms for each topic discovered by clustering algorithms. In BERTopic, the classic TF–IDF weighting formula:

$$W_{t,d} = tf_{t,d} \times log\left(\frac{N}{df_t}\right) \#(3.7)$$

is generalized to operate over clusters (“classes”) of documents rather than individual documents. Each topic class *c* is treated as a single aggregated document by concatenating all its member texts, and the class-based TF–IDF weight is defined as:

$$W_{t,c} = tf_{t,c} \times log\left(1 + \frac{A}{\Sigma_{c'=1}^{C} tf_{t,c'}}\right). \#(3.8)$$

In this last formulation, $tf_{t,c}$ represents the raw count of term t within the aggregated text of class *c*, C is the total number of topic classes. $\Sigma_{c'=1}^{C} tf_{t,c'}$ is the total frequency of term t across all classes. The parameter A represents the average number of words per class (i.e., the mean length of these aggregated “class documents”). The “1 +” inside the logarithm ensures that all weights remain strictly positive. This c-TF-IDF formulation emphasizes terms that are both frequent within a given topic and comparatively rare across other topics, yielding coherent and distinctive keyword representations for each cluster (Grootendorst, 2022).

In article 2, BERTopic was leveraged for the needed an unsupervised method that could mine thousands of heterogeneous corporate-website snapshots and surface coherent, business-level themes without any labelled training data. BERTopic meets the goal because it (i) embeds each chunk with a transformer (“all-mpnet-base-v2”), (ii) groups semantically similar chunks via UMAP + HDBSCAN, and (iii) builds human-readable topic representations with class-based TF-IDF, retaining the full contextual signal that frequency-based models (LDA) lose. In article 2 a grid search was performed to set the hyperparameters (n_neighbors, n_components, cluster_selection_epsilon, min_cluster_size) to minimise the proportion of outlier chunks (label −1) while preserving topic coherence.

Supervised method

As validation method, Article 2 opted for a comparison among Random Forest (RF), Extreme Gradient Boosting (XGBoost), and a Multilayer Perceptron (MLP) to establish which topics are associated with the companies that receive private fundings.

RF (Breiman, 2001) is an ensemble method that combines three key techniques to produce robust predictions. First, it uses bootstrap sampling (bagging) to train multiple decision trees, each on a different random sample of the training data drawn with replacement. Second, at each node during tree construction, only a random subset of features is considered for splitting, which reduces correlation between trees while maintaining predictive strength. Third, predictions are aggregated through majority voting (classification) or averaging (regression). This ensemble approach effectively reduces variance and prevents overfitting. RF also provides out-of-bag (OOB) estimates from data points excluded from each tree's bootstrap sample, offering internal measures of generalization error and variable importance without requiring a separate validation set. A key advantage is that generalization error converges as more trees are added, providing natural protection against overfitting.

XGBoost (T. Chen & Guestrin, 2016) represents an advanced gradient boosting framework that builds trees sequentially. Each new tree is trained to correct the residual errors of its predecessors, with additional regularization terms included in the objective function to control model complexity and overfitting (Grinsztajn et al., 2022). The quality of a tree structure is evaluated using a scoring function derived from this objective. To further prevent overfitting, XGBoost utilizes shrinkage, which scales the weights of newly added trees, and column (feature) subsampling. Key algorithmic innovations contribute to XGBoost's performance. XGBoost incorporates a novel algorithm that handles sparse data by assigning a default direction in each tree node for missing values or zero entries. This method only visits non-missing entries, making computation complexity linear to the number of non-missing entries and significantly faster than naive implementations. Moreover, when building decision trees on large datasets, XGBoost does not check every possible way to split the data. Instead, it uses an efficient algorithm (weighted quantile sketch) to select the most promising split points. The algorithm assigns weights to data points based on how much they affect the loss function (using second-order gradient information) and ensures the split points are chosen in a balanced way, even across distributed systems.

An MLP (LeCun et al., 2015) is a fully connected, feed-forward neural network composed of an input layer, one or more hidden layers with non-linear activations, and an output layer. MLP uncovers complex structures in large datasets using the backpropagation algorithm, which uses the chain rule to efficiently compute gradients across multiple layers. During training, the MLP processes inputs to produce class probability scores, and an objective function measures prediction error. Then, weights are updated using stochastic gradient descent (SGD), typically processing small batches of training examples, a common optimization procedure where weights are adjusted based on the average gradient computed from small batches of training examples. The backpropagation algorithm is a practical application of the chain rule for derivatives, allowing the computation of these gradients efficiently through the multiple layers of the network. Modern deep networks often use non-linear activation functions like the rectified linear unit (ReLU), as it typically allows for faster learning in networks with many layers compared to older functions like tanh or sigmoid.

Article 2 leverages the capacity of these methods to model more than 100 variables when predicting the dependent variable. These three algorithms provide complementary approaches to the classification task: RF offers interpretability and robustness, XGBoost provides high performance through sequential error correction, and MLP enables complex pattern recognition through deep learning. This diverse algorithmic comparison ensures comprehensive evaluation of the relationship between topics and private funding. Regarding model evaluation, Article 2 provides a comprehensive explanation of the evaluation methodology and the rationale behind the selected approaches (see p.).

Imbalance Correction

To improve the performances of our supervised method, we perform an imbalance correction. It has been widely reported that the class imbalance heavily compromises the process of learning, because the model tends to focus on the prevalent class and to ignore the rare events (Japkowicz & Stephen, 2002). Moreover, it is not clear which tool to use in many cases, (Wasikowski & Chen, 2010; Menardi & Torelli, 2014), and only heuristic reasons justify the decision of one algorithm or another. Although we attempted to mitigate the imbalance issue at an algorithmic level, our best results were obtained by combining hyperparameter selection with data imbalance correction. In Article 2, the models' performances are compared after using the following imbalance correction algorithms: ROSE, SMOTE, ADASYN, K-means SMOTE, and SVM-SMOTE.

Random Over-Sampling Examples (ROSE) (Menardi & Torelli, 2014) tackles the imbalance problem issue by generating new synthetic examples for both minority and majority classes using a smoothed bootstrap method based on kernel density estimation. The algorithm starts by picking a data point from either the minority or majority class in the training set. Once it has chosen this data point, it does not just make a simple copy, but it creates a new, slightly altered version of that point by adding a small amount of “random noise.” This noise is drawn from a probability distribution, often a normal (Gaussian) distribution, that is centered exactly at the chosen data point. The amount of variation or “spread” in this noise is controlled by what is called a smoothing parameter.

Building on the concept of synthetic generation, Synthetic Minority Over-sampling Technique (SMOTE) (Chawla et al., 2002) focuses on generating additional minority class samples rather than merely duplicating existing examples. This synthetic generation is performed by interpolating new examples along the lines connecting each minority class instance to its nearest neighbors avoiding overfitting, typically caused by exact replication.

Adaptive Synthetic Sampling Approach for Imbalanced Learning (ADASYN) (He et al., 2008) extends SMOTE’s interpolation approach by incorporating adaptive allocation strategies. It adaptively allocates more synthetic data to minority samples that are harder to classify, as identified by their proximity to majority-class instances. This adaptive strategy helps in effectively shifting the decision boundary towards challenging minority examples. The ADASYN process involves evaluating the degree of class imbalance, calculating the required synthetic examples, identifying difficult minority cases using nearest neighbor analysis, and then generating new samples strategically around these difficult instances. ADASYN’s particular strength is enhancing minority-class prediction without significantly compromising majority-class accuracy.

While ADASYN addresses boundary challenges, the “Oversampling for Imbalanced Learning Based on K-Means and SMOTE” (Last et al., 2017) (K-means SMOTE) combines the K-means clustering algorithm with the Synthetic Minority Over-sampling Technique (SMOTE) overcoming the noise generation problem inherent to traditional SMOTE. Unlike traditional SMOTE, which can randomly amplify noise and create redundant data points, K-means SMOTE strategically targets safe, minority-dominated clusters to generate synthetic data, thereby reducing noise generation. Additionally, it addresses within-class imbalances by allocating more synthetic samples to sparse minority regions identified through clustering. The algorithm operates in three distinct phases. First, the entire dataset is clustered via K-means. Second, clusters predominantly consisting of minority class samples are selected, especially focusing on those with lower density of minority instances. Third, SMOTE is applied within these selected clusters to synthetically create minority examples, effectively addressing both inter-class and intra-class imbalances. The proposed method notably includes the behavior of standard SMOTE and random oversampling as special limiting cases, thereby ensuring at least equivalent performance.

Finally, the “Borderline Over-sampling for Imbalanced Data Classification” (Nguyen et al., 2011) or the SVMSMOTE[14] focuses specifically on the critical boundary area between minority and majority classes. The motivation is that the borderline instances, particularly the support vectors identified through a Support Vector Machine (SVM), are crucial for determining the decision boundary. Thus, the method generates synthetic minority class samples by using interpolation, extrapolation or both techniques. Extrapolation is used particularly in regions where the density of majority class samples is low, allowing the minority class to expand into areas previously dominated by the majority class. This approach improves the model’s ability to recognize the true decision boundary. Interpolation (as in classic SMOTE) is applied in areas where majority class instances are more densely clustered, consolidating the minority class along the existing boundary. By leveraging the SVM decision boundary and adjusting the sampling strategy based on local class densities, SVM SMOTE enhances the classifier’s ability to distinguish between classes in imbalanced datasets.

[14] (as called in https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SVMSMOTE.html)

CHAPTER 4 ARTICLE 1: First impressions on sustainable innovation matter: Using NLP to replicate B-Lab environmental index by analyzing companies' homepages

Pietro Cruciata, Davide Pulizzotto and Catherine Beaudry

Published in the Journal of Technological Forecasting and Social Change (May 15th 2024)

Keywords: Zero-shot text classification, B-Corp data, sustainability, sustainable innovation, natural language processing, signal theory

Abstract

This study explores the potential for developing web-based environmental culture indicators by analyzing signals extracted from the homepages of company websites. The primary aim is to assess the proposed method's ability to generate indicators that can serve as proxies for real environmental measures by leveraging the homepage content. We performed a Zero-Shot Text Classification (ZSTC) using a BERT-type Natural Language Processing (NLP) model, followed by a regression analysis to test the ability of these web-based indicators to replicate the B-Lab environmental index and comprehend the dynamics behind the results. This pilot study explains 57% of the variance of the B-Lab environmental index using the results of the ZSTC score and companies' characteristics. This research makes two significant contributions. First, the text content of a company's homepage seems to provide insights into its environmental performance. Second, it introduces a generalizable methodology for studying the performance of companies through their websites without the need for heavy pre-processing, significantly reducing the time and cost of research. Furthermore, the method could provide policymakers with a real-time landscape to create and finetune policies about specific topics, partially addressing the problems associated with questionnaire-based surveys.

4.1 Introduction

Traditional innovation indicators built using public databases generally supplemented by questionnaire-based data are important sources of information for governments, academics, and the private sector. These sources of information are often incomplete (e.g., representative samples much smaller than the population of firms) or not specific, whereas questionnaire-based surveys (especially large-scale as the biennial European CIS or the annual MIP) lack regional granularity, coverage, timeliness, and more importantly, they are costly to run (Axenbeck & Breithaupt, 2021). Moreover, the number of low-cost web-based surveys sent to firms has sky-rocketed to the extent that obtaining a representative response rate has plummeted to lower than 5-10% in most cases. For all these reasons, innovation indicators built using traditionally collected data hardly provide the full picture (Kinne & Lenz, 2021b).

Alternative or complementary to these sources are web-based unstructured textual data. The increasing amount of data available in the form of digitalized text indeed offers new avenues for innovation studies. Among their noteworthy advantages, the rapidity of their evolution, their increasing quantity, variety, and availability opened new possibilities for policymakers and researchers (Gök et al., 2015). Although it seems difficult to measure and interpret “signals” of innovation dynamics in corporate websites or other web sources, researchers in innovation and technology management have obtained good results by building new indicators with large amounts of text. For example, Gök et al. (2015) created web indicators of R&D activities by extracting the keywords from companies’ websites. Their study suggested that R&D activities captured through the web indicators were significantly more numerous, compared to the R&D activities documented in other sources. Libaers et al. (2016) harnessed the data from companies’ websites to develop a taxonomy that identified strategies used by small firms to commercialize their innovations. The authors analyzed the content of firms’ websites to extract the keywords related to possible strategies used by companies. Blazquez and Domenech (2018) used web-based variables built with keywords to predict firm export orientation. Héroux-Vaillancourt et al. (2020) built innovation indicators based on four core concepts (R&D, IP protection, collaboration, and external financing) from the complete texts of corporate websites of Canadian nanotechnology and advanced materials firms using keywords frequency analysis. Other researchers specifically studied different dimensions of sustainable development in companies through their environmental performances using their websites. For instance, Fernández-Vázquez and Sancho-Rodríguez (2020) analyzed texts and images from the websites of the Spanish IBEX 35 to investigate to which extent the companies address climate change in the construction of their reputational identity and to explore the types of narratives. Calabrese et al. (2021) examined the websites of 23 manufacturers from the fast-growing fitness equipment industry to study firms' strategies and their contribution to the SDGs.

All these pilot studies highlight the strong potential that these new sources of data bring to the field of innovation studies. Building on these encouraging findings, can we develop web-based environmental indices that mirror a real environmental index by analyzing the content of companies' homepages?

Companies' communications are divided into two main channels: external, towards clients and stakeholders; and internal, towards the workers who are part of the company. Through internal communication, a company expects to generate know-how necessary to fuel operational procedures, as well as the loyalty of employees, which motivates them to apply their expertise to the company's processes (Mazzei, 2010). External communication revolves around the company's relational network, serving to provide vital information to the business intelligence system, influence project specifications, facilitate industrial and financial package development, and foster trust with clients and partners (Goczol & Scoubeau, 2003). Therefore, an official website serves as a platform for conveying authentic, precise, and current information about companies, enabling visitors to make more informed decisions (Jiang et al., 2023). As a result, the information contained in a website provide a general understanding of the relevance of a particular element for the company (Héroux-Vaillancourt et al., 2020).

As policymakers shift their focus to adapt to climate change, mitigating its effects, and striving for a more positive socio-environmental impact through their policies, sustainable innovation (SI) emerges as a key solution. This paper examines the potential of developing web-based environmental culture indicators that analyze signals gleaned from the homepage of companies' websites. The primary objective is to explore the proposed method' capacity to create indicators as proxies for real environmental measures. This pilot study focuses on the environmental index of the B-Corp database, to evaluate the approach. It is one of 5 indices developed by B-Lab to assess various Environmental, Social, and Governance (ESG) dimensions of companies. The B-Corp certification has gained recognition in helping organizations stand out in the 'green revolution' (Kim & Schifeling, 2016, p. 32), establishing legitimacy (Blasi & Sedita, 2022; Cormier &

Magnan, 2015), and projecting an authentic commitment to triple bottom line (TBL) practices (Cao et al., 2017; S. Kim & Schifeling, 2016). The database is therefore particularly well suited for our purposes: measuring the correlation between a company's website and the B-Lab indicator is the main goal for this pilot study.

The methodology comprises two steps: first, a Zero-Shot Text Classification (ZSTC) score is obtained using a BERT-type Natural Language Processing (NLP) model to extrapolate and study the environmental signal of each company's website; second a regression model is estimated to evaluate to what extent these web-based environmental culture indicators (the signal) explain the value of the environmental index attributed by B-Lab to the company. The results of the ZSTC score together with the companies' characteristics explain 57% of the variance[15] of the B-Lab environmental index obtained by companies, thereby showing great promise for the proposed method.

The remainder of the paper is organized as follows. Section 2 presents the pertinent literature on sustainable innovation and signal theory. Section 3 describes the data collected and explains the methodology. Section 4 analyzes the results of the ZSTC, the correlation between the ZSTC scores and the B-Lab environmental index, and the Principal Component Analysis. Section 5 presents the Ordinary Least Squares (OLS) regression results while Section 6 discusses their implications. Finally, Section 7 concludes and highlights the limitations of the research and possible future works.

4.2 Literature review

The widely accepted viewpoint that innovation is driven solely by the combination of scientific research, technological advancements, their implementation by businesses, and distribution in the market, has evolved considerably. Innovation is no longer solely about enhancing market competitiveness and advancing technology in various industries. Instead, it is increasingly seen as a means to address social issues, improve quality of life, and enhance overall societal and environmental health. For instance, policymakers are now working to define and support the concept of SI, among other ideas linked to environmental, social, and governance (ESG)

[15] This is simply measured by the R-squared value of the regressions.

considerations. The origin of the SI concept can be dated back to the publication of the "Brundtland Report" (WCED, 1987), in which the World Commission on Environment and Development (WCED) coined the term *Sustainable Development*, which the report defined as "development that meets the requirements of the present without jeopardizing future generations' ability to meet their own needs" (Zhu & Hua, 2017, p. 893). Over time, governments have gradually placed greater emphasis on reducing the environmental footprint of economic activities. It was previously believed that economic objectives and environmental concerns were incompatible, but this notion was challenged by Weale's paper (1992). Moreover, the "triple bottom line" concept introduced by John Elkington in the 1990s has become the cornerstone of sustainable development. This concept seeks to harmonize environmental, economic, and social performance – a challenge that businesses must now address (Bossle et al., 2016).

Furthermore, the belief that companies can simultaneously pursue economic, environmental, and social goals has been reinforced by shifts in customer demands and stakeholder requirements. These changes are exerting increasing pressure on companies to implement sustainable initiatives and to measure, monitor, and report on sustainability performance. Customers and stakeholders are showing a growing interest in sustainable brands, with ethical and sustainable certification becoming a crucial factor that consumers consider when making purchasing decisions. Additionally, studies have shown the growth of B-Corp businesses after obtaining certification (Romi et al., 2018; Paelman et al., 2020). This underscores the importance of external communication in attracting customers.

In this context, several studies within the field of signal theory shed light on how companies strategically use their official websites to shape stakeholders' perceptions (e.g., Mavlanova et al., 2012; Yildiz et al., 2023). Signal theory defines a "signal" as an action initiated by a better-informed party in situations characterized by information asymmetry. The purpose of this signal is to effectively and credibly communicate the party's true characteristics to a less-informed counterpart (Connelly et al., 2011). Scholars in management have leveraged signal theory to elucidate the impact of information asymmetry across a variety of research domains. For instance, Mavlanova et al. (2012) conducted a study on the role of website signals as a means for online retailers to communicate their product quality, proposing and validating a three-dimensional framework. Jiang et al. (2023) argued that a corporate official website serves as a credible source of non-financial information for assessing the credit risk of Small and Medium Enterprises (SMEs).

SMEs equipped with comprehensive official website information are less likely to default and are better positioned to secure financial support for further development. Yildiz et al. (2023) found that the presence of a "green label" on a hotel enhances the trustworthiness of the eco-conscious tourist brand. Lastly, Eccles et al. (2014) concluded in their research that companies with a strong emphasis on sustainability demonstrate increased levels of information transparency and accountability. Based on these findings, we put forth the following proposition:

Proposition 1: Environmental compliance indices are positively correlated with the environmental culture indices built using the text contained in companies' websites.

Several studies have highlighted the impact of both internal and external factors on companies' sustainability efforts (Hermundsdottir & Aspelund, 2021b). In addition to the pressure from stakeholders and customers, national regulations, incentives, society's awareness, industrial norms, and regulations are some of the several factors that might directly impact companies in their pursuit of sustainable initiatives (Hermundsdottir & Aspelund, 2021b). Doran and Ryan (2012) discovered that regulation and industrial agreements significantly influence a firm's decision to engage in eco-innovation. Aguilera-Caracuel and Ortiz-de-Mandojana (2013) proposed that policymakers play a crucial role in a firm's ability to transform SI into competitive advantages. Moreover, the authors suggested that countries with stricter environmental regulations tend to have a higher prevalence of green innovative firms. Finally, de Azevedo Rezende et al. (2019) demonstrated that there are differences in green innovation performances between Europe and North America due to their distinct approaches to regulations. Given the variations in performance and in the willingness to pursue SI highlighted in the literature, we suggest the following proposition:

Proposition 2: The country in which a company is located influences its environmental compliance index.

Additionally, Magnusson et al. (2011) suggested that the reputation of a brand's country of origin serves as a conspicuous and consistent signal that can shape consumer perceptions of corporate brand reputation. On the other hand, corporate brands originating from countries with more favorable sustainability reputations may not experience the same benefits from engaging in corporate social responsibility (CSR) or sustainability efforts, as these reputation-building strategies may be expected (Cowan & Guzman, 2020). Thus, we posit the following proposition:

Proposition 2m: The relationship between a company's web-based environmental culture indicators and its environmental compliance index is moderated by the country in which it is located.

It is reasonable to assume that a company's size may also affect its propensity to pursue sustainable initiatives. Aguilar-Fernández & Otegi-Olaso (2018) investigated the impact of size on the likelihood of firms to pursue SI, concluding that there is no consensus on this impact. On the one hand, while large companies may have advantages in pursuing SI from both a supply chain and financial perspective. SMEs, on the other hand, may have more flexibility to adapt and change their business models. Additionally, large companies may face more pressure from stakeholders to achieve socio-environmental goals due to their greater exposure. On the other hand, the lack of resources and capacities can be a limit for SMEs (Aguilar-Fernández & Otegi-Olaso, 2018). Ketata et al. (2015) highlighted the positive impact of firm size in their study on SI in Germany. De Azevedo Rezende et al. (2019) also identified differences in green innovation performance according to company size in their analysis. The interplay between a company's size and its digital communication strategy has been a focal point of various studies. For instance, Kinne & Axenbeck (2020) found a correlation between the size of a company and the number of pages of its website. In a complementary vein, Callison (2003) posits that companies positioned higher in the rankings often possess greater financial and professional resources, which they can leverage to enhance their web presence. This perspective is further corroborated by the research of Jung Moon & Hyun (2014), who observed that large firms tend to have robust marketing teams dedicated to the upkeep of their websites. In light of the evidence presented, our proposition is as follows:

Proposition 3: The size of a company influences its environmental compliance index.

Furthermore, Hoehn-Weiss and Karim (2014) shed light on the advantages that young firms gain when they signal alliances with larger partners. This strategy can attract the general market and make an Initial Public Offering (IPO) a more appealing option than an acquisition. Given this intriguing finding on the signaling of small companies, we posit the following proposition:

Proposition 3m: The relationship between a company's web-based environmental culture indicators and its environmental compliance index is moderated by its size.

As previously mentioned, it is commonly understood that economic benefits drive companies to adopt sustainable behavior, and the sector in which a company operates is also significant. De

Azevedo Rezende et al. (2019) demonstrated differences in green innovation performance between manufacturing and non-manufacturing companies. The latter face more challenges in implementing green technologies in sectors such as services or information, while for manufacturing companies, green innovation can attract clients or increase efficiency, thereby generating a direct economic impact. Hermundsdottir and Aspelund (2021) highlighted how some industries adopt sustainable practices as standard, while others respond differently to environmental obstacles to SI. For these reasons, it is commonly asserted that the impact of various factors on the development of environmentally friendly products and processes varies depending on the industrial sector under examination. Numerous research endeavors have acknowledged that their conclusions are applicable solely within a specific industry, and have specified that their outcomes are limited to the context of that industry (Tariq et al., 2017). For the reasons mentioned above, the following proposition is suggested:

Proposition 4: The industrial sector in which a company operates influences its environmental compliance index.

Moreover, as previously mentioned, Yildiz et al. (2023) provided empirical evidence of how environmental efforts signaled by hotels are advantageous for eco-tourists. The trust instilled by the “green label” significantly mediates the perceived green risk in online booking intentions within the hotel sector. This research exemplifies how a company’s signaling approach is personalized and dependent on the specific needs of the sector in which the company operates. In line with this finding, we formulate the following proposition:

Proposition 4m: The relationship between a company’s web-based environmental culture indicators and its environmental compliance index is moderated by the industrial sector in which it operates.

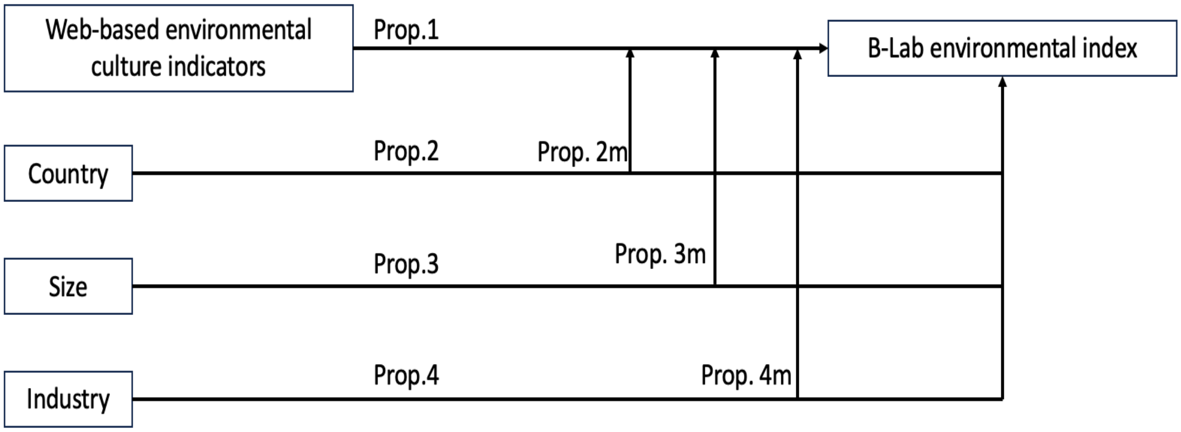
Web-based environmental culture indicators
Prop.1
B-Lab environmental index
Country
Prop.2
Prop. 2m
Size
Prop.3
Prop. 3m
Industry
Prop.4
Prop. 4m

Figure 4.1 Summary of the propositions to be tested.

4.3 Data and Methodology

4.3.1 Data

Two types of data are required to assess whether new web-based environmental culture indicators are good proxies for more traditionally built environmental compliance or certification indicators. For the latter, we selected the certification of the B-Corporation, henceforth referred to as B-Corp, which is a type of for-profit corporation that has been certified by the non-profit organization B-Lab to meet certain standards of social and environmental performance, accountability, and transparency. These standards are set by the B-Lab and are verified through a rigorous assessment process. The B-Corp Certification is comprehensive and adopts a holistic approach to environmental, social, and governance (ESG) issues. Furthermore, obtaining and maintaining accreditation is a rigorous procedure that involves teams and departments from across the organisation. B-Corp firms are committed to making a positive impact on society and the environment, and to conducting in a way that is transparent, accountable, and sustainable. Unlike traditional corporations, B-Corps are required to consider the impact of their decisions on their employees, customers, suppliers, community, and the environment.

B-Corp data includes one main index, “overall score”, which is an aggregation of five other indices evaluating specific dimensions: governance, customers, workers, community, and environment. These dimensions are further divided into several items. In this paper, we focus on the B-Corp indicator concerning the “impact area environment”. We refer to this environmental compliance index variable “B-Lab environmental Index” (*BLabEnvIndex*). The B-Lab Environmental Index is derived from the B Impact Assessment, a tool that assesses a company’s social and environmental performance. Specifically, the B-Lab environmental Index evaluates a company’s overall environmental stewardship. This includes how the company manages general environmental impacts, air and climate issues, water sustainability, and impacts on land and life. The scoring system used in the assessment allows for comparability across companies and identifies areas for improvement over time. The scoring criteria are customized and evolve with each version of the assessment, based on the specific track of the company being evaluated. We incorporated control variables related to two other specific ESG areas, community and governance, because the minimum score required to pass the assessment is determined by the sum of the scores for the aforementioned five ESG areas. This means that a company could potentially invest more in other areas than the environmental one and still qualify as a B-Corp (Liute & De Giacomo, 2022). Additionally, we introduced a dummy variable representing the assessment year—the year in which the company completed the B-Lab test designed to measure, manage, and enhance positive impact performance for the environment, communities, customers, suppliers, employees, and shareholders. The control variables related to the assessment year account for changes that can affect the assessment test[16].

This pilot study uses only a subset of the B-Corp data limited to Canadian and US companies. The reason for this choice is straightforward. We aimed to ensure a certain degree of homogeneity in the sample considered, i.e., that the results are not affected by widely different national systems, or languages, or by considering several dimensions simultaneously. As of March 2022, B-Corp had 8,799 certifications from 5,631 companies (the certification lasts 3 years) across 86 countries.

Figure 4.2[17] illustrates the primary steps of our data collection process. As mentioned earlier, we began with 8,799 certifications of companies[18], narrowing down our selection to 1,741 certified companies in Canada and the USA. For textual data collection, we utilized the website URLs

[16]See https://kb.bimpactassessment.net/support/solutions/articles/43000547789-overview-of-changes-in-version-6-of-the-b-impact-assessment, accessed on 26th Nov. 2024.

[17]To assess the comparability between the 1,256 companies with a snapshot in the online archive and the 485 excluded companies, we conducted a non-parametric Mann-Whitney Anova test. The results revealed no significant differences in averages between the two groups. Furthermore, we performed the same test between the final sample of 1,110 companies and the 631 eliminated companies (due to no snapshot found and specific criteria such as no more than a sentence, non-English language, and manual elimination for lack of meaningful content).

[18] Some companies have been certified multiple times.

provided by B-Corp, leveraging the Wayback Machine, a web archive tool. This allowed us to obtain snapshots of company homepages corresponding to the certification years between 2007 and 2022 for each Canadian or US company in the B-Corp data. The objective was to collect data from the company's website for the year of the assessment recorded in the B-Corp database. In instances where a company's snapshot for the specific year was unavailable, we gathered data within a three-year range, encompassing the target year as well as the preceding and following years. We extracted the text from the homepages using specific HTML tags (i.e., <p>, <li>, <h1>, <h2>, etc.), effectively matching the companies with their respective snapshots and reducing the sample to 1,256. Subsequently, we filtered out snapshots deemed irrelevant based on the following criteria: non-English websites[19], instances with less than one sentence of text, and manual removal of snapshots resulting from errors in the Wayback Machine. Our final sample comprises 1,110 companies, with 195 of them being Canadian firms.

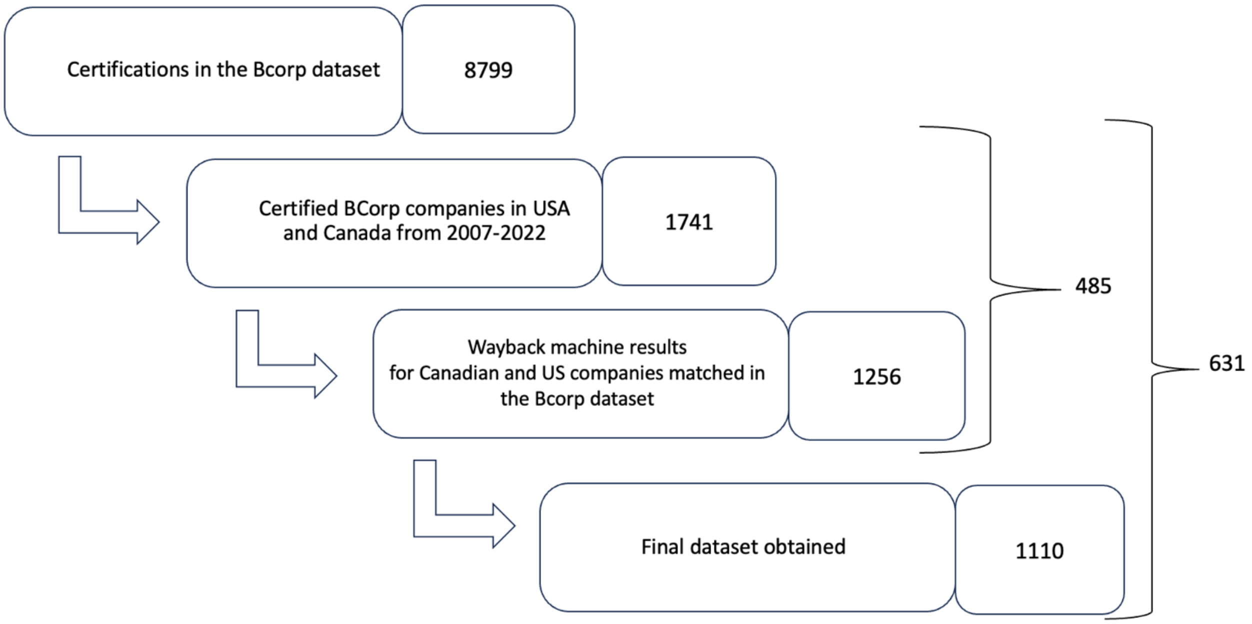
Certifications in the Bcorp dataset
8799
Certified BCorp companies in USA and Canada from 2007-2022
1741
Wayback machine results for Canadian and US companies matched in the Bcorp dataset
1256
Final dataset obtained
1110
485
631

Figure 4.2 Pre-processing steps

4.3.2 Methodology

Once the data is prepared, the first step of the analysis involves "understanding" the text of the corporate websites. Instead of counting specific keywords related to predetermined topics, as most of the literature mentioned in the introduction does, we employ the Zero-Shot Text Classification

[19] Our methodology will use a NLP model that is only trained in English document.

(ZSTC) method. This method is a Natural Language Processing (NLP) task designed to answer the question: “Is this text about label X?”. The response to this question serves as an indicator of the confidence that the given text pertains to the label X. The 31 labels used for this purpose correspond to the names of the 31 items that compose the B-Corp environmental certification.

Within the realm of the Natural Language Understanding (NLU), ZSTC is a challenging task that necessitates the use of syntactic and semantic analysis to comprehend the actual meaning and sentiment of human language. More specifically, ZSTC refers to a task where the model classifies text into classes that were not present in the training corpus. In other words, ZSTC aims to associate an appropriate label with a piece of text, regardless of the text’s domain or the aspects specified by the label. ZSTC was initially applied in a Dataless Classification scenario, similar to the problem we are currently addressing, where it was used to select the appropriate label for a text through Explicit Semantic Analysis. With the rise of word embeddings, various approaches have been proposed for this purpose. For instance, generative Long Short-Term Memory has been used to generate text given the vector labels, and the vector representation of the label has been used to represent the text in multilabel classifiers (Yin et al., 2019).

The core of ZSTC is the NLP model “Bidirectional and Auto-Regressive Transformers” or BART (M. Lewis et al., 2019), a transformer-based deep learning model for NLP developed by Facebook AI. This model combines the most significant characteristics of BERT[20] and GPT[21]. BART was pre-trained on the English Wikipedia and BooksCorpus, using a two-step processes: first, the text is altered by adding a noise factor (e.g., changing the words randomly); then, the model learns to reconstruct the original text. This innovative approach allows BART to reach state-of-the-art performances in several NLP challenges. Indeed, BART excels in text generation, but it has also been tested in a wide range of tasks, including discriminative tasks such as General Language Understanding and the Stanford Question Answering Dataset (M. Lewis et al., 2019).

Performing the ZSTC requires the selection of both the labels and the corpus. Since our aim is to create an environmental culture indicator, we utilized the labels of the items that constitute the “impact area environment” index of the B-Corp data (Table 1). After experimenting with several

[20] Bidirectional Encoder Representations from Transformers (Devlin et al., 2018).

[21] Generative Pre-Training (Radford et al., 2018).

settings, we decided to divide each website into groups of three sentences to create the corpus[22]. This decision was based on the observation that the ZSTC performed better when the input was a text longer than a single sentence but shorter than the full homepage. Consequently, for each website, we applied the ZSTC on each group of sentences. Then, to prepare the results for the subsequent Pearson correlation test, we calculated the average ZSTC scores for each label listed in Table 4.1, for each website.

Table 4.1 Labels used in the Zero-Shot Text Classification

- Air climate
- Certification
- Community
- Construction practices
- Designed to conserve agriculture process
- Designed to conserve manufacturing process
- Designed to conserve wholesale process
- Energy water efficiency
- Environment products services introduction
- Environmental education information
- Environmental management
- Environmentally innovative agricultural process
- Environmentally innovative manufacturing process
- Environmentally innovative wholesale process
- Green investing
- Green lending
- Inputs
- Land life
- Land office plant
- Land wildlife conservation
- Material energy use
- Materials codes
- Outputs
- Renewable energy
- Cleaner burning energy
- Resource conservation
- Safety
- Toxin reduction remediation
- Training collaboration
- Transportation distribution suppliers
- Water

Source: https://data.world/blab/b-corp-impact-data/workspace/data-dictionary

It is crucial to note that we configured the multilabel output for the ZSTC. When the output is multilabel, the ZSTC generates a score among the class using cosine similarity metrics between the word-embedding vectors created by BART, representing the label and the target corpus. The cosine similarity is a metric in a range from -1 to 1, where a value of 1 signifies identical vectors, 0 indicates they are orthogonal (i.e., completely dissimilar), and -1 implies they are diametrically opposed. Generally, the closer the cosine similarity is to 1, the more similar the vectors are to each other. In this task, the corpus and the label are transformed into vectors using BART's word embedding representation, and then the cosine similarity is calculated. Consequently, when we ask the model "Is this text about label X?" The score within the 0 to 1 range can be identical for multiple labels.

To examine the correlation between the web-based environmental culture indicators and B-Lab environmental index (the environmental compliance index of the propositions), we calculate Pearson correlations. This calculation measures which different items of the B-Corp environmental

[22] We use the python package spacy for this purpose.

certification, detected with the ZSTC in the text of the corporate websites, serves as good proxies for the environmental score obtained by these firms. From the list of labels in Table 4.1, we expect the ZSTC to generate a series of indicators that are likely to correlated to various extents. From this point, we want to comprehend the different dynamics behind the indicator produced by exploiting the B-Corp subset. Therefore, we leverage these correlations by performing a Principal Component Analysis (PCA). This analysis serves two purposes: firstly, it reduces the 31 indicators (for each of the 31 labels) to a manageable number; and secondly, it constructs a series of aggregated web-based environmental culture indicators that are orthogonal to one another.

Finally, the concluding step of our analysis, which aims to verify our propositions, involves estimating a series of Ordinary Least Squares (OLS) regressions. The goal is to estimate the proportion of the variance of the B-Lab environmental index that can be explained using our web-based environmental culture indicators. The structure of the regression model to be estimated is as follows:

$$Y_i = \alpha + \rho CommInd + \tau GovernInd + \sum_{k}^{K} \eta_k dyear_k + \sum_{j}^{J} \beta_j pca_{ij} + \varepsilon_i. \qquad (4.1)$$

$$\begin{aligned} Y_i = \alpha + \rho CommInd + \tau GovernInd + \sum_{k}^{K} \eta_k dyear_k + \sum_{j}^{J} \beta_j pca_{ij} + \gamma dCanada_i \\ + \sum_{l}^{L} \delta_{1l}\, dsize_{il} + \sum_{m}^{M} \theta_{\mathfrak{J}}\, dIndustry_{\mathfrak{J}} + \varepsilon_i \ \#(4.2) \end{aligned}$$

where Y_i is the dependent variable B-Lab environmental index of firm i that we are trying to predict, with $i \in N, 1 \leq i \leq 1{,}110$, $CommInd \wedge GovernInd$ are the two control variables representing respectively the log of the B-Lab impact on the community indicator and the B-Lab impact on the governance indicator, the variable *dyear* with $k \in \{2015 \wedge 2019, 2016, 2017, 2018, 2020, 2021\}$ represents the dummy variables related to the year of the assessment test filed by the company[23]. The variables pca_{ij} represent the results of the factor analysis, with $j \in N, 1 \leq j \leq 6$, where 6 is the number of the factors, $\alpha, \rho, \tau, \beta, \delta, \theta \in R$, are the coefficients in the regression model, and $\varepsilon \in$

[23] Upon realizing that there was only one observation in *d2015* and noting a correlation of -0.39 between *d2019* and d2018 in the correlation matrix, we decided to omit both the two assessment years, d2015 and *d2019*.

R indicates the error term of the regression. The dummy variable $dCanada_i$ takes the value 1 if the firm is located in Canada, and 0 otherwise (i.e., it is an American company). The dummy variables, il represent the company sizes $l \in N, 1 \leq l \leq 3$[24]. The industry dummy variables,$industry_m$represent each industry category $m \in N, 1 \leq m \leq 9$[25].

4.4 Web-based variable construction

4.4.1 Zero-Shot Text Classification

Table 4.2 presents the results of the initial step of our analysis. For all the 1,110 websites, the ZSTC provides the average score, which ranges from 0 to 1, for each of the 31 web-based environmental labels. Considering their mean score, "designed to conserve wholesale process" and "inputs" are the labels with the highest average. In other words, in the full text of the websites, there are, on average, more groups of sentences that, according to the model, refer to these labels. While all variables have a minimum score close to 0, the maximum value varies across the labels. Out of the 31 labels, 17 have a maximum score above 0.90. This indicates that, according to the model, each of these labels was the dominant one on at least one website. Conversely, the labels "clear burning energy", "green lending", "community", "designed to conserve manufacturing process" and "land wildlife conservation" display the lowest maximum values in the table, with the latter showing a maximum score of less than 0.5. This suggests that generally, our sample does not include websites where the "land wildlife conservation" label has a ZSTC score higher than the other labels. Moreover, "land wildlife conservation", "environmentally innovative manufacturing process" and "green investing" all exhibit a low average score. This implies that the model infrequently identifies groups of sentences that correspond to these labels.

Generally, the minimum score value is closer to the mean than to the maximum value, except for the labels "designed to conserve wholesale process" and "inputs". In other words, while the other

[24] We verified that there were no significant differences between the *size_0* (no employees) and the *size_1_9* (1 to 9 employees) on the dependent variable using the Mann-Whitney test before merging the two B-Lab classifications into *dmicro* (0 to 9 employees). Likewise, we created the dummy variable *dlarge* by merging the *size_250_999* (250 to 999 employees) and *size_1000+* (more than 1,000 employees). Finally, we obtained 4 size categories and *dmicro* is the omitted size dummy in the regression analysis.

[25] We repeated the same Mann-Whitney tests for the industrial classifications provided by B-Lab. The results allowed to merge 4 of the B-Lab industry categories: Media with Restaurant, Hospitality & Travel; Legal Services with Finance Services; and Retail with Transportation & Logistics. Finally, we obtained 10 industry categories and *dConsPdct* is the omitted industry category in the regression analysis.

labels are found on several websites with low scores or are not found at all, the model considers the labels “designed to conserve wholesale process” and “inputs” only when their value is significantly higher than the other scores. Additionally, the scores displayed in Table 4.2 do not follow a normal distribution: the mean and median are not equal and for most of the scores, the third quartile is closer to the minimum value, suggesting possible outliers. Consequently, we perform several transformations of the web-based indices before proceeding with the Pearson Correlation test.

Table 4.2 Zero-Shot Text Classification (ZSTC) results

Labels	**mean**	**std. dev.**	**min**	**25%**	**50%**	**75%**	**max**
Inputs	0.435	0.107	0.020	0.375	0.445	0.505	0.793
Outputs	0.103	0.104	0.000	0.046	0.076	0.127	0.969
Community	0.104	0.082	0.000	0.043	0.087	0.145	0.668
Designed to conserve wholesale process	0.698	0.169	0.042	0.602	0.710	0.824	0.996
Land office plant	0.138	0.168	0.000	0.043	0.090	0.169	0.991
Designed to conserve manufacturing process	0.142	0.079	0.001	0.088	0.138	0.188	0.543
Green investing	0.073	0.065	0.000	0.033	0.060	0.095	0.945
Water	0.223	0.115	0.001	0.143	0.211	0.289	0.861
Training collaboration	0.153	0.106	0.001	0.079	0.135	0.207	0.754
Energy water efficiency	0.217	0.214	0.000	0.059	0.135	0.315	0.973
Green lending	0.080	0.076	0.000	0.020	0.058	0.117	0.512
Air climate	0.272	0.213	0.001	0.111	0.216	0.377	0.986
Designed to conserve agriculture process	0.212	0.101	0.000	0.145	0.203	0.270	0.777
Renewable energy	0.125	0.113	0.000	0.064	0.094	0.143	0.993

Table 4.2 Zero-Shot Text Classification (ZSTC) results (continued)

Labels	**mean**	**std. dev.**	**min**	**25%**	**50%**	**75%**	**max**
Construction practices	0.208	0.137	0.001	0.114	0.179	0.264	0.995
Land life	0.147	0.109	0.000	0.079	0.125	0.182	0.986
Environment products services introduction	0.100	0.104	0.000	0.033	0.074	0.133	0.848
Environmentally innovative wholesale process	0.243	0.130	0.000	0.158	0.224	0.306	0.873
Environmentally innovative manufacturing process	0.059	0.068	0.000	0.020	0.038	0.074	0.783
Material energy use	0.175	0.145	0.002	0.077	0.125	0.223	0.951
Certification	0.104	0.115	0.000	0.038	0.074	0.119	0.991
Cleaner burning energy	0.207	0.088	0.010	0.143	0.197	0.255	0.624
Environmental management	0.252	0.200	0.001	0.098	0.188	0.360	0.939
Resource conservation	0.259	0.147	0.004	0.149	0.237	0.351	0.898
Materials codes	0.259	0.146	0.002	0.155	0.238	0.342	0.952
Land wildlife conservation	0.059	0.046	0.000	0.027	0.049	0.078	0.404
Environmentally innovative agricultural process	0.291	0.131	0.010	0.202	0.273	0.366	0.905
Transportation distribution suppliers	0.208	0.122	0.004	0.127	0.182	0.265	0.924
Safety	0.316	0.152	0.002	0.205	0.291	0.398	0.979
Environmental education information	0.225	0.100	0.003	0.157	0.212	0.277	0.918
Toxin reduction remediation	0.094	0.086	0.000	0.037	0.069	0.122	0.761

4.4.2 Correlation Tests

Table 4.3 shows the Pearson correlations[26] between each web-based environmental culture indicator and the B-Lab environmental index, the dependent variable of our regression model. Normality of all the variables is ensured by using a natural logarithm transformation (see the notes of Table 4.3).

[26] The Pearson Correlation matrix and the descriptive statistics were performed by STATA software v. 16.1

To ease the interpretation of the results, we divide Table 4.3 into three sections. The lower section of the table contains the labels that have either a negative or null correlation with the B-Lab environmental index. Only the last 5 have p-values < 0.005, with the last two presenting p-values <0.001. The last two labels have a score that is weakly inversely related to the B-Lab environmental index. As we move up to the middle of the table, in the positive but lesser than 0.3 correlation portion, only the bottom two labels exhibit low and non-significant correlations. The most interesting section is located in the top part of Table 4.3. There, we find the labels that have the strongest and most significant correlation with the B-Lab environmental index. Among these labels, the top four have a correlation higher than 0.4 with green investing reaching nearly 0.5 (0.497).

Table 4.3 Pearson Correlation results

Labels	**Correlation**	**p-value**
Green investing**	0.497	0.000
Resource conservation**	0.472	0.000
Environmentally innovative wholesale process**	0.467	0.000
Green lending**	0.430	0.000
Environmental management**	0.390	0.000
Designed to conserve wholesale process**	0.356	0.000
Designed to conserve agriculture process**	0.349	0.000
Environmental education information**	0.320	0.000
Environmentally innovative manufacturing process***	0.297	0.000
Materials codes**	0.284	0.000
Designed to conserve manufacturing process**	0.283	0.000
Environment products services introduction**	0.266	0.000
Certification**	0.248	0.000

Table 4.3 Pearson Correlation results (continued)

Labels	Correlation	p-value
Environmentally innovative agricultural process***	0.215	0.000
Material energy use**	0.214	0.000
Outputs**	0.161	0.000
Land life**	0.108	0.000
Community*	0.081	0.007
Cleaner burning energy***	0.052	0.086
Renewable energy***	0.039	0.199
Inputs*	0.007	0.812
Air climate**	-0.003	0.922
Water***	-0.022	0.460
Safety**	-0.024	0.433
Land office plant**	-0.036	0.233
Toxin reduction remediation**	-0.067	0.025
Construction practices**	-0.079	0.008
Transportation distribution suppliers***	-0.080	0.008
Energy water efficiency**	-0.098	0.001
Training collaboration**	-0.161	0.000
Land wildlife conservation**	-0.220	0.000

Notes: The labels with * are transformed with the formula ln((label)+1)
The labels with ** are transformed with the formula ln ((label *10) +1)
The labels with *** are transformed with the formula ln((label*100) +1)

4.4.3 Principal Component Analysis

To further investigate the relationship between the indices produced and the B-Lab environment index, we first conducted a Principal Component Analysis (PCA). This analysis groups the web-based indicators created with the ZSTC into latent variables. This step has a twofold effect: 1) reduce the correlation among the labels; and 2) it groups them into a smaller set of components maintaining trends and characteristics. The PCA analysis[27] groups the 31 items into 6 components or dimensions. The PCA, presented in Table 4.4, yields a high Kaiser-Meyer-Olkin (KMO) of 0,865 and a cumulative variance of 68,828%. The resulting six factors, interpreted using their B-Lab description[28], are as follows: 1) The first component relates to management and finance (*ManFin*); 2) The second focuses on the environmental impact of companies on water and land (*WatLand*); 3) The third covers the energy efficiency of companies (*EnergyEff*); 4) The fourth concerns the impact of companies on the community, including air pollution, economic impact on the area, diversity, civic engagement, and public collaboration (*ComImp*); 5) The fifth is associated with the impact of the companies in agriculture processes and practices (*Agri*); and 6) Lastly, the sixth deals with the processes put in place by the companies to reduce the impact on the environment of manufacturing and transportation process, as well as the safety measure applied by the companies (*ManTransSaf*).

Five of the six components present a high level of reliability with a Cronbach's alpha score greater than 0.70. Although the general rule of thumb suggests that a Cronbach's alpha > 0.60 is acceptable, we retain this last component (*ManTransSaf*), despite its Cronbach's alpha on the low side (0,577), because this study is an exploratory analysis (Hair et al., 1998). Before proceeding to the regression analysis, we calculate the score of the PCA for each factor using the SPSS option regression factor to obtain orthogonal components. The six resulting factors will be used as web-based environmental culture indicators in the regression analysis.

[27] The Principal Component Analysis (PCA) was performed using the IBM SPSS software v.29.

[28] See https://data.world/blab/b-corp-impact-data/workspace/data-dictionary.

Table 4.4 PCA solution and factor loadings

Labels	FinMan	WatLand	EnergyEff	ComImp	Agri	ManTransSaf
Environment products services introduction	0.696	0.244	0.202	0.256	-0.035	0.161
Environmental management	0.822	0.251	0.162	-0.048	0.095	-0.062
Resource conservation	0.882	0.110	0.052	0.015	0.228	0.105
Environmentally innovative manufacturing process	0.629	-0.052	0.194	-0.111	0.157	0.425
Environmentally innovative wholesale process	0.811	-0.081	0.295	-0.019	0.123	0.107
Green investing	0.829	-0.037	0.317	0.091	0.220	0.000
Green lending	0.701	0.057	0.177	0.137	0.270	0.167
Environmental education information	0.718	0.250	-0.047	0.147	0.025	-0.084
Water	0.400	0.628	-0.167	0.066	-0.037	-0.098
Energy water efficiency	0.181	0.771	0.330	0.121	0.042	0.057
Land office plant	0.038	0.628	0.303	0.305	0.228	0.152
Land wildlife conservation	0.007	0.782	0.113	0.065	0.130	0.075
Toxin reduction remediation	0.142	0.708	0.130	0.068	0.067	0.339
Renewable energy	0.313	0.075	0.719	0.107	-0.079	-0.156
Material energy use	0.398	0.208	0.674	0.049	-0.094	0.296
Construction practices	0.006	0.434	0.509	0.211	0.020	0.133
Cleaner burning energy	0.270	0.218	0.647	-0.042	0.162	0.097
Community	0.195	0.021	-0.067	0.786	0.171	0.062
Air climate	0.077	0.200	0.380	0.630	0.018	0.217
Training collaboration	-0.037	0.139	0.059	0.766	-0.009	-0.024

Table 4.4 PCA solution and factor loadings(continued)

Labels	**FinMan**	**WatLand**	**EnergyEff**	**ComImp**	**Agri**	**ManTransSaf**
Designed to conserve agriculture process	0.491	0.134	-0.002	0.135	0.688	0.257
Land life	0.134	0.277	0.075	0.469	0.641	-0.063
Environmentally innovative agricultural process	0.276	0.077	-0.027	-0.035	0.844	0.020
Transportation distribution suppliers	-0.129	0.356	0.161	-0.103	0.100	0.580
Designed to conserve manufacturing process	0.412	-0.021	0.160	0.041	0.199	0.695
Safety	0.093	0.194	-0.120	0.307	-0.125	0.720
KMO						0.865
Eigen Values	5.805	3.317	2.460	2.258	2.033	2.023
% Var	22.327	12.756	9.463	8.684	7.819	7.780
% Var. Cum	22.327	35.082	44.546	53.230	61.048	68.828
Cronbach's alpha	0.917	0.797	0.698	0.676	0.787	0.547
Cronbach's alpha based on standardized items	0.926	0.826	0.747	0.689	0.790	0.577
Number of Items	8.000	5.000	4.000	3.000	3.000	3.000

4.5 Regression Results

Table 4.8, and Table 4.9 present the results of the OLS regressions[29,30,31,32]. In Table 4.7, Reg0 presents the basic regression including only control variables and six factors from the PCA analysis. Reg1 shows the results of the regression that includes variables related to industry, size, and country, while the others present the results of the regression with both the direct and moderating effects of industry (Reg1 to Reg8 in Table 4.7), firm size (Reg 9 to Reg14 in Table 4.8), and country (Reg15 to Reg20 in Table 4.9). The regression exhibits a very high R^2 of approximately 0.58, i.e., we can predict 58% of the variance of the dependent variable.

4.5.1 Direct effects

Table 4.7 shows 9 different regressions: the basic regression (Reg1) that includes the control variables and the six orthogonal components, the complete regression (Reg2) used to test the propositions for the direct effect, and Reg 3-8, which are the regressions with the moderating effect for the industry category. Reg0 demonstrates the significant and negative impact of the two ESG variables, *CommInd* (impact area community) and *GovernInd* (impact area governance). This result aligns with the idea that companies do not exert the same efforts in all the ESG areas (Liute & De Giacomo, 2022). The factors associated with Management and Finance (*FinMan*), Agriculture (*Agri*) and Energy efficiency (*EnergyEff*) have a positive and significant impact on the B-Lab environmental index. More specifically, companies that mention specific topics related to green investing or their good management of the company resources or renewable energy on their website are also those that exhibit a higher environmental index as measured by B-Lab. Likewise, highlighting good stewardship of the land through environmentally friendly processes, such as conserving natural resources or developing innovative agricultural processes, yields a higher score on this environmental index. However, the association is negative and significant for Water and

[29] The Multivariate Ordinary Least Square regression was performed by STATA software v. 16.1.

[30] Three of the PCA web-based environmental indicators had to be normalised prior to the regression analysis: *EnergyEff*, *Agri*, and *ManTransSaf*. Details of the transformations on Table A. 1.

[31] We conducted normality tests, confirming that residuals fall within acceptable bounds for skewness and kurtosis. We also examined autocorrelation between residuals using the Durbin-Watson statistic, observing no autocorrelation within the limits. However, the Breusch-Pagan test rejected the hypothesis of constant variance (homoskedasticity). Due to the heteroskedastic residuals, we employed the “vce robust” option in Stata to mitigate this effect (https://www.stata.com/manuals/semintro8.pdf, Nov. 24th 2023).

[32] We performed a Tobit regression to ensure robustness, yielding results highly similar to the linear regression, which are presented in Annex 3. This might be caused by the fact that we encountered only 7 observations in the left-censored category (0) and one observation in the right-censored category (66.10).

Land (*WatLand*), which is built from the labels "water", "energy water efficiency", "land office plant", "land wildlife conservation" and "toxin reduction remediation", as well as Community impact (*ComImp*) built from the labels "air climate", "training collaboration" and "community". This suggests that some key topics for the environment, such as "energy water efficiency", "land wildlife conservation", and "toxin reduction remediation" do not positively impact the B-lab environment index as expected. While this result requires further analysis, one possible interpretation is that these topics refer to long-term projects. Specifically, companies may only be discussing future projects that do not reflect the current state of the B-Lab environmental index, which evaluates projects already implemented by the company.

The factors related to Management, Transportation, and Safety (*ManTransSaf*) show no impact on the environmental culture. In contrast, and quite surprisingly, companies that emphasize their sense of community and collaboration have a negative impact on the B-Lab environmental index. To explore the dynamics identified through the Pearson correlation in Proposition 1, we examined Reg1, which incorporates the variables from the basic regression and those related to the country, size, and industry category of the companies. In Reg1, it is observed that among the ESG control variables, only *CommInd* (impact area community) is significant, exhibiting a negative association. This suggests that a company with a higher impact community score typically has a lower B-Lab environmental index. Concerning the six PCA factors, the *EnergyEff* factor becomes insignificant when introducing variables related to industry, size, and country. As suspected in the theoretical framework that led to our propositions, industrial differences are probably at play here. Before exploring the moderating effects, let us first examine the direct effects. Depending on the industry category, companies may have more interest in applying good environmental practices when a specific industry category is known for its pollution, or depending on the category, companies could have a direct economic advantage in applying good environmental practices (e.g., de Azevedo Rezende et al., 2019; Hermundsdottir and Aspelund, 2021). In general, compared to the Consumer product and service industry category (the omitted category or baseline), all industry categories but Retail, Transportation and Logistics (*dRetTransLog*) yield significant coefficients. The only industries that have a negative association with the dependent variable compared to the baseline are Media, Restaurant, and Hospitality (*dMediaRestHosp*), and Business Product and Service (*dBusinessPro*) industry categories. As mentioned in Section 3, the impact of these industry categories on the environmental index of B-lab was expected.

Surprisingly, the country does not seem to matter in our model (Table 4.9), i.e., once we have controlled for industry categories and size: the coefficient of *dCanada* compared to the US as the baseline is not significant. This result is not aligned with the literature (e.g., Doran & Ryan, 2012; Aguilera-Caracuel & Ortiz-de-Mandojana, 2013), as researchers studying SI and evidence of green innovation performances have generally shown how several factors contribute to differentiated impacts of the country on its environmental performances. For example, policymakers can incentivize and reinforce positive environmental behavior using subsidies and increasing the regulatory constraints on pollution.

Compared to the micro companies (0 to 9 employees), our baseline, Table 4.8 indicates that medium and large firms have a positive and significant association with the B-Lab environmental index while small firms do not show a significant difference compared to micro firms. Increasing the size of a company from a range of 0-9 to a number of employees higher than 49 results in a higher B-Lab environmental index. The results align perfectly with the literature (e.g., Aguilar-Fernández & Otegi-Olaso, 2018) that suggests firm size impacts a companies' pursuit of SI. While the impact of medium and large enterprises is not statistically different from one another, their differentiated influence contrasts with both small and very small firms.

4.5.2 Moderating effects

This section focuses on the coefficients of the interaction terms presented in the lower part of each regression results table. Table 4.5 presents the moderating effects of the different industry categories. In Reg2, the negative coefficient associated with Business Product and Service industry category is partially mitigated by a strong signal regarding environmental management and manufacturing processes, green investing, and lending (*FinMan*). In Reg3, the industry categories where water conservation and good stewardship of the land matter most, i.e., Agriculture (*dAgricul*) and Building (*dBuild*), are the main contributors to the positive association of this web-based environmental culture indicator with the B-Lab environmental index. However, *WatLand* does not have a significant impact when moderating the industry. Reg5 shows that most of the industry categories exhibit a negative moderating effect on the importance of community and collaboration. With the exception of Agriculture, Building, and Energy Environment industry categories where we observe a positive effect of this web-based environmental culture indicator on the B-Lab environmental index. Additionally, Reg5 displays no significant moderating effect of

ComImp for the industry categories. In Reg6, all the industry categories are not significant and only the moderating effect of Finance and Legal service category, when moderating the relation between the agriculture factor (*Agri*) and the B-Lab environmental index, becomes slightly significant and negative. Thus, the items associated with the factor Agriculture, when addressed by companies in Finance and Legal Services have a negative impact on their environmental score.

Table 4.5 Moderating effect of the industry category for the EnergEff factor

Industry		**Coeff.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
			-46.880	20.419	2.038	-7.394	10.960	2.366	36.997	-8.814	8.429	-2.454
dAgricul	1	-46.880										
dBuild	2	20.419	++									
dConsumPrd	3	2.038										
dEducationTr	4	-7.394										
dEnergyEnvir	5	10.960	++									
dFinLegservic	6	2.366	++									
dHealthHuman	7	36.997	+++		+	+						
dMedRestHosp	8	-8.814							-			
dRetTransLog	9	8.429	++									
dBusinessPro	10	-2.454	+									

+ if the difference between two cells is positive and the p-value ≤ 0.1

++ if the difference between two cells is positive and the p-value ≤ 0.05

+++ if the difference between two cells is positive and the p-value ≤ 0.01

- if the difference between two cells is negative and the p-value l ≤ 0.1

-- if the difference between two cells is negative and the p-value ≤ 0.05

--- if the difference between two cells is negative and the p-value ≤ 0.01

Given the non-significant impact of the *EnergyEff* and *ManTransSaf* factors, further analysis was warranted. Specifically, we compared the moderating effect of industry categories on the impact of these factors on the B-Lab environmental index. The web-based environmental culture indicator linked to Energy Efficiency (*EnergyEff*), becomes negative and significant when moderated by the Agriculture industry category (*dAgricul*) and positive and significant when moderated by the Health and Human industry (*dHealthHuman*) (see Reg5). From an industrial perspective, Agriculture firms show a reduced correlation with the B-Lab environmental index when influenced by a robust web-based environmental culture indicator related to *EnergyEff*, although the overall impact remains positive. Conversely, in the case of *dHealthHuman*, the *EnergyEff* moderated by *dHealthHuman* is positive, but the general effect remains negative. Table 4.5 Moderating effect of the industry category for the EnergEff factor shows that Agriculture has a negative and significant moderating effect on the relationship between *EnergyEff* and environmental impact, compared to other industry categories. The results suggest that Agriculture companies that include concepts related to "renewable energy" (driven items of the *EnergyEff* factor according to the loading of the PCA analysis) in their websites tend to have a lower environmental index. Most other sectors show a positive effect, except for Media, Restaurant and Hospitality (*dMediaRestHosp*), Business Products and Services (*dBusinessProd*), and Education and Training (*dEducationTr*).

Although the Agriculture industry category is broad, encompassing companies producing tractors to shops selling fruits, we observe that companies tend to focus more on the products they are selling or producing, and less on the sustainability of their company (e.g., energy usage to run their activities). Thus, for this industry, the average score of the ZSTC for the items composing *EnergyEff* is less than the average score of the items for companies in other industry categories Additionally, the distinctly negative coefficient for *EnergyEff* moderated by Agriculture, as seen in the table, consistently reflects a positive delta compared to *dConsumPrd*, *dEducationTr*, *dMedRestHosp,* and *dHealthHuman*. For the latter (see Reg4), we observe a positive moderating effect for an industry that includes companies related to health and care. This sector encompasses companies such as veterinary, mental health, and homecare, and the average score of the ZSTC for the items composing *EnergyEff* is generally low due to the lesser importance given to this topic compared to other topics such as community and safety.

The regression results in Reg 7 reveal a substantial negative impact of the cluster comprising the Media, Restaurant, and Hospitality industry categories on the B-Lab environmental index. Similarly, Finance and Legal Services demonstrate a significant negative effect on the B-Lab environmental index. In terms of moderating effects, Reg 7 underscores Agriculture and, Health and Human industry categories as notable for their significant moderating influence on this factor.

Companies in the Agriculture industry category, emphasizing concepts related to the items comprising *ManTransSaf*, exhibit a lower B-Lab environmental index. Conversely, a positive effect is observed when the moderating industry is Health and Human. Table 4.6 illustrates that, when moderated by the Agriculture industry category, the difference in effects is positive compared to Media, Restaurant and Hospitality, Consumer Products, and Services. Notably, Media, Restaurant and Hospitality also exhibit a positive and significant difference compared to Building Consumer Products and Services and Education and Training.

The PCA reveals that "safety" is the driving item for the *ManTransSaf* factor. Although the Media, Restaurant, and Hospitality industry category includes companies ranging from lobster sellers to book publishers, we observed that companies performing well in this category tend to emphasize the concept of "safety". As B-Lab does not provide a precise definition of "safety", we used the knowledge of the NLP model BART to understand the meaning of the word. Our exploration revealed that the context of 'safety' is remarkably broad, encompassing social dimensions like protection from bullies in the Education and Training industry category to physical security concerns, such as ensuring a safe safari experience in the Media, Restaurant, and Hospitality industry category. This broad interpretation of "safety" is reflected in the significant result observed for *ManTransSaf* moderated by *dHealthHuman*, aligning with the nature of companies in this category that prioritize safety measures for the well-being of individuals. However, it's essential to acknowledge that the lack of a precise definition for the label "safety" may contribute to explaining the contrasting results obtained in Reg6.

Regressions 9 to 14 (in Table 4.8) show the moderating effect of the size categories. Small firms that signal a strong web-based environmental culture index regarding green finance and environmental management (*FinMan*) are associated with a higher B-Lab environmental index, hence partially compensating for their small size status compared to their larger counterparts. In contrast, web-based indicators related to water energy conservation and better land management (*WatLand*) have a negative moderating impact for medium-sized companies, implying that the negative association between the web-based measures and the B-Lab index is due to the signals sent by medium-sized companies on their websites. The web-based indicator related to the projects put in place by the companies to reduce the impact on the environment of the manufacturing and transportation process, as well as the safety measures applied (*ManTransSaf*) has a weakly significant and positive impact on the large firms. Overall, the moderating effects due to the size of firms are very small. We suspect that with a larger sample of firms, most of these effects may disappear.

Table 4.6 Moderating effect of the industry categories for the ManTransSaf

Industry		**Coeff.**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
			-4.376	-5.454	1.466	-9.736	1.581	6.315	7.032	16.260	-9.372	-2.233
dAgricul	1	-4.376										
dBuild	2	-5.454										
dConsumPrd	3	1.466	+++									
dEducationTr	4	-9.736										
dEnergyEnvir	5	1.581										
dFinLegservic	6	6.315				++						
dHealthHuman	7	7.032				+						
dMedRestHosp	8	16.260	+	++	+++	+++						
dRetTransLog	9	-9.372			---							
dBusinessPro	10	-2.233										

+ if the difference between two cells is positive and the p-value ≤ 0.1

++ if the difference between two cells is positive and the p-value ≤ 0.05

+++ if the difference between two cells is positive and the p-value ≤ 0.01

- if the difference between two cells is negative and the p-value l ≤ 0.1

-- if the difference between two cells is negative and the p-value ≤ 0.05

--- if the difference between two cells is negative and the p-value ≤ 0.01

Finally, the last six regressions (Table 4.9) explore the moderating effects of the country (Canada compared with the US) on the relationship between the web-based environmental culture indices and the B-Corp environmental index. Table 4.9 shows that there is no moderating effect of the country on the web-based environmental culture indices. The introduction of these moderating effects in the regression models does not significantly influence the sign or level of significance of the coefficients of the variables *EnergyEff*, and *ManTransSaf,* which remain non-significant. In other words, Canadian firms do not have a different effect on the web-based environmental culture indices compared to the American.

Table 4.7 Basic regression results and moderating effect of industry

Variables	Reg1	Reg2	Reg3	Reg4	Reg5	Reg6	Reg7	Reg8
CommInd	-1.658 **	-1.613 **	-1.590 **	-1.623 **	-1.559 **	-1.661 **	-1.551 **	-1.662 **
	(0.805)	(0.737)	(0.701)	(0.730)	(0.737)	(0.739)	(0.745)	(0.744)
GovernInd	-0.306 ***	-0.050	-0.053	-0.045	-0.038	-0.047	-0.046	-0.051
	(0.077)	(0.065)	(0.063)	(0.066)	(0.064)	(0.064)	(0.063)	(0.066)
dsmall		0.885	0.762	0.823	0.717	0.818	0.923	0.989 *
		(0.583)	(0.565)	(0.580)	(0.585)	(0.573)	(0.587)	(0.585)
dmedium		3.276 ***	3.139 ***	3.334 ***	3.203 ***	3.258 ***	3.312 ***	3.306 ***
		(0.760)	(0.747)	(0.752)	(0.756)	(0.756)	(0.767)	(0.760)
dlarge		3.088 ***	2.645 **	3.062 ***	3.092 ***	3.110 ***	2.987 **	3.089 ***
		(1.140)	(1.125)	(1.140)	(1.120)	(1.161)	(1.167)	(1.136)
dCanada		0.794	0.758	0.742	0.809	0.757	0.745	0.828
		(0.663)	(0.638)	(0.672)	(0.669)	(0.673)	(0.666)	(0.664)
dAgricul		7.407 ***	6.846 ***	10.416 ***	117.483 **	8.104 ***	-6.024	24.427
		(2.589)	(2.571)	(2.749)	(52.064)	(2.563)	(47.853)	(31.968)
dBuild		7.430 ***	6.881 ***	6.685 ***	-43.125	9.295 ***	-81.099	28.268
		(2.097)	(2.077)	(1.930)	(58.975)	(2.328)	(61.393)	(27.241)
dEducationTr		-12.385 ***	-15.068 ***	-12.412 ***	5.265	-13.160 ***	-42.702	24.950
		(1.171)	(1.790)	(1.164)	(38.459)	(1.577)	(53.580)	(25.625)
dEnergyEnvir		9.457 ***	12.068 ***	8.275 ***	-17.907	9.073 ***	-16.541	3.276
		(1.818)	(2.755)	(2.016)	(32.702)	(1.939)	(41.319)	(22.737)
dFinLegservic		-13.002 ***	-14.223 ***	-12.725 ***	-18.692	-12.897 ***	26.542	-37.783 **
		(0.772)	(0.873)	(0.785)	(31.415)	(0.766)	(20.662)	(16.363)
dHealthHuman		-10.176 ***	-11.812 ***	-10.053 ***	-97.652 *	-9.709 ***	51.703	-37.587
		(1.177)	(1.417)	(1.166)	(52.633)	(1.208)	(45.102)	(25.465)
dMedRestHosp		-7.773 ***	-8.330 ***	-7.481 ***	12.919	-7.776 ***	10.741	-69.786 **
		(1.895)	(1.864)	(1.872)	(29.606)	(1.843)	(31.144)	(29.882)
dRetTransLog		-1.736	-2.452	-2.039	-21.697	-1.563	-42.772	35.406
		(1.565)	(1.505)	(1.564)	(40.807)	(1.552)	(36.767)	(26.178)

Table 4.7 Basic regression results and moderating effect of industry(continued)

Variables	Reg1	Reg2	Reg3	Reg4	Reg5	Reg6	Reg7	Reg8
dBusinessPro		-9.056***	-8.717***	-9.129***	-3.138	-8.857***	25.453	-0.402
		(0.802)	(0.848)	(0.801)	(33.765)	(0.801)	(27.629)	(16.533)
FinMan	6.667***	4.144***	2.780***	4.181***	4.105***	4.111***	4.165***	4.170***
	(0.367)	(0.392)	(0.502)	(0.394)	(0.387)	(0.387)	(0.397)	(0.394)
WatLand	-2.430***	-0.664**	-0.698**	-0.679	-0.649*	-0.680**	-0.560	-0.631*
	(0.362)	(0.324)	(0.338)	(0.548)	(0.348)	(0.323)	(0.345)	(0.329)
EnergEff	10.514**	2.515	2.385	0.080	2.038	1.319	2.982	3.522
	(4.575)	(4.484)	(4.100)	(4.623)	(7.005)	(4.324)	(4.541)	(4.616)
ComImp	-0.750**	-0.588**	-0.576**	-0.650**	-0.601**	-1.293**	-0.603**	-0.611**
	(0.315)	(0.283)	(0.274)	(0.280)	(0.278)	(0.583)	(0.283)	(0.284)
Agri	25.714***	15.548***	14.226***	15.574***	16.104***	15.830***	16.932***	14.642***
	(4.302)	(3.862)	(3.808)	(4.021)	(3.973)	(3.864)	(5.573)	(3.918)
ManTransSaf	2.346	1.249	1.745	1.593	1.691	1.080	1.340	1.466
	(1.787)	(1.603)	(1.675)	(1.633)	(1.659)	(1.607)	(1.647)	(2.493)
DummyYears	yes	yes	yes	yes	yes	yes	yes	yes
Constant	-69.551***	-21.839	-19.636	-17.646	-24.154	-18.959	-26.931	-22.809
	(14.958)	(14.758)	(14.402)	(15.148)	(19.502)	(14.449)	(17.602)	(16.835)
interaction x ...			…FinMan	…WatLand	…EnergyEff	…ComImp	…Agri	…ManTransSaf
dAgricul			2.487	4.852	-46.880**	3.922	5.105	-4.376**
			(2.890)	(3.296)	(21.969)	(2.920)	(18.557)	(8.272)
dBuild			2.502	3.574	20.419	-3.242	36.904	-5.454
			(2.269)	(1.989)	(23.972)	(2.386)	(25.555)	(7.139)
dConsumProdu								
dEducationTr			-2.733	-1.905	-7.394	1.638	12.907	-9.736
			(2.400)	(1.069)	(16.340)	(1.253)	(22.676)	(6.655)
dEnergyEnvir			-1.032	-2.083	10.960	2.384	11.238	1.581
			(1.676)	(1.327)	(13.074)	(1.473)	(17.566)	(5.974)
dFinLegservic			-0.716	-1.037	2.366	0.959	-16.519*	6.315
			(0.981)	(0.980)	(13.176)	(0.758)	(8.624)	(4.118)
dHealthHuman			-0.133	0.115	36.997*	-0.103	-26.036	7.032*
			(1.849)	(1.200)	(22.278)	(0.939)	(18.899)	(6.453)
dMedRestHosp			2.220	-1.892	-8.814	1.452	-7.863	16.260
			(1.812)	(1.912)	(12.384)	(2.122)	(13.156)	(7.894)

Table 4.7 Basic regression results and moderating effect of industry(continued)

Variables	Reg1	Reg2	Reg3	Reg4	Reg5	Reg6	Reg7	Reg8
dRetTransLog			1.781	-1.373	8.429	1.143	17.159	-9.372
			(1.323)	(1.733)	(17.105)	(1.386)	(15.348)	(6.634)
dBusinessPro			4.313***	0.574	-2.454	1.091	-14.496	-2.233
			(0.988)	(0.806)	(14.123)	(0.736)	(11.519)	(4.195)
Nb obs.	1,110	1,110	1,110	1,110	1,110	1,110	1,110	1,110
F	37.620	61.566	51.944	49.877	52.904	48.604	46.821	47.780
R^2	0.370	0.587	0.605	0.595	0.592	0.593	0.591	0.591
Adjusted R^2	0.363	0.577	0.593	0.581	0.579	0.580	0.578	0.578
Kurtosis	4.148							
Durbin-Watson	2.004							
d_l	1.855							
d_u	1.937							
4-du	2.063							
4-dl	2.145							
Breusch-Pagan	72.690***							

Notes: ***p≤0.001, **p≤0.05, *p≤0.1. The Breusch-Pagan test is a χ^2 with 1 degree of freedom. “dl” and “du” are the lower and upper critical values of the Durbin-Watson test. Since ‘2.004’ falls between “du” and “4-du” there is no autocorrelation.

DummyYears refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions but Reg1

dMicro, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category.

EnergyEff represents the transformation ln (EnergyEff+11)

Agri represents the transformation ln (Agri+11)

*ManTransSaf represents the transformation ln((ManTransSaf+5) *10+1)*

CommInd represents the transformation ln (CommInd+1)

Table 4.8 Regression results exploring the moderating effect of size

Variable	Reg9	Reg10	Reg11	Reg12	Reg13	Reg14
CommInd	-1.574**	-1.661**	-1.614**	-1.612**	-1.684**	-1.607**
	(0.732)	(0.739)	(0.735)	(0.737)	(0.739)	(0.739)
GovernInd	-0.046	-0.045	-0.058	-0.050	-0.051	-0.045
	(0.065)	(0.066)	(0.065)	(0.065)	(0.065)	(0.066)
dAgricul	7.280***	7.357***	7.447***	7.402***	7.382***	7.463***
	(2.579)	(2.594)	(2.582)	(2.581)	(2.550)	(2.591)
dBuild	7.545***	7.355***	7.401***	7.465***	7.486***	7.406***
	(2.082)	(2.091)	(2.090)	(2.108)	(2.116)	(2.106)
dEducationTr	-12.374***	-12.373***	-12.242***	-12.451***	-12.307***	-12.326***
	(1.155)	(1.185)	(1.183)	(1.177)	(1.172)	(1.179)
dEnergyEnvir	9.128***	9.483***	9.245***	9.506***	9.771***	9.361***
	(1.833)	(1.800)	(1.845)	(1.815)	(1.828)	(1.822)
dFinLegservic	-12.967***	-13.065***	-12.967***	-12.997***	-12.924***	-13.053***
	(0.773)	(0.777)	(0.778)	(0.778)	(0.775)	(0.770)
dHealthHuman	-10.197***	-10.300***	-10.204***	-10.273***	-10.183***	-9.917***
	(1.190)	(1.191)	(1.186)	(1.200)	(1.181)	(1.153)
dMedRestHosp	-7.856***	-7.861***	-7.780***	-7.814***	-7.782***	-7.802***
	(1.910)	(1.865)	(1.911)	(1.899)	(1.901)	(1.881)
dRetTransLog	-1.654	-1.631	-1.608	-1.694	-1.546	-1.687
	(1.584)	(1.546)	(1.573)	(1.569)	(1.567)	(1.598)
dBusinessPro	-9.077***	-9.091***	-9.058***	-9.051***	-9.016***	-9.031***
	(0.797)	(0.802)	(0.801)	(0.803)	(0.802)	(0.799)
dCanada	0.785	0.750	0.809	0.796	0.813	0.751
	(0.662)	(0.659)	(0.664)	(0.668)	(0.661)	(0.666)
dsmall	0.845	0.908	-3.457	0.881	-17.652	8.952
	(0.573)	(0.584)	(21.111)	(0.583)	(19.475)	(12.906)
dmedium	3.286***	2.946***	-25.990	3.239***	-29.394	-10.630
	(0.773)	(0.776)	(24.519)	(0.763)	(23.957)	(16.511)
dlarge	2.971	3.096***	42.135	3.104***	-2.059	-40.373
	(1.893)	(1.136)	(55.087)	(1.154)	(34.307)	(24.864)
FinMan	3.524***	4.154***	4.169***	4.143***	4.146***	4.160***
	(0.483)	(0.390)	(0.392)	(0.393)	(0.391)	(0.394)

Table 4.8 Regression results exploring the moderating effect of size (continued)

Variable	Reg9	Reg10	Reg11	Reg12	Reg13	Reg14
WatLand	-0.664**	-0.387	-0.610*	-0.657**	-0.646**	-0.739**
	(0.328)	(0.474)	(0.329)	(0.325)	(0.323)	(0.329)
EnergEff	2.712	1.946	0.554	2.342	2.163	2.579
	(4.557)	(4.468)	(6.361)	(4.480)	(4.460)	(4.511)
ComImp	-0.611**	-0.617**	-0.621**	-0.564	-0.587**	-0.592**
	(0.281)	(0.284)	(0.282)	(0.426)	(0.282)	(0.283)
Agri	15.606***	15.493***	15.357***	15.502***	10.774**	15.515***
	(3.845)	(3.828)	(3.877)	(3.857)	(5.407)	(3.906)
ManTransSaf	1.404	1.453	1.228	1.230	1.211	0.895
	(1.622)	(1.613)	(1.619)	(1.605)	(1.601)	(2.389)
DummyYears	yes	yes	yes	yes	yes	yes
Constant	-23.228	-21.013	-16.535	-21.231	-9.175	-20.613
	(15.052)	(14.596)	(19.411)	(14.676)	(18.502)	(17.559)
interaction x …	…FinMan	…WatLand	…EnergyEff	…ComImp	…Agri	…ManTransSaf
dsmall	1.470**	-0.128	1.823	-0.251	7.727	-2.058
	(0.697)	(0.689)	(8.817)	(0.618)	(8.128)	(3.282)
dmedium	0.940	-1.966*	12.287	0.413	13.681	3.550
	(1.015)	(1.048)	(10.264)	(0.789)	(10.047)	(4.210)
dlarge	0.459	-0.252	-16.330	-0.093	2.156	10.896*
	(2.708)	(0.897)	(23.088)	(0.995)	(14.367)	(6.299)
Nb obs.	1,110	1,110	1,110	1,110	1,110	1,110
F	55.775	56.444	55.938	55.299	56.206	55.737
R^2	0.589	0.589	0.588	0.587	0.588	0.589
Adjusted R^2	0.578	0.578	0.577	0.576	0.577	0.578

Notes: ***p≤0.001, **p≤0.05, *p≤0.1. *dMicro*, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category.

DummyYears refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions

EnergyEff represents the transformation ln (EnergyEff+11)

Agri represents the transformation ln (Agri+11)

*ManTransSaf represents the transformation ln((ManTransSaf+5) *10+1)*

CommInd represents the transformation ln (CommInd+1)

Table 4.9 Regression results exploring the moderating effect of country

Variables	Reg15	Reg16	Reg17	Reg18	Reg19	Reg20
CommInd	-1.593**	-1.639**	-1.613**	-1.613**	-1.616**	-1.610**
	(0.737)	(0.739)	(0.736)	(0.738)	(0.737)	(0.737)
GovernInd	-0.050	-0.053	-0.045	-0.050	-0.051	-0.049
	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)	(0.065)
dsmall	0.888	0.864	0.835	0.885	0.891	0.880
	(0.583)	(0.584)	(0.585)	(0.585)	(0.584)	(0.583)
dmedium	3.289***	3.273***	3.263***	3.276***	3.277***	3.311***
	(0.758)	(0.760)	(0.759)	(0.762)	(0.760)	(0.761)
dlarge	3.100***	3.046***	3.106***	3.087***	3.096***	3.173***
	(1.140)	(1.143)	(1.137)	(1.140)	(1.141)	(1.132)
dCanada	0.766	0.707	27.868	0.794	-5.851	24.223
	(0.638)	(0.644)	(23.496)	(0.663)	(25.589)	(17.495)
dAgricul	7.454***	7.346***	7.477***	7.407***	7.474***	7.486***
	(2.595)	(2.592)	(2.593)	(2.591)	(2.591)	(2.591)
dBuild	7.490***	7.460***	7.377***	7.431***	7.435***	7.524***
	(2.096)	(2.101)	(2.093)	(2.098)	(2.098)	(2.096)
dEducationTr	-12.345***	-12.373***	-12.384***	-12.386***	-12.389***	-12.395***
	(1.168)	(1.165)	(1.179)	(1.172)	(1.169)	(1.169)
dEnergyEnvir	9.466***	9.436***	9.460***	9.457***	9.461***	9.559***
	(1.821)	(1.828)	(1.806)	(1.819)	(1.820)	(1.818)
dFinLegservic	-12.973***	-12.956***	-13.064***	-13.002***	-12.992***	-12.964***
	(0.770)	(0.773)	(0.772)	(0.774)	(0.775)	(0.771)
dHealthHuman	-10.140***	-10.150***	-10.182***	-10.177***	-10.174***	-10.214***
	(1.174)	(1.178)	(1.181)	(1.179)	(1.176)	(1.183)
dMedRestHosp	-7.714***	-7.725***	-7.721***	-7.773***	-7.771***	-7.729***
	(1.904)	(1.892)	(1.886)	(1.897)	(1.897)	(1.880)
dRetTransLog	-1.757	-1.699	-1.800	-1.736	-1.729	-1.644
	(1.565)	(1.563)	(1.560)	(1.566)	(1.570)	(1.555)
dBusinessPro	-9.029***	-9.048***	-9.054***	-9.056***	-9.050***	-9.003***
	(0.801)	(0.805)	(0.803)	(0.805)	(0.803)	(0.805)
FinMan	4.081***	4.140***	4.134***	4.144***	4.145***	4.149***
	(0.421)	(0.392)	(0.392)	(0.392)	(0.392)	(0.391)

Table 4.9 Regression results exploring the moderating effect of country (continued)

Variables	Reg15	Reg16	Reg17	Reg18	Reg19	Reg20
WatLand	-0.672**	-0.791**	-0.640**	-0.664**	-0.661**	-0.689**
	(0.323)	(0.341)	(0.323)	(0.328)	(0.324)	(0.322)
EnergEff	2.553	2.309	4.471	2.515	2.486	2.535
	(4.507)	(4.493)	(4.873)	(4.486)	(4.485)	(4.472)
ComImp	-0.589**	-0.572**	-0.579**	-0.586*	-0.585**	-0.568**
	(0.283)	(0.286)	(0.282)	(0.311)	(0.283)	(0.284)
Agri	15.544***	15.684***	15.632***	15.548***	15.141***	15.489***
	(3.853)	(3.849)	(3.825)	(3.858)	(4.073)	(3.845)
ManTransSaf	1.251	1.323	1.305	1.249	1.262	2.124
	(1.603)	(1.600)	(1.606)	(1.607)	(1.607)	(1.684)
DummyYears	yes	yes	yes	yes	yes	yes
Constant	-22.026	-21.846	-26.983	-21.840	-20.835	-25.240
	(14.817)	(14.714)	(15.557)	(14.773)	(15.082)	(14.783)
Interaction x …	…FinMan	…WatLand	…EnergyEff	…ComImp	…Agri	…ManTransSaf
dCanada	0.391	0.662	-11.319	-0.008	2.780	-5.976
	(0.781)	(0.871)	(9.809)	(0.739)	(10.682)	(4.440)
Nb obs.	1,110	1,110	1,110	1,110	1,110	1,110
F	59.254	59.167	60.045	59.276	59.272	59.494
R^2	0.587	0.587	0.587	0.587	0.587	0.588
Adjusted R^2	0.576	0.577	0.577	0.576	0.576	0.577

Notes: ***p≤0.001, **p≤0.05, *p≤0.1. *dMicro*, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category.

DummyYears refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions

EnergyEff represents the transformation ln (EnergyEff+11)

Agri represents the transformation ln (Agri+11)

*ManTransSaf represents the transformation ln((ManTransSaf+5) *10+1)*

CommInd represents the transformation ln (CommInd+1)

4.6 Discussion and conclusion

The main objective of this research was to determine whether web-based sustainable innovation indicators can serve as a proxy for performance measures and indices built using traditional survey-based and administrative data. This pilot study compares web-based environmental culture indicators with the environmental index created by B-Lab.

The first proposition was based on three main assumptions: first, nowadays companies feel greater pressure to implement environmental initiatives; second, environmentally friendly companies pursuing green goals have competitive advantages (Paelman et al., 2020); and third an official website serves as a platform for delivering authentic, precise, and up-to-date information about companies (Jiang et al., 2023). This proposition was accepted based on the moderate correlation of certain topics with the environmental index (Table 4.3) and the results of the regression analysis. The latter shows that the web-based indicators created from ZSTC, together with control variables, industry, size, and country attributes, contribute to explaining over 57% of the variance of the B-Lab environmental index.

The second proposition concerns the direct impact of the country in which the company operates on the evidence of its appropriate environmental approach. This proposition is rejected because our study shows no significant differences between the firms located in Canada and the United States on the B-Lab environmental index. Arguably, the well-aligned green policies and long-standing agreements in protecting the environment and decarbonizing the industry between Canada and the United States contribute to explaining this result[33]. The results by no means imply that the country does not influence the way companies address environmental issues, take actions, and write about them on their external communication channels such as their corporate websites. Also, proposition 2m is rejected. Although Magnusson et al. (2011) suggested that the country shapes the perception of corporate sustainability of the customers, we found no differences in the signal captured on the website of Canadian and United States companies.

The third proposition examined the impact of a company's size on its approach to sustainable environmental practices and is accepted. Researchers (e.g., Ketata et al., 2015) suggest that the size of a company may impact its approach to environmental practices, but there is no consensus in the literature on whether the impact is positive or negative. Our study clearly shows that larger companies have a higher environmental index, indicating a positive impact of size. There are likely both internal and external reasons for this finding. As highlighted by previous research (e.g., Aguilar-Fernández & Otegi-Olaso, 2018), larger companies are more exposed to criticism from customers and stakeholders, which can pressure them to pursue eco-conscious practices. Larger

[33] See for instance the joint statement : https://www.canada.ca/en/environment-climate-change/news/2021/04/joint-statement-by-the-us-environmental-protection-agency-and-environment-and-climate-change-canada-on-environment-and-climate-change.html.

companies have more control over the market, as they can more easily choose suppliers that adopt green practices. Furthermore, they generally have more resources to invest compared to smaller companies (Aguilar-Fernández & Otegi-Olaso, 2018). However, we found that the moderating effect of size is weak or infrequent and the evidence generated only partially supports proposition 3m. Some industry categories are likely to be dominated by firms of specific sizes. We suspect that a triple interaction between web-based environmental indicators, industry category, and size is needed to disentangle this.

Lastly, the fourth proposition regarding the direct effect of industry categories on the B-Lab environmental index is accepted. With the exception of Retail, Transportation, and Logistics, all industry categories showed a significant impact compared to the baseline. There are specific industrial characteristics that influence firms' approach to sustainable innovation, with some industries having more interest in pursuing environmental goals due to economic or social pressure (Hermundsdottir & Aspelund, 2021b). Proposition 4m, which aimed to explore the moderating effect of industry category on the relationship between our web-based environmental indicators and the B-Lab index, is also accepted. All six web-based indicators have a significant impact on the B-Lab environmental index when the industry category is changed. Both Agriculture and Health and Human industry categories moderate the relationship between the B-Lab index and the web-based indicators regarding water and land stewardship, energy efficiency, and community implication. The web-based indicator for Manufacturing, Transportation, and Safety moderates that relationship for the industry categories of Agriculture and Health and Human.

This study makes significant advancements both in theoretical frameworks and methodological approaches, enriching the discourse in Sustainability Research and Signal Theory. On the theoretical front, our paper makes a dual contribution to sustainability studies. Primarily, by establishing a correlation between our web indices and the B-Lab environmental index, our research underscores the potential of website communication to reflect a company's environmental performance. Furthermore, through the regression analysis, we disentangle the effect of the web indices on the B-Lab environment index shedding light on the subjects correlated to a higher B-Lab environment index. Concerning Signal Theory, our work responds to the critical inquiry posed by Connelly et al. (2011): "Does the signal represent a valid and reliable measure of the underlying quality that the signaler is attempting to communicate?" In answering this, our investigation confirms the feasibility of extracting meaningful signals from company websites, validating these within the sustainability domain via the B-Corp environmental index, thus bridging a vital research gap.

From a methodological perspective, our study brings to light the potential to accurately replicate the B-Lab index, which can have a game-changing impact in several aspects. Leveraging Zero-Shot Text Classification (ZSTC) alongside a pre-trained language model, our approach yields impressive results, explaining upwards of 57% of the variance in the B-Lab index. This efficiency is achieved without extensive textual preprocessing or reliance on an annotated, which streamlines research efforts in terms of time and cost.

Moreover, with further validation and considering the limitations presented in the following section, our methodology holds promise for forecasting the B-Lab scores assigned to corporations. Furthermore, by relying on a pre-trained language model and employing a semantic similarity approach, we maintain a consistent representation of the labels proposed by B-Lab This capability, potentially extendable across various fields and research inquiries, paves the way for real-time analysis. Such a tool could provide policymakers with a dynamic overview of specific issues, offering a preliminary assessment of policy impacts ahead of more traditional evaluation techniques. This represents a significant stride towards mitigating the challenges posed by conventional questionnaire-based surveys, providing early insights into the efficacy of environmental policies and strategies.

4.7 Limitation and future works

Despite the promising results, the methodology used in this study is nonetheless subject to multiple limitations. The first limitation is inherent to the ZSTC task, which is considered the most challenging task for NLP models (T. Brown et al., 2020). The model only has access to the label and the text, without any examples or further explanations, which forces it to interpret everything by itself. It also increases the misinterpretation and ambiguity of the already complex natural language. A second implicit limitation arises the nature of natural language itself. By definition, language is inherently ambiguous, presenting challenges in the study of semantics. This is what we are doing in this study: attempting to understand the meaning behind a paragraph to be able to label it.

Another limitation is related to the labels used. We chose the labels directly from the items that B-Lab uses to evaluate the environmental culture of a company. However, these items may not be precise enough for the concept we are seeking, and contacting B-Lab could help us obtain more appropriate labels to better target certain topics. A further limitation arises from the NLP model used in the ZSTC task. The model represents words in vectors based on the knowledge acquired during training, which is not specific to the SI problem addressed in this research. As mentioned in Section 3.2, this knowledge is derived from English Wikipedia and BooksCorpus.

Finally, there are some limitations that stem from the data. The research conducted is an exploratory study about the capacity of the applied methods to predict the environmental readiness of the companies. We chose for our study only Canadian and American companies. This limits our research and our results since this sample, as mentioned in the data description, was chosen to be homogeneous from a country point of view. Additionally, we lost data because we could not find the companies' websites using the Wayback Machine. Because we needed to gather the websites of data near the certification date, we needed to recover the old website of the companies. For these reasons, our sample was downsized.

Building from these limitations, new avenues to improve this research are possible. First, expanding the sample by adding other countries will help to understand if the results can be generalizable for the B-Corp data. Moreover, using other datasets than the website of the B-Corp companies would be needed to increase the reliability of our results. Second, our literature review revealed that researchers have uncovered intriguing findings about the environmental strategies employed by companies. These findings were obtained by analyzing the companies' entire websites. Indeed, some researchers have found that sometimes the companies dedicate a full page to their environmental initiatives and strategies (Calabrese et al., 2021; Fernández-Vázquez & Sancho-Rodríguez, 2020). Additionally, by analyzing images on the website using their descriptions, researchers may better understand the messages that companies are conveying (Fernández-Vázquez & Sancho-Rodríguez, 2020). Thus, it could be interesting to compare our results obtained from analyzing only the homepage with those obtained from analyzing the full website including the images. This analysis could enhance our understanding of the results. Third, the NLP model used in this research is not specialized. Thus, creating a specialized model could be beneficial, especially when investigating more granular concepts in a precise topic. Indeed, advanced NLP models, like BART offer the possibility to be fine-tuned to different tasks. In other words, BART can be trained to accomplish tasks that have never been done before. Moreover, one can use the fine-tuning process to specialize BART to a particular topic. For example, BioBERT

(J. Lee et al., 2020) was built to be used for several NLP tasks in the biomedical domain, outperforming the previous models. Taking the lead from this research, it is possible to create a specific model for SI that might have a greater comprehension of certain specific concepts. Lastly, we used the algorithm for the ZSTC, which, to the best of our knowledge, is the best to perform this task. However, the continued expansion and increasing work done in the field of NLP will possibly bring more advanced techniques that can better tackle this task.

CHAPTER 5 ARTICLE 2: The Digital Pitch: A Hybrid AI Approach to Decoding Funding Signals on Corporate Websites

Pietro Cruciata, Davide Pulizzotto and Catherine Beaudry

Submitted to the Journal of Business Venturing (July 22th 2025)

Keywords: RAG, Private funding, Venture Capital, Bertopic, Signalling Theory

Abstract

Effectively monitoring private funding trends in real-time presents a significant challenge for policymakers and investors, as traditional data sources often lag. This study addresses this gap by developing and testing a novel framework to predict private funding outcomes for new ventures using unstructured data from their corporate websites. Drawing on signaling theory, we analyze a sample of Canadian companies founded between 2020 and 2024. Our hybrid methodology combines a top-down, hypothesis-driven approach using Retrieval-Augmented Generation (RAG) to create targeted web indicators for founder background, collaborations, and previous funding mentioned, with a bottom-up, exploratory approach using topic modeling to discover emergent textual patterns.

Our analysis yields several key findings. We find strong evidence that signals of founders' human capital (academic and industry experience) are associated with a higher likelihood of securing private funding. In contrast, signals for firm collaboration show highly sector-dependent associations. Our bottom-up analysis identifies that service-oriented topics, such as digital marketing and IT support, are negatively associated with receiving capital, though the overall predictive power of general website text is modest (AUC 0.547), suggesting a "noisy channel".

This pilot study makes a dual contribution. Methodologically, it pioneers the combined use of RAG and topic modeling for the scalable analysis of web data, offering a more nuanced tool for researchers. Practically, it provides a framework for the timely monitoring of venture dynamics, extending signaling theory by showing that corporate websites act as complex signaling environments where both intentional signals and broader content are associated with funding outcomes.

5.1 Introduction

The venture capital (VC) and private equity (PE) industries have experienced remarkable growth over the past two decades. In 1994, VC and PE funds managed approximately $100 billion globally (Metrick & Yasuda, 2011), with expectations of reaching $6 trillion by the end of 2024. This substantial growth underscores the increasing reliance on VC as a pivotal financing mechanism for innovative ventures. Venture capitalists have become critical intermediaries for funding young, offering not only capital but also strategic direction and specialized knowledge crucial for addressing the intricate obstacles these businesses face (Metrick & Yasuda, 2021; Van den Heuvel & Popp, 2023). Empirical evidence consistently highlights the beneficial impact of VC investments, demonstrating improved firm-level outcomes, such as enhanced growth and innovation (Puri & Zarutskie, 2012), as well as broader economic gains, including increased entrepreneurship rates, employment, and overall income (Samila & Sorenson, 2011).

However, VC and PE firms, like other resource providers, often encounter selection challenges due to intense competition among entrepreneurs seeking funding (Vanacker et al., 2020). The entrepreneur-investor relationship is inherently characterized by information asymmetry, where entrepreneurs possess superior insights into their ventures' true qualities, leaving investors uncertain about these attributes. This imbalance poses difficulties for high-quality entrepreneurs whose strengths might remain unrecognized by investors. Signalling theory addresses this by describing how entrepreneurs with superior but hidden qualities can credibly communicate their strengths to reduce investor uncertainty (Bafera & Kleinert, 2023; O. Colombo, 2021; Spence, 1973). Effective signals, such as past funds received (Vanacker et al., 2020) or functional prototypes (Steigenberger & Wilhelm, 2018), enable investors to infer entrepreneurial quality. Nevertheless, signalling theory is based on assumptions that are frequently not satisfied in entrepreneurial contexts. Under Knightian uncertainty (Packard et al., 2017), where future outcomes are unknowable, entrepreneurs themselves might misjudge their ventures' potential, making signals unreliable indicators of future success.

These challenges are compounded by the inability of traditional data sources, such as surveys, to deliver timely insights into entrepreneurial signals or funding trends. Surveys are often slow and expensive, making it challenging for policymakers and strategic managers to stay informed in a timely, ideally real-time, manner (Axenbeck & Breithaupt, 2021). Recognizing these limitations, innovation studies have increasingly explored the use of non-intrusive and publicly available datasets, such as company websites, as alternative information sources (e.g., Gök et al., 2015). While websites have been used to study innovation or customer engagement, their role in signaling private funding activities remains largely unexplored, creating a critical gap in real-time funding analysis (Bafera & Kleinert, 2023).

To address this gap, this study proposes a comprehensive approach, integrating top-down and bottom-up methodologies, to explore the potential of using company websites for the real-time monitoring of private funding dynamics. Drawing on signalling theory, we posit that companies use their websites to signal their characteristics and activities to stakeholders (including investors).

Specifically, our top-down approach investigates the potential of Generative AI, leveraging the Retrieval-Augmented Generation (RAG) framework, to construct web-based indicators from company website signals. These signals are related to concepts and indicators solidly grounded in the literature, such as funding announcements or investor relations sections. This approach aims to assess whether these AI-derived indicators reliably replicate or correspond to classical indicators derived from traditional data sources, such as funding amounts or growth metrics. Thus, we seek to determine if this method can be used to monitor private funding dynamics for recent startups.

In contrast, the bottom-up approach leverages topic models and supervised machine learning methods to analyze the linguistic patterns that differentiate privately funded companies from those without private funding. This method allows us to identify unanticipated signals that can lead to novel indicators for private funding trend analysis.

These approaches complement each other: the top-down method validates website signals against established metrics, while the bottom-up method uncovers novel signals, together enabling a robust real-time monitoring system. Ultimately, the development of real-time monitoring tools for entrepreneurship, specifically for private funding dynamics, could offer valuable insights for policymakers and strategic managers, enabling them to track investment and entrepreneurial trends within specific ecosystems or domains.

This paper is structured as follows. First, we draw upon signalling theory to develop four testable propositions. Second, we describe our research design, including the data collection and the top-down and bottom-up analytical approaches. Third, we present the results of our empirical analysis and discuss their significance. We conclude with a summary of our findings and their implications for future research.

5.2 Literature review

Signalling theory defines a “signal” as a deliberate action undertaken by a more informed party to effectively and credibly convey information in contexts characterized by information asymmetry. This theory posits that high-quality ventures can distinguish themselves through observable activities or characteristics that are difficult for lower-quality ventures to imitate (Spence, 1973). This theoretical framework has been widely applied across various management disciplines, with increasing prominence particularly within entrepreneurship research. This development is unsurprising, considering how signal theory helps resolve one of the main obstacles faced by new ventures: reducing severe information asymmetries for important audiences such as potential customers, strategic partners, and investors (Bafera & Kleinert, 2023). Signals reveal information and thus reduce the level of resource-holder uncertainty (Bergh et al., 2014; Spence, 1973; Wesley II et al., 2022). Uncertainty reduction is thus a necessary consequence of a signaling mechanism. If uncertainty is reduced, then applying the theory may be appropriate. If communication does not reduce resource-holder uncertainty, it cannot be understood as a signal in the sense of signalling theory (Steigenberger et al., 2024).

Confirming a growing interest for signaling theory, Bafera and Kleinert (2023) analyzed more than 200 studies that focus on successful entrepreneurship signals. While the entrepreneurship literature has increasingly incorporated signalling theory, it has often overlooked the strategic importance of corporate websites as key communication tools. In this context, research within the field of signalling theory provides instructive insights into how companies deliberately utilize their official websites to influence stakeholders’ perceptions. Studies such as those by Mavlanova et al. (2012) and Yildiz et al. (2023) highlight the deliberate strategies companies employ through their online presence to communicate credibility and competence to their audiences. For instance, Mavlanova et al. (2012) conducted a study on the role of website signals as a means for online retailers to communicate their product quality, proposing and validating a three-dimensional framework. Jiang et al. (2023) argued that a corporate official website serves as a credible source of non-financial information for assessing the credit risk of Small and Medium Enterprises (SMEs). Cruciata et al. (2024) demonstrated that website-based signals correlate positively with the B Corp environmental index, revealing that firms actively showcase their commitment to environmental values through their online presence.

Considering the application of signal theory in the study of private funding, Islam et al. (2018) discovered that receiving government grants increased the likelihood of follow-on VC funding, especially for startups lacking other strong signals. They explained that such grants help distinguish startups from their peers, enabling them to overcome information asymmetries and establish ties with VCs. This finding is confirmed by other researchers. Wu et al. (2020), for instance, stated that Government R&D subsidies served as a signal to attract additional venture capital for renewable energy enterprises. The underlying mechanism is often described as a "certification effect," where the rigorous evaluation for a competitive subsidy signals a firm's quality and reduces information gaps for investors (Bellucci et al., 2023). However, adding nuance to the impact of grants, Stevenson et al. (2021) discovered that while they speed up a venture's ability to attract private investment, they do not necessarily increase its future revenue. The power of signaling extends beyond grants, as Vanacker et al. (2020) found that strong performance in prior funds helps new private equity firms raise subsequent funds, regardless of market uncertainty. In light of this evidence, we propose the following:

Proposition 1: Companies that have received private funding are more likely to mention other funding received on their website

Ko & McKelvie (2018) discovered that education and prior founding experience are crucial signals for first-round financing, while investor prominence gains importance in subsequent rounds. They also found that the interaction between these signals shows that prominence magnifies the effects of human capital, significantly influencing funding amounts over time. In a similar vein, Piva & Rossi-Lamastra (2018) posited that entrepreneurs' business education and entrepreneurial experience significantly contribute to their success in raising funds through equity crowdfunding. The importance of founder background as a key signal is further highlighted in the research of Bhattacharyya & Subrahmanya, (2024), who analyzed digital startups in India and found that VCs prioritize signals from prestigious educational backgrounds, incubator affiliations, and digital business models. This aligns with the argument from Bellavitis et al. (2019), who suggested that founder education and venture patents serve as signaling mechanisms which VCs use in the pre-investment phase to reduce information asymmetries and thus mitigate adverse selection. Given these intriguing findings on the impact of founder backgrounds, we posit the following proposition:

Proposition 2: Companies that receive private funding are more likely to highlight founder experience and background on their website.

Finally, Caviggioli et al. (2020) analyzed over a thousand young innovative companies to discover that for VC funding, qualitative aspects such as R&D collaboration for patents are more valued. Baum and Silverman (2004) identified startups' alliances (i.e., social capital), patents (i.e., intellectual capital), and top management (i.e., human capital) as key signals of startup potential. Miloud et al. (2012) found that the number of alliances among other factors, has a positive impact on the chances of fundraising and receiving higher evaluation. According to Hoenig and Henkel (2015), VCs seem to interpret existing research alliances as signals of a startup's technological quality. Conversely, Ozmel et al. (2013) suggested that while previous alliances may crowd out VC investments, the number and size of a startup's alliances still significantly affect its likelihood of going public. Based on these findings, we put forth the following proposition:

Proposition 3: Companies that receive private funding are more likely to highlight collaboration on their website.

Venture-capital investments naturally shift in response to factors that steer funding toward particular sectors or regions (Corea et al., 2021). Thus, VC funds are influenced by market developments. Governments also wield regulation to shape VC market liquidity: for example, U.S. banking sector limits on VC investments reduced fund sizes, follow-on fundraising, and ultimately harmed startups (J. Chen & Ewens, 2021), whereas rules allowing pension fund participation fueled the late-1990s VC surge (P. Gompers & Lerner, 2001a). At the same time, VC activity follows distinct boom-and-bust cycles tied to broader economic trends, investor sentiment, and overall capital availability. During economic downturns deal counts, total capital deployed, and average deal sizes contract sharply, particularly for early-stage ventures (Howell et al., 2020). Similarly, during the COVID-19 pandemic, limited partners' pressure to conserve capital caused many VCs to scale back on new financing rounds (P. Gompers et al., 2021). Considering these findings, we posit the following proposition:

Proposition 4: Topics communicated by companies that receive private funding differ from those communicated by companies that did not receive private funding.

5.3 Data

To validate these four propositions, we use the CrunchBase company data, to which we merge information from the Wayback Machine, a digital archive of the World Wide Web, that allows users to retrieve web pages across different points in time.

The CrunchBase dataset, created by CrunchBase[34], is a business intelligence platform that provides information about startups, tech companies, investment, funding rounds, acquisitions, mergers, founders, investors, and key players in the business and technology ecosystem. We chose a subsample of the full dataset that include the Canadian companies founded between 2020 and 2024, our results are based on the November 2024 update. The main variable used is 'Last Founding Date' which we leverage to determine whether a company was privately financed. If there is a date in this column, it indicates that the company received private funding. Additionally, we limit our sample to companies founded between 2020 and 2024, and use the 'website' (URL) variable to retrieve each company's website through the Wayback Machine for every year from their founding year through 2024.

Following Cruciata et al. (2024), we meticulously extract textual content from websites using targeted HTML tags to build a comprehensive dataset for analysis. Figure 5.1 depicts the steps that lead to the final sample. From Crunchbase, we retrieved companies founded in Canada between 2020 and 2024, resulting in a sample of 6,086 companies. Then, we attempted to retrieve the website snapshots for each company for each year in the considered range. Using this time frame, we obtained 23,694 snapshots of 6,086 websites for all the companies. At this point, we merged the website content because we aim to evaluate the companies' communication as a whole, resulting in 5,710 companies. Finally, before proceeding with the analysis, we removed the websites of companies that contained only Cascading Style Sheets (CSS) code or Wayback Machine errors, leading to a final sample of 5,696 companies, 1,761 of which have received private fundings.

[34] https://www.crunchbase.com

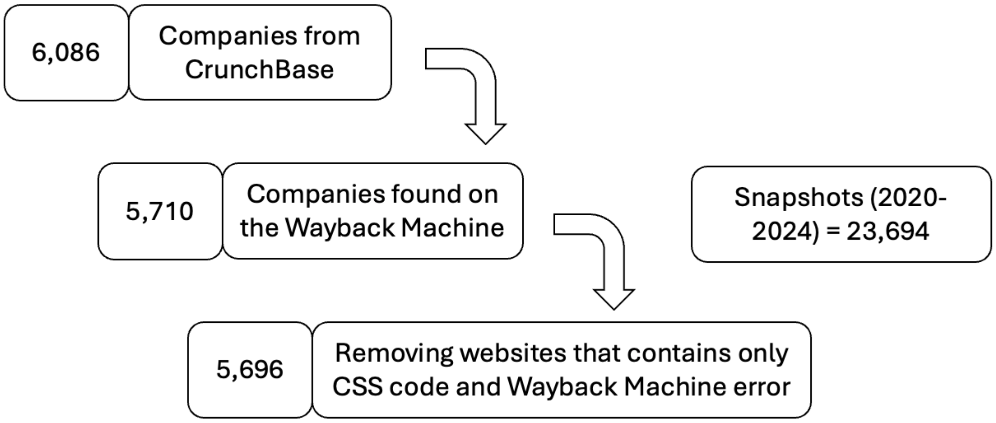
6,086
Companies from CrunchBase
5,710
Companies found on the Wayback Machine
Snapshots (2020-2024) = 23,694
5,696
Removing websites that contains only CSS code and Wayback Machine error

Figure 5.1 Dataset creation steps

5.4 Methodology

The methodology section is divided into two parts and discussed in the following order: first, the top-down approach, in which we combined the Retrieval-Augmented Generation (RAG) framework (P. Lewis et al., 2020), followed by a logistic regression to answer the first three propositions; second, the bottom-up approach in which we use BERTopic (Grootendorst, 2022) and supervised machine learning methods to address the 4th proposition.

5.5 Top-down approach

5.5.1 Website Analysis using LLM

Following data preparation, the analytical process begins with a comprehensive examination of the corporate websites' text utilizing a state-of-the-art LLM, selected based on our computational capabilities and performances. The advent of transformer architecture (Vaswani et al., 2017) has paved the way for the development of LLMs, a category of artificial intelligence known for exceptional performance across a multitude of natural language processing (NLP) tasks. The enhancements in these models' capabilities have led to significant progress in text synthesis and a variety of downstream NLP applications (T. Brown et al., 2020).

For our analysis, we selected Deepseek-r132B (DeepSeek-AI et al., 2025) to analyze the signal related to the selected key topics introduced earlier. Deepseek-r1 operates based on a user-generated prompt. We choose Deepseek-r1 for two main reasons: First and foremost, the model is open access, ensuring wide availability and ease of use for researchers and practitioners alike. Secondly, Deepseek-r1 is known for its outstanding performance across a broad spectrum of tasks, demonstrating exceptional reasoning capabilities. Indeed, this model is part of the reasoning models. These models are trained using the Chain-of-thought prompting, namely a series of intermediate reasoning steps significantly improves the ability of large language models to perform complex reasoning (Wei, Wang, et al., 2022).

Although LLMs have achieved remarkable success, they still face significant limitations, particularly in domain-specific or knowledge-intensive tasks, where they may produce "hallucinations" (Y. Zhang et al., 2023) when handling queries beyond their training data or requiring current information. These hallucinations can arise from an overload of data, a lack of contextual relevance, or both, compromising the reliability of LLMs in practical applications (Y. Gao et al., 2023). To overcome challenges, RAG (P. Lewis et al., 2020) enhances LLMs by retrieving relevant document chunks from external knowledge based on semantic similarity calculation. Exclusively using external knowledge, the RAG framework avoids the problem of generating factually incorrect content. Its integration into LLMs has resulted in widespread adoption, establishing RAG as a key technology in advancing chatbots and enhancing the suitability of LLMs for real-world applications (Y. Gao et al., 2023). While the integration of RAG with LLMs has achieved notable technological advancements (Y. Gao et al., 2023; H. Li et al., 2022; Zhao et al., 2023), the existing literature primarily focuses on these innovations themselves, overlooking their practical applications in the business sector.

To improve information retrieval accuracy for our specific use case, we adopt an advanced RAG pipeline (Y. Gao et al., 2023) displayed in Figure 3.2. This advanced framework builds upon the standard Naïve RAG architecture, which is divided into three steps: indexing, retrieving, and generation. The indexer embeds and organizes data into a searchable format. The retriever calculates the semantic similarity between the user query and texts in the indexed corpus. The generation step combines the user query with the retrieved documents to form a coherent prompt, that is then used by the LLM to provide a response.

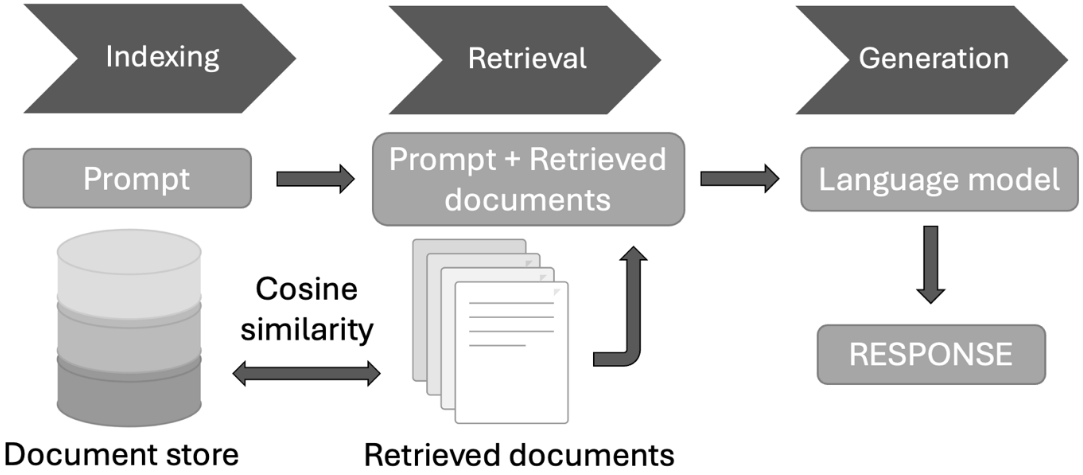

Figure 5.2 Pipeline RAG

To enhance retrieval relevance, we modified the standard approach by adding a threshold on the retriever as pre-processing so that returns only the website chunks that have a cosine similarity higher than 0.5, with a maximum of 10 chunks. The retriever performs this cosine similarity filtering using an alternative and simplified version of the query (see Table B 6 in the Appendix), specifically, a version that focuses on the key topics we want to retrieve rather than the full user prompt. Thus, we also fine-tuned the hyperparameters to optimize the performance of our RAG pipeline. To obtain structured and consistent results, we followed an iterative process involving hyperparameter tuning and careful prompt design. The selected hyperparameters and their final values are presented in Table 5.1.

Table 5.1: Hyperparameter used in the RAG.

Parameters	Meaning	Values chosen
Chunk_size	Maximum number of characters or tokens allowed in a single segment of text when breaking down documents.	1,024
Top_k	Number of most relevant chunks of text that the retriever will return based on their similarity to the user's query.	10
Temperature	This parameter controls the randomness or creativity of the model's output during the text generation process. Lower temperature (e.g., 0.1 or 0.2) makes the model more deterministic.	0
Embedding model	This model is responsible for the creation of the Document store.	BAAI LLM-Embedder
Context Windows	The amount of text measured by tokens that the LLM can process at once to generate a response.	3,900

After determining the hyperparameters, we tested various prompts through multiple iterations and ultimately selected those shown in Table 5.2. This phase required considerable effort because the wording of the prompts significantly influences the consistency and structure of the outcomes. We performed extensive manual checks and revisions to ensure that the prompts effectively guided the model toward the desired structured results.

Our analysis focuses on three key areas: founder experience, funding sources, and collaboration indicators. For each area, the LLM first answers 'Yes' or 'No' to indicate the presence of the information on the website, followed by a justification provided in the 'Explanation' field. This field facilitates manual verification of the answers given by the model.

For founder experience analysis, we ask the LLM model to provide additional information about whether the founder's background comes from academia, industry, or both. Similarly, for funding analysis, we include an extra field asking the LLM to specify whether the funding comes from government sources, private organizations, or both. The collaboration analysis required a more nuanced approach due to the complexity of defining collaboration clearly. We implemented a two-step process: first, we identified companies that mentioned collaboration on their websites using the RAG approach used for founder experience and for funding received. Then, focusing only on these pre-selected companies, we asked the model to return a score on a scale from 1 to 10, based on the clarity and explicitness with which collaboration was mentioned (Table B 3). The processes above generated three types of web indicators: dummy variables indicating the presence or absence of different founder backgrounds, dummy variables indicating the different types of funding received by companies, and a qualitative indicator measuring collaboration on a 1–10 scale. Before proceeding to the logistic regression, we addressed the score distribution of the collaboration results by splitting them into four categories: No WebCollab, Low WebCollab, Medium WebCollab, High WebCollab.

Table 5.2: Prompts used to extract the information

Topic	Prompt
Collaboration	Please evaluate the company's website to determine if the website explicitly mentions any partnership or collaboration. Structure your response as a JSON object with two keys: 'Response' and 'Explanation'. - In the 'Response', indicate one of the following: 'Yes' or 'No'. - In the 'Explanation', provide precise and concise evidence from the website to support your answer. If there is no relevant information, state 'No information available'. **Important**: Provide **only** the JSON object in the following format and **do not include any additional text**:
Fundings	Please carefully review the company's website to identify any explicit mentions of the company receiving a grant, subvention, or financial support from external sources. Structure your findings strictly as a JSON object with three keys: 'Response', 'Explanation', and 'Source'. - Under 'Response', explicitly state either 'Yes' or 'No'. - Under 'Explanation', provide a precise quotation or a concise summary from the website as evidence supporting your response. If no relevant information is found, explicitly state: 'No information available'. - Under 'Source', specify the entity providing the funding by choosing exactly one of these categories: 'Government', 'Private organization', 'Both', or 'Not specified'. If no information about the funding source is provided, select 'Not specified'. **Important Instructions:** - Do not include any additional commentary, context, or explanations outside the JSON structure.

Table 5.2: Prompts used to extract the information (continued)

Topic	Prompt
Founder Experience	Please evaluate the company’s website to determine if the website mentions any information about the founder work experience or academic background. Structure your response as a JSON object with three keys: ‘Response’, ‘Explanation’, ‘Type’. “Response”: Provide “Yes” if the website mentions such information, or “No” if it does not. “Explanation”: Clearly and concisely reference the specific evidence from the website supporting your answer. If no relevant information is found, state “No information available.” “Type”: Categorize the founder’s experience explicitly mentioned as one of the following: - “Academic” if only academic background is mentioned. - “Industry” if only industry - related experience is mentioned. - “Both” if both academic and industry experiences are mentioned. - “Not specified” if there is no relevant information.

Regression analysis

Using these indicators, we then estimated a logistic regression model to verify the first three propositions. The regression model structure is as follows:

$$y_i = \alpha + \sum_i \theta_i x_i + \sum_w \lambda_w Web_w + \sum_j \delta_j D_j + \varepsilon_i \qquad (5.1)$$

where $ln\left[\frac{p_i}{(1-p_i)}\right] = y_i$, x_i represent CrunchBase variables (*dGrowth*, *dSize*[35], *dHConf*, *dProvince*[36]), Web_i represent the web-based variables (*NoWebFounderExp, WebFounderExpAcademic, WebFounderExpIndustry, WebFounderExpBoth, WebFundingFPTGov, WebFundingNonFPT, WebFundingBoth, NoWebFunding, No WebCollab,*

[35] dSize indicates the dummy variable for each size category: *extremely small* (1-10), *very small* (11-50), *small* (51-100), *small-medium* (101-250), *medium & large* (>250). From the initial configuration of Crunchbase, we grouped companies with >250 employees into a single category because we had a small number of companies in the larger size categories.

[36] dProvince represents the dummy variables for the grouped provinces of Canada. We started with the headquarter location provided by CrunchBase, and given the low number of companies some Canadian provinces, we decide to group them as follows: Eastern Canada (including Newfoundland, Nova Scotia, New Brunswick, Prince Edward Island) Prairies provinces (including Saskatchewan, Manitoba and Alberta) Quebec, and British Columbia.

Low WebCollab, Medium WebCollab, High WebCollab), D_i represent industry dummy variables (*dIndustry*)[37]. y_i is the dependent variable $dFunded_i \in \{0,1\}$ that we are trying to predict, with $i \in N, 1 \leq i \leq$ 5,696. $WebFounderExpAcademic_i \in \{0,1\}$ represents the web signal related to the founder experience in academy, $WebFounderExpIndustry_i \in \{0,1\}$ represents the web signal related to the founder experience in industry, $WebFounderExpBoth_i \in \{0,1\}$ denotes the web signal related to the founder experience in academy and in the industry, $WebFundingFPTGov_i \in \{0,1\}$ refers to the web signal related associated to the funding received from federal, provincial or territorial (FPT) government sources. $WebFundingNonFPT_i \in \{0,1\}$ indicates the web signal not related to FPT government funding (i.e., private sources). $WebFundingBoth_i \in \{0,1\}$ represents the web signal related to companies that receive both FPT government funding and private funding. $NoLow\ WebCollab_i \in \{0,1\}$ reflects the companies that have no signal about their collaboration. $Medium\ WebCollab_i \in \{0,1\}$ indicates the companies that have medium intensity signal about their collaboration while $High\ WebCollab_i \in \{0,1\}$ refers to the companies that have a high intensity signal about their collaboration. Finally, $\alpha, \theta, \lambda, \delta \in R$ are the coefficients in the regression model while $\varepsilon_i \in R$ indicates the error term.

5.6 Bottom-up approach

The bottom approach consists of two steps. First, we use BERTopic[38] to uncover hidden thematic structures in large text corpora. Then, we apply supervised methods to identify which topics are most determinant for the classification task.

5.6.1 BERTopic

BERTopic is an integrated topic modeling technique that uses embedding vectors and class-based TF-IDF[39] (c-TF-IDF) to create dense clusters, allowing for interpretable topics from text data. BERTopic employs UMAP[40] (McInnes et al., 2020) to reduce the dimensionality of embeddings, which are then clustered using HDBSCAN[41] (McInnes et al., 2017). Classical topic modeling techniques, such as Latent Dirichlet Allocation (LDA), and Dynamic Topic Models, rely on

[37] The industry variable was created based on the industrial sector description provided through the variable "Industry" variable in CrunchBase. We mapped the 3748 descriptions to their corresponding 2-digits NAICS code.

[38] BERT: Bidirectional Encoder Representations from Transformers.

[39] TF-IDF: Term Frequency-Inverse Document Frequency

[40] UMAP : Uniform Manifold Approximation and Projection

[41] HDBSCAN : Hierarchical Density-Based Spatial Clustering of Applications with Noise

frequency-based approaches to derive unobserved topics from large text corpora. However, these methods remove context by focusing solely on term frequencies. In contrast, BERTopic generates document embeddings using pre-trained transformer-based language models, clusters these embeddings, and creates topic representations through the c-TF-IDF procedure. In other words, BERTopic enables us to incorporate contextual knowledge from large text datasets.

More specifically, we utilized BERTopic with "all-mpnet-base-v2" as the embedding model. Before proceeding with clustering, we performed several preprocessing steps to emphasize the words most relevant to the clustering process. First, we segmented the documents into chunks of sentences, ensuring each chunk contained between 150 and 250 tokens. This segmentation was carried out using the SpaCy package. During this process, SpaCy was also employed to refine the chunks by removing stopwords (via the NLTK package) and conducting a part-of-speech (POS) analysis to retain only nouns, verbs, adverbs, and adjectives. With preprocessing complete, we moved on to clustering. Since BERTopic relies on UMAP and HDBSCAN, we conducted a grid search across four parameters, detailed in Table 5.3.

Table 5.3: BERTopic parameters tested and their meaning

Parameters	**Meaning**	**Values tested**
n_neighbors	Defines how many nearest neighbors each point considers when it builds the manifold.	30, 50, 75
n_components	Number of dimensions in the resulting embedding space.	300, 500, 600, 750
epsilon	Maximum distance two points can be apart to still be considered "neighbors"	0.3, 0.4, 0.5
min_cluster_size	Minimum number of chunks part of the same cluster to be considered valid in the analysis	30, 75, 100

We evaluated the models using two criteria. Initially, we selected the parameter combination that minimized the number of documents assigned to the outlier cluster (-1). This process yielded two configurations with comparable performance. Consequently, we manually reviewed the topics generated by each configuration to ensure their word representations were coherent. Next, to align with our objective of creating cluster representations that distinguish between companies receiving private funding and those that do not, we employed a supervised method to finalize the BERTopic configuration. Before applying this supervised method, we performed additional preprocessing. BERTopic assigns each chunk a soft membership score for every topic by computing the cosine similarity between the chunk's embedding and each topic-centroid embedding, then normalising those similarities so they sum to 1. Finally, before proceeding with the classification task, we performed a pre-processing step on the results of BERTopic. For each company, we aggregated the chunks from all available website data across the 2020-2024 years period, computing the average probability of each topic across all chunks. We averaged topic probabilities over period for two reasons: first, our analysis revealed that website content remained largely stable over short timeframes; second, we sought to maximize contextual depth to identify topics more prevalent among companies that received private funding.

5.6.2 Supervised methods

For the supervised approach, we compare three different models: Random Forest, Extreme Gradient Boosting (XGBoost) and Neural Networks (all the models are used through the package scikit-learn in python). These models are widely adopted in classification tasks due to their complementary characteristics: Random Forest (RF) provides interpretability and robustness against overfitting, XGBoost delivers high performance through sequential error correction, and Multilayer Perceptron (MLP) leverages deep learning to capture complex non-linear patterns. The models are represented as follow:

$$W = \{topic_i, \dots, topic_I\}\underset{\rightarrow}{f(.)}dFunded, \tag{5.2}$$

with $topic_i$, $i \in \{1,2,3, \dots ,118\}$[42], being the variable representing topic associated to the clusters created using BERTopic, $f(.)$ being the machine learning function used and $dFunded$ that takes value 1 if the companies received a private funding and 0 otherwise. To address class imbalance, we applied the following oversampling techniques: (i) Random Over-Sampling Examples (ROSE): generates synthetic samples using a smoothed bootstrap approach (Lunardon et al., 2014); (ii)

[42] The 118 topics are the results of the most performant BERTopic setting (see section 5.2)

Synthetic Minority Oversampling Technique (SMOTE) produces synthetic minority-class samples by interpolating between randomly selected instances from the k-nearest neighbors (Chawla et al., 2002); (iii) Adaptive Synthetic Sampling (ADASYN) generates synthetic samples based on the neighborhood of each minority instance (He et al., 2008); (iv) KMeansSMOTE combines k-means clustering with SMOTE to mitigate between-class and within-class imbalances (Last et al., 2017); (v) SVMSMOTE utilizes Support Vector Machines (SVM) to identify minority instances near the decision boundary and interpolates those to create synthetic samples (Nguyen et al., 2011).

To evaluate the results, we use two measures that stem from the Confusion matrix (Table 5.4), a well-known tool for a clear representation of the elements correctly and incorrectly classified (Maimon & Rokach, 2005).

Table 5.4: Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	True Negative	False Positive
Actual Positive	False Negative	True Positive

In our binary classification problem, we aim to predict whether companies receive private funding, defining the positive class as companies that received funding and the negative class as those that did not (the null hypothesis assumes no funding). We evaluate model performance using three key metrics: the Receiver Operating Characteristic (ROC) curve, the Precision-Recall (PR) curve, and the weighted F1 score. The ROC curve demonstrates the classifier's diagnostic ability by plotting the True Positive Rate (TPR) on the y-axis against the False Positive Rate (FPR) on the x-axis as the discrimination threshold varies. TPR, also known as sensitivity or recall, measures the proportion of funded companies correctly identified:

$$TPR = \frac{TruePositive}{TruePositive + FalseNegative} \tag{5.3}$$

The false positive rate is the ratio between the number of non-defaults wrongly categorized as companies private funded (false positives) and the total number of actual negative samples (companies that did not receive private fundings). FPR is also known as fall-out or false alarm ratio and is the probability of falsely rejecting the null hypothesis:

$$FPR = \frac{FalsePositive}{TrueNegative + FalsePositive}. \tag{5.4}$$

Additionally, we also use the weighted version of the F1 score to evaluate the performance of the model:

$$F1score_{\text{weighted}} = \sum_{k=1}^{K} \frac{s_k}{N} \frac{2P_k R_k}{P_k + R_k} \tag{5.5}$$

where P_k is the precision for class k, R_k is the recall for class k, s_k is number of true instances (support) of class k, $N = \sum_{k=1}^{K} s_k$ total instances. The weighted F1 score adjusts for class imbalance by assigning weights proportional to class frequencies in the test set. This makes it particularly valuable when misclassifying the minority class (funded companies) carries significant costs. These metrics collectively guide us in selecting the best-performing model for predicting funding outcomes.

5.7 Results

5.7.1 Top-down approach

After creating the web indicators through the RAG (see Table B 3 and Table B 4 in the Appendix for the prompts used), Table 5.5 presents the results of the logistic regressions[43,44]. From left to right, we present four different logistic regressions: Reg1 contains only the control variables, Reg2 presents only the web indicators, Reg3 combines control variables and web indicators, and Reg4 adds an interaction term between the collaboration signal and the companies' NAICS code compared to Reg3. The results show that the highest Pseudo R^2 of 15,58% is reached by Reg3,

[43] The logistic regression was performed by STATA software v. 16.1

[44] The Breusch-Pagan test rejected the hypothesis of constant variance (homoskedasticity). Due to the heteroskedastic residuals, we employed the "vce robust" option in Stata to mitigate this effect (https://www.stata.com/manuals/semintro8.pdf, June 18th 2025).

meaning that our best regression explains almost 16% of the variance of the in private funding received by the companies.

We begin our analysis of the direct effects by examining differences across industry categories, using manufacturing (NAICS 31–33) as the baseline. We first discuss sectors with statistically significant differences compared to this baseline, followed by those without significant differences. Ventures in agriculture, forestry, fishing, and hunting (NAICS 11) have twice the odds of receiving private funding compared to manufacturing companies. Similarly, companies in mining, oil and gas extraction, utilities, and construction (NAICS 21–23) exhibit 88% higher likelihood of receiving funding compared to manufacturers. Conversely, businesses in wholesale trade, retail trade, transportation, and warehousing (NAICS 41, 44–45, 48–49) experience about 37% lower odds of securing private funding than the baseline. Real estate and leasing companies (NAICS 53) have approximately 57% lower odds compared to manufacturing firms. The odds decrease by roughly 75% for companies in professional, scientific, and technical services (NAICS 54). Firms in administrative, support, and waste-remediation services (NAICS 56) perform slightly better but still have 62% lower odds of obtaining private funds. Finally, the combined residual category of "other services" and public administration (NAICS 81, 91) has odds approximately 46% below the manufacturing benchmark. The results also highlight industry sectors whose odds of receiving private funding are not statistically different from those of manufacturing. These include information and cultural industries (NAICS 51), finance and insurance (NAICS 52), the combined sector of educational services and healthcare (NAICS 61–62), and the leisure-oriented sector consisting of arts, entertainment, recreation, accommodation, and food services (NAICS 71–72).

Increasing firm size negatively impacts the likelihood of receiving private funding. Compared to the baseline category of *micro* companies, all other firm size categories show statistically significant odds ratios below one. Moving from micro-enterprises to very small-sized firms, the odds decrease by approximately 18%. Compared to small companies the odds decrease by about 70%. Medium-sized companies experience about a 63% reduction, while large companies see their likelihood of obtaining funding reduced by approximately 68%. This reinforces the observation of a funding environment primarily geared toward the smallest, earliest-stage businesses.

Turning to growth dynamics (*dGrowth*) and growth confidence indicators (*dHConfidence*) as estimated by CrunchBase, we set as the baseline companies categorized with no or low growth and no confidence. Compared to this baseline, companies experiencing growth but without high confidence ratings have odds 14 times greater for receiving private funding. Companies that are both growing and rated with high confidence have even higher odds, approximately 30 times greater, compared to companies with no or low growth and no confidence ratings.

The geographical location of a company's headquarters also significantly influences funding odds. Relative to firms headquartered in Ontario, those based in the Atlantic provinces are three times more likely to secure private funding, and companies in British Columbia see a 29% higher likelihood. Quebec-based firms show no statistically significant difference compared to Ontario. Firms in the Prairie provinces demonstrate a marginal positive effect of approximately 17%, significant only at the 10% significance level (p-value < 0.1). However, in Reg3, the effect for Prairie provinces loses statistical significance.

Turning now to the web-based variables, we begin the analysis with the founder-experience signals, proceed to funding signals, and conclude with collaboration signals. Across this sequence, the logistic regressions evidence makes clear that founder's previous experience is the strongest predictor of private financing. Using websites that reveal no founder experience as the baseline, ventures whose pages highlight academic credentials enjoy odds of private investment that are 85% higher in the pure web-signal model (Reg 2, odds ratio = 1.850, p < 0.05) that is only marginally reduced to 69% when standard controls are added (Reg 3, odds ratio= 1.685, p < 0.10) and to 67% when moderating effects are included (Reg 4, odds ratio = 1.665, p < 0.10). The most powerful signal appears when sites showcase founders with both academic and industry experience: the likelihood of funding nearly doubles in Reg 2 (odds ratio = 1.856, p < 0.01) and remains roughly 55% higher in the subsequent specifications (Reg 3 odds ratio = 1.549 and Reg 4 odds ratio = 1.547, p < 0.01 and p < 0.05, respectively). By contrast, signalling industry experience alone leaves the probability of receiving private funding prospects essentially unchanged, never reaching statistical significance.

Prior public financing paints a nuanced but internally consistent picture: ventures that have secured only non-FPT grants, namely funding streams that originate outside federal, provincial, or territorial programmes, enjoy a roughly 70–74 percent increase in the odds of attracting private capital, rising to an odds ratio of 1.736 in the pure web-signal model (Reg 2, p < 0.05), holding at 1.705 once controls variables are introduced (Reg 3, p < 0.10), and returning to 1.736 when moderating effects are added (Reg 4, $p < 0.10$). In contrast, firms that signal FPT fundings awards, show odds ratios clustered close to one and never reach statistical significance. Similarly, web collaboration signals prove equally inconsequential. Visible collaboration activity on the web, whether signals are low (odd ratio = 1.066), medium (odd ratio = 0.906), or high (odd ratio = 1.121) has no effects on the odds of securing private investment once founder credentials and public-funding history are accounted for, with all coefficients being near unity and lacking statistical significance. This neutrality underscores that investors care far more about founders' academic signals, especially when paired with industry experience, and about endorsements from diverse non-governmental policy makers, whereas industry experience alone, flagship FPT grants, and online collaboration footprints contribute little to swaying private-capital decisions.

The moderating analysis in Reg 4 reveals that the payoff to web-visible collaboration differs sharply across industries once manufacturing firms that display no or low collaboration signals is the omitted category. Among firms that show no website collaboration signal at all, those in agriculture, forestry, fishing and hunting (NAICS 11) are already more than twice as likely to raise private capital (odds ratio = 2.295, $p < 0.05$); a similarly strong advantage is seen for the resource-and-construction cluster comprising mining, utilities and construction (NAICS 21-23, odds ratio = 1.857, $p < 0.01$). When these same sectors broadcast low levels of web collaboration signal, their odds remain elevated: 2.061 for NAICS 11 and an even larger 2.357 for NAICS 21-23 (both $p < 0.05$). These results suggest that a modest collaboration web signal neither adds to nor undermines the intrinsic appeal of primary-industry ventures. At the opposite extreme, high web collaboration signals amplify the chances of receiving private funding still further (odds ratio = 3.120, $p < 0.05$), whereas they wipe out the benefit for agriculture (odds ratio = 0.707, ns). Service sectors present a mirror image: professional, scientific and technical services (NAICS 54) are already penalized without collaboration (odds ratio = 0.268, $p < 0.01$) and suffer a greater drop when they move to medium collaboration (odds ratio = 0.105, $p < 0.001$), though the effect eases to odds ratio = 0.217 ($p < 0.05$) under high web collaboration signals. Administrative and support services (NAICS 56) odds range from 0.187 to 0.451 (all $p < 0.10$), while wholesale–retail–transport (NAICS 41, 44-45, 48-49) switch from a 34% odds decrease in the absence of web collaboration signal (odds ratio = 0.658, $p < 0.05$) to a statistically non-significant results once any collaboration is signalled. Finally, the leisure and hospitality ensemble (NAICS 71-72) show no odds increase unless web collaboration signals are high, where their odds more than double (odds ratio = 2.822, $p < 0.10$).

In summary, these results suggest that web-visible collaboration amplifies existing funding advantages in resource-based industries, slightly reduces disadvantages in low-tech service sectors, and significantly boosts the odds for leisure-oriented businesses, while manufacturing remains consistently neutral regardless of collaboration intensity.

Table 5.5: Regression results (odds ratios) and moderating effect of industry.

Variables	REG1		REG2	REG3		REG4	
Size [dExtremely small (1-10 employees) omitted]							
dVery small (11-50)	0.821	***		0.828	***	0.830	**
	(0.059)			(0.060)		(0.060)	
dSmall (51-100)	0.297	***		0.297	***	0.298	***
	(0.057)			(0.058)		(0.058)	
dMedium (101-250)	0.373	***		0.365	***	0.370	***
	(0.121)			(0.119)		(0.121)	
dLarge (> 250)	0.311	***		0.319	***	0.328	***
	(0.132)			(0.135)		(0.133)	
Growth [dNo-Low Growth omitted]							
dMedium-High Growth	14.405	***		14.52	***	14.63	***
x dNot High Confidence	(2.665)			(2.686)		(2.738)	
dMedium-High Growth	30.946	***		30.97	***	32.73	***
x dHigh Confidence	(10.465)			(10.502)		(10.818)	
Region [Ontario omitted]							
dAtlantic Provinces	2.951	***		2.994	***	3.01	***
	(0.482)			(0.492)		(0.502)	
dQuebec	1.163			1.135		1.141	
	(0.115)			(0.114)		(0.115)	
dPrairies	1.203	*		1.195	*	1.17	
	(0.119)			(0.119)		(0.117)	
dBritish Columbia	1.292	***		1.283	***	1.280	***
	(0.112)			(0.112)		(0.112)	

Table 5.5: Regression results (odds ratios) and moderating effect of industry.

Variables	REG1	REG2		REG3		REG4	
WebFounderExp [No WebFounderExp omitted]							
WebFounderExpAcademic		1.850	**	1.684	*	1.663	*
		(0.498)		(0.485)		(0.489)	
WebFounderExpBoth		1.856	***	1.549	***	1.549	**
		(0.275)		(0.260)		(0.265)	
WebFounderExpIndustry		0.970		0.948		0.913	
		(0.130)		(0.145)		(0.143)	
WebFunding [No WebFunding omitted]							
WebFundingBoth		1.036		0.97		0.945	
		(0.623)		(0.749)		(0.742)	
WebFundingNonFPT		1.736	**	1.705	*	1.73	*
		(0.430)		(0.481)		(0.506)	
WebFundingFPTGov		0.860		0.805		0.78	
		(0.294)		(0.318)		(0.303)	
WebCollab[No WebCollab omitted]							
Low WebCollab signal		1.066		1.13			
		(0.113)		(0.136)			
Medium WebCollab signal		0.906		0.85			
		(0.103)		(0.113)			
High WebCollab signal		1.121		1.07			
		(0.147)		(0.160)			

Table 5.5: Regression results (odds ratios) and moderating effect of industry.

VARIABLES	REG1		REG2		REG3		REG4	
Industry [31-33 Manufacturing omitted]								
[11] Agriculture, forestry, fishing & hunting	2.086	**			2.027	**		
	(0.650)				(0.627)			
[21] Mining, quarrying, & oil and gas extraction, [22] Utilities, [23] Construction	1.882	***			1.898	***		
	(0.323)				(0.326)			
[41] Wholesale trade, [44-45] Retail trade, [48-49] Transp. & warehousing	0.635	***			0.644	***		
	(0.107)				(0.109)			
[51] Information and cultural industries	1.077				1.090			
	(0.153)				(0.155)			
[52] Finance & insurance	1.003				1.018			
	(0.163)				(0.167)			
[53] Real estate & rental & leasing	0.432	***			0.437	***		
	(0.122)				(0.124)			
[54] Professional, S&T services	0.267	***			0.272	***		
	(0.042)				(0.043)			
[56] Admin. and support, waste management & remediation serv.	0.384	***			0.395	***		
	(0.101)				(0.104)			
[61] Educational services, [62] Health care and social assistance	1.280				1.264			
	(0.193)				(0.191)			
[71] Arts, entertainment & recreation, [72] Accom. & food serv.	0.768				0.784			
	(0.144)				(0.148)			
[81] Other services (except public admin.),	0.540	**			0.550	**		
[92] Public admin.	(0.141)				(0.145)			
Constant	0.46	***	0.43	***	0.45	***	0.45	***
	(0.063)		(0.015)		(0.062)		(0.069)	

Table 5.5: Regression results (odds ratios) and moderating effect of industry.

Interaction x	WebCollab	None x	Low x	Medium x	High x
WebCollab x Industry [No-Low Webcollab x Manufacturing omitted]					
[11]		2.294 **	2.060		0.708
[21, 22, 23]		1.874 ***	2.354 **	1.132	3.116 **
[31-33]			1.149	1.424	0.585
[41, 44-45, 48-49]		0.655 **	0.482	0.940	0.518
[51]		1.101	1.224	0.807	1.216
[52]		0.945	1.596	1.047	1.679
[53]		0.347 ***	0.575	0.380	2.772
[54]		0.265 ***	0.580 *	0.118 ***	0.217 **
[56]		0.451 ***	0.187	0.185	0.279 *
[61-62]		1.352 *	0.963	1.117	0.884
[71-72]		0.722	0.521	1.121	2.819 *
[81-92]		0.570 *	0.698	0.496	
Pseudo R^2		0.148	0.004	0.150	0.156
Log pseudolikelihood		-2799.397	-3508.318	-2791.188	-2771.960
Number of observations		5,231	5,696	5,231	5,277

Notes: ***p ≤0.001, **p ≤0.05, *p ≤0.1.
Odds ratios presented, and standard errors in parentheses.

5.7.2 Bottom-up approach

Table 5.6 presents the configurations of two BERTopic models[45], which share similar topics and comparable document frequencies for topic -1. As shown, the sole difference in hyperparameters between these models lies in the number of components. Specifically, this refers to the

[45] To run BERTopic model, we use the package bertopic in python.

dimensionality considered when analyzing the embedding space. These two BERTopic models serve as the foundation for the supervised analysis conducted to identify the best-performing model.

Table 5.7 compares the performance of these two optimal BERTopic configurations using AUC (Area Under the Curve, a measure of model accuracy) and weighted F1 scores (a metric that balances precision and recall). We evaluated these models using various techniques to address dataset imbalance, employing XGBoost as the comparison model. The BERTopic1 configuration achieved superior results, demonstrating the highest scores in both AUC and weighted F1 metrics. Moreover, among the imbalance-correction methods tested, the SMOTE technique produced the best performance.

Table 5.6: The settings of the two best BERTopic models.

	BERTopic 1	**BERTopic 2**
n_neighbors	30	30
n_components	500	750
epsilon	0.3	0.3
min_cluster_size	100	100

Table 5.7: Comparison BERTopic 1 and BERTopic 2 using XGBoost.

	BERTopic 1		BERTopic 2	
	AUC	F1 weighted	AUC	F1 weighted
Random Over Sampling	0.529	0.617	0.50	0.594
SMOTE	0.537	0.623	0.50	0.588
ADASYN	0.522	0.609	0.51	0.602
SVMSMOTE	0.515	0.599	0.509	0.600
K-means SMOTE	0.504	0.600	0.486	0.582

Table 5.8 summarizes the AUC and weighted F1 scores for three different supervised models. The Random Forest model, when coupled with the SMOTE imbalance-correction method, yielded the highest overall performance, achieving an AUC of 0.547 and a weighted F1 score of 0.627. These results represent approximately a 5% improvement over the baseline model that generates random predictions aligned with the class distribution in the training set. While this improvement alone might not definitively prove the predictive power of BERTopic1-generated topics for identifying companies receiving private funding, the findings enable us to pinpoint the topics most influential in guiding the Random Forest model's predictions. Figure 5.3 presents the ten most significant topics considered by the Random Forest model in its decision-making process for each observation. As shown, Topic_2 carries the greatest importance, contributing a weight of approximately 0.035. Following closely are Topic_10, with a weight slightly above 0.03, Topic_0 nearing 0.03, and Topic_23, which surpasses 0.025. Due to a notable drop-off in significance thereafter, we concentrate solely on these four primary topics.

Table 5.8:Comparison between XGBoost, Random Forest and a Neural Network on the BERTopic 1.

	Random Forest		XGBoost		Neural Network	
	AUC	F1 weighted	AUC	F1 weighted	AUC	F1 weighted
Random Over Sampling	0.525	0.6212	0.529	0.617	0.512	0.558
SMOTE	0.547	0.6257	0.537	0.623	0.490	0.537
ADASYN	0.520	0.6150	0.522	0.609	0.4980	0.583
SVMSMOTE	0.522	0.6133	0.515	0.599	0.499	0.589
K-means SMOTE	0.503	0.599	0.504	0.600	0.505	0.589
Baseline imbalanced	0.5	0.576	0.5	0.576	0.504	0.589

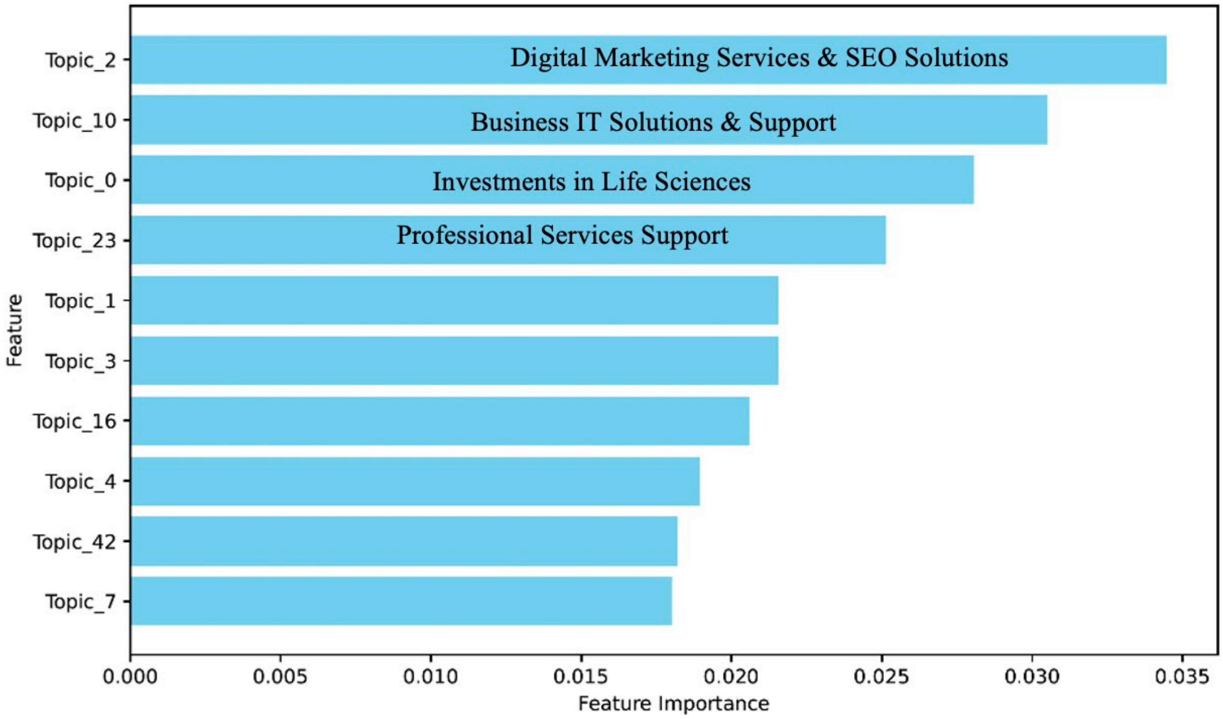

Figure 5.3 Top 10 Most Important Topics – Random Forest Results.

To deepen our understanding of these four key topics, we once again leverage the capabilities of the LLM, specifically using the BERTopic output alongside the Deepseek r-1 reasoning model

(Table 10). Our analysis combined two key elements: the top 20 keywords along with documents exhibiting a similarity to the topic centroid greater than 0.80. We perform this analysis on the first 4 topics. The results show that Topic_2 pertains to Digital Marketing Services & SEO Solutions, Topic_10 addresses Business IT Solutions & Support, Topic_0 relates to Investments in Life Sciences, and Topic_23 focuses on Professional Services Support.

Finally, Figure 5.4 presents the Partial Dependence Plot (PDP), a visualization tool commonly that shows how individual features affect model predictions while holding other variables constant. The PDP clarifies the relationship between a selected feature and the model's predicted outcome, averaging out the influence of all other variables. We illustrate the PDPs for the first four features, with the y-axis representing the predicted probability, starting around a baseline near 50%. The x-axis denotes the value of each respective topic. Each graph captures the marginal effect of the specific topic noted below it.

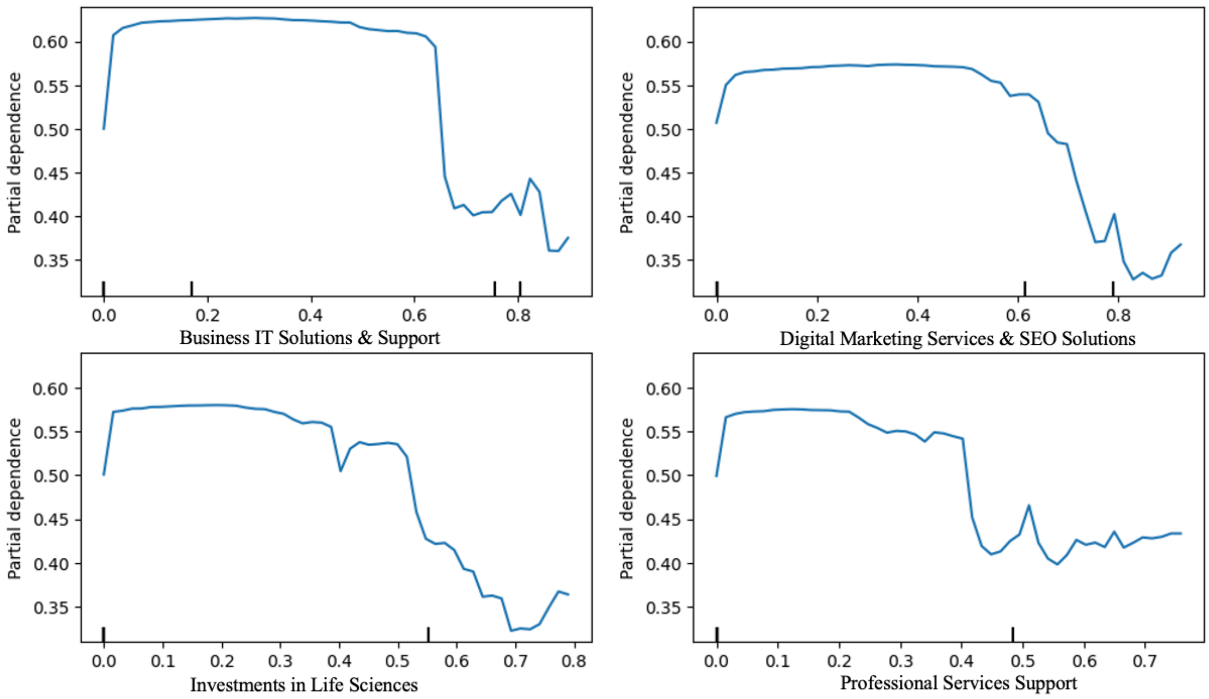

Figure 5.4 Partial Dependence plot of the 4 most important topics.

The results reveal that the four topics have clear negative marginal effects. For Professional Services Support and Business IT Solutions & Support, higher topic values correspond to lower probabilities of receiving private funding, with probabilities stabilizing around 40%. Topics related to Digital Marketing Services & SEO Solutions, and particularly the one related to Life Sciences shows even stronger negative impacts, significantly decreasing the probability to values as low as around 35%.

Summarizing the results we have that Companies with higher concentrations of topics related to Digital Marketing Services & SEO Solutions (Topic_2) or Business IT Solutions & Support (Topic_10) are significantly less likely to receive private funding. This suggests that businesses in these service-oriented sectors may face different funding landscapes compared to companies in other industries.

5.8 Discussion

Our analysis provides mixed support for Proposition 1, which states that companies that receive private funding have more likelihood of mentioning other funding on their websites. The logistic regressions show that the web-based funding signal is positive and significant for privately funded companies, partially validating our proposition. However, the strength of evidence varies significantly by funding type. For federal, provincial or territorial (FPT) funded companies, our findings reveal limited web-based funding disclosure, with descriptive statistics in Table B 2 showing that only 44 of these firms displayed funding information on their websites, creating a weak signal that limits our statistical power to detect meaningful effects. In contrast, results for privately funded companies are more robust and statistically meaningful, although we must note an important methodological consideration: the coefficient *PrivateFundig* partially mirrors the dependent variable *dFunding*, which records whether a company has already received private funding. This finding is theoretically meaningful and consistent with prior research. Extensive literature has documented the positive effects of public funding disclosure (e.g., Bellucci et al., 2023; Islam et al., 2018; Wu et al., 2020), while our results specifically support studies showing that disclosing private funding achievements serves as a credible signal of project legitimacy and increases the likelihood of attracting additional capital (Vanacker et al., 2020).

Our findings offer strong support for the second proposition that founder background signal is correlated with the capacity of attracting private fundings. The logistic regressions demonstrate that website signals for a founder's academic background (*WebFounderExpAcademic*) and, even more powerfully, for combined academic-and-industry experience (*WebFounderExpBoth*), both yield odds ratios significantly greater than one. This result directly confirms the core tenet of our proposition and aligns with the argument from Bellavitis et al. (2019), who identified founder education as a key signaling mechanism used by VCs to mitigate adverse selection in the pre-investment phase. Our specific finding that educational credentials, either alone or in combination with industry experience, enhance the likelihood of funding resonates with multiple studies. It supports the work of Piva & Rossi-Lamastra (2018), who highlighted the contribution of business education to fundraising success, and (Bhattacharyya & Subrahmanya, 2024), who found that VCs in India prioritize signals from prestigious educational backgrounds. Furthermore, the overall importance of these human capital signals in our analysis is consistent with Ko & McKelvie (2018), who discovered that education and prior experience are crucial for securing first-round financing. Our results empirically demonstrate that firms which actively disclose these credentials on their websites are correlated with a higher likelihood of securing private capital, effectively reducing the information asymmetry that investors face.

The third proposition tests whether companies receiving private funding are more inclined to display collaboration signals on their websites. We first identified websites that mention collaboration-related signals and then we used the retrieved text to calculate the intensity of collaboration mentioned. Logistic regressions reveal two key findings. First, in Reg2 none of the WebCollab variables are statistically significant. Second, when we examine sectoral moderation, several patterns emerge. In resource-based industries (mining, utilities and construction), intense collaboration signals are correlated to substantial benefits. Similarly, this intensity benefits leisure-oriented firms, slightly narrows the funding gap for low-tech services, has mixed effects in high-tech professional services, and is neutral for manufacturing. These results support the proposition that, compared with manufacturing, leisure-oriented firms are correlated to stronger collaboration signals. By contrast, professional, scientific and technical services firms are penalized when no collaboration is visible and perform even worse at moderate levels, though high-intensity collaboration partially offsets the disadvantage. This finding contradicts the original proposition yet aligns with work arguing that equity-based alliances may deter venture capitalists by diminishing their governance leverage (Jolink & Niesten, 2021) and introducing governance complexity (Grilli & Murtinu, 2014). Firms engaged in multi-stakeholder, equity-based collaborations may therefore appear riskier and less attractive to VC investors.

Finally, proposition 4 posits that there are different topics on the website of companies that receive and do not receive private funding. Our approach, developed using BERTopic and supervised methods, allows us to identify four topics that are used by the supervised method to discriminate between companies. These four topics are related to the digitalization of the company and specifically include: Topic_2 pertains to Digital Marketing Services & SEO Solutions, Topic_10 addresses Business IT Solutions & Support, Topic_0 relates to Investments in Life Sciences, and Topic_23 focuses on Professional Services Support. Companies that offer digital marketing and SEO solutions and IT business solutions and support are correlated with lower chances of receiving private fundings, with these chances decreasing by approximately 20-30% when the topics increase their presence on their websites. Similarly, companies that invest in Life Science and professional service support are correlated with lower chances of receiving private funding when increasing the presence of these topics on their websites. These findings indicate involvement in three different sectors and are aligned with the results that we find in Reg3, as companies in the sectors related to Professional, Scientific & Technical services, (NAICS 54) are correlated with less likely to obtain private funding (Table 5.5). This convergence between our bottom-up topic modeling approach and the top-down approach strengthens our conclusions. Thus, using the bottom-up approach, we identify the topics that are more associated with companies that do not receive private funding. These findings are particularly noteworthy given the current Canadian venture capital landscape. In the 2024 the information and communications technology (ICT) sector led the way, attracting CAD $4.49 billion, and life sciences followed with CAD $1.38 billion, with cleantech securing CAD $1.07 billion in investments[46]. Investment in ICT is more than triple the Life science investment, which represents the second-largest sector for VC investment. However, it's important to note that Canada's ICT umbrella encompasses computer hardware/software & services, internet software & services, ecommerce, electronic & semiconductor, mobile & telecom and services that are IP-heavy. While digital-marketing agencies and IT support providers do raise money, they represent a smaller subset within the broader ICT category and are not the primary targets of major VC investments, which explains why our analysis identifies these specific service-oriented topics as associated with lower funding probability, even though the broader ICT sector attracts significant investment.

[46] https://www.cvca.ca/insights/market-reports/q4-2024/

5.9 Conclusion

Our research goal is to develop a real-time analysis tool for funding trends that serves policymakers and strategy managers employing both top-down and bottom-up approaches. The top-down approach uses AI-derived web indicators, validated against Crunchbase data, to replicate prior literature findings. The bottom-up approach uses topic modeling to identify website content patterns that distinguish Canadian companies receiving private funding. We tested these approaches through a pilot study of Canadian ventures founded between 2020-2024, a period of significant funding activity. Drawing on signaling theory, we developed four propositions based on evidence that corporate websites reveal firms' strategic choices (e.g., Cruciata et al., 2024; Jiang et al., 2023). Our findings contribute to signaling theory literature by demonstrating the breadth of information retrievable from websites and enabling stakeholders to identify private funding patterns.

Our pilot study shows that a combined top-down pipeline based on the RAG and the bottom-up topic modeling can mine unstructured website text at scale and reproduce, or even extend, insights drawn from structured datasets. Because websites are updated continuously, the method provides more timely signals of funding, collaboration, founder background and other factors, giving scholars, investors and policymakers an earlier view of venture dynamics. To the best of our knowledge, this is the first paper to deploy both approaches together and one of the first to use a RAG framework and a LLM to create targeted web indicators. In terms of top-down approach, the closest methodology is the one used by Cruciata et al., (2024). They employed the zero-shot text classification for a similar purpose of building web-indicators. However, our method improves the web indicator quality thanks to the RAG approach combined with the LLM offering a controlled output and deeper understanding of complex website corpora. From a signalling-theory perspective, our findings confirm that the analyzed signals (funding announcements and founders' human capital and collaboration) affect a firm's likelihood of securing private capital, demonstrating the power of websites as a data source and, to our knowledge, marking the first time that signals for private funding have been systematically extracted and analyzed from corporate websites at this scale.

While this study establishes a novel framework for analyzing web-based signals, its findings also delineate the scope of the pilot and should therefore be viewed alongside their inherent limitations, which provide a foundation for the next phase of research.

First, our analysis is constrained by the characteristics of the sample and the resulting sparsity of specific web-based signals. Although our dataset comprised several thousand companies, the number of firms exhibiting the signals of interest, particularly those generated through the RAG framework, was modest. For instance, a clear signal for prior government funding was detected on only 44 company websites. This limited statistical power may have prevented us from detecting significant relationships for certain variables, potentially leading to Type II errors.

Furthermore, the modest predictive power of our bottom-up supervised model (AUC 0.547) is a key finding. This suggests that a company's general website text is likely a "noisy channel" for predicting private funding, where the signal-to-noise ratio is low. While we successfully identified influential negative topics, the overall weakness implies that only precise linguistic patterns serve as reliable indicators.

Finally, as a pilot study, this research was bounded by the scope of our data access and the specific time frame (2020–2024). To ensure the generalizability and temporal stability of our results, further validation is essential.

Future works should build directly on the findings and limitations of this pilot study. The first priority is to expand the dataset significantly to assemble a critical mass of observations for each web-based indicator. This includes covering a larger cross-section of industries and, crucially, replicating the study in different venture capital markets beyond Canada to test the generalizability of our model and build a more comprehensive tool for monitoring global private funding dynamics.

Methodologically, future research should focus on refining NLP techniques to better isolate signals from the noisy environment of general website text. This could involve developing more sophisticated classifiers or exploring different architectures designed to enhance the signal-to-noise ratio. Such work could also contribute to the theoretical understanding of these noisy communication channels within the field of signalling theory.

Finally, our findings lay the groundwork for future research aimed at moving beyond correlational insights toward a more robust, causal-informed understanding. While this study identifies strong associations, a key limitation is the difficulty in controlling for unobserved firm characteristics and establishing temporal precedence. To address this, future research could employ a panel data approach, tracking firms longitudinally from their inception through subsequent funding rounds. Such a study should aim to collect not only web-based signals but also other pertinent data, such as products development and revenue metrics. By modeling these variables together, we could more effectively isolate the marginal contribution of web-based signaling and test whether its evolution is a meaningful predictor of funding outcomes, rather than merely a reflection of other underlying business developments. This would allow for a more nuanced understanding of how, and to what extent, these signals function within the complex information environment faced by investors.

CHAPTER 6 The Discrepancy Between Public Disclosure and Private Reporting: Validating Web-Based Innovation Indicators Against Survey Data

6.1 Introduction

Building indicators of sustainable innovation for near real-time monitoring remains a significant challenge for policymakers and industry analysts. The previous two chapters began to address this by developing novel environmental and private funding indicators, though the studies yielded mixed results. We had great results concerning the environment while the private fundings indicators results were not at the same level.

This chapter present a comparison between aggregated results from Statistics Canada's Survey of Innovation and Business Strategy (SIBS) [47] and indicators from the websites of a subset of the surveyed companies. We leveraged the proved methodology of Article 2 to determine if our approach can extract reliable, innovation and environmental related signals to monitor industry trends. The validity of these web-derived signals, which represent a crucial point of this research, will be tested against data from the SIBS. This survey was selected for two primary reasons. First, its sample, designed by Statistics Canada, provides a representative cross-section of Canadian companies against which to benchmark our findings. Second, its inclusion of direct questions on sustainable innovation allows us to perform a crucial deep-validation. The results provide an insight into the promising use of corporate websites as a source of data in this domain.

6.2 Methodology

6.2.1 Data

This research employs two primary datasets to achieve its objective:

1. SIBS 2019, survey from Statistics Canada with data collected between December 2019–March 2020 (electronic questionnaire plus telephone follow-up) on the period spanning 2017 to 2019 (see section 3.3);

[47] SIBS aggregate results are available at the following links: https://www150.statcan.gc.ca/n1/daily-quotidien/210426/dq210426a-eng.htm and https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2021020-eng.htm. Note: these results are weighted for the no responses.

2. The Wayback Machine, the digital archive of the World Wide Web, that allows users to retrieve web pages across different points in time.

For the purpose of this project, we leverage the Wayback Machine and the crawler explained in Chapter 3, to extract the textual content of the websites. From the original list of public URLs that Statistic Canada provided us, which consists of the 10,407 websites of the enterprises surveyed, we removed the websites that contained mistakes or that were duplicated, resulting in a list of 8,081 websites. Figure 6.1 depicts the further steps that lead to the final sample. Starting from this list of 8,081 websites of the surveyed companies in SIBS, we first retrieved companies' websites from the Wayback Machine once for each year taken into consideration by the survey (2019, 2018, 2017). Using this time frame, we obtained 19,740 snapshots of the original sample. Given that SIBS ask questions based on these three years as an aggregated period, we merged the website content extracted from the Wayback Machine, resulting in 6,726 websites.

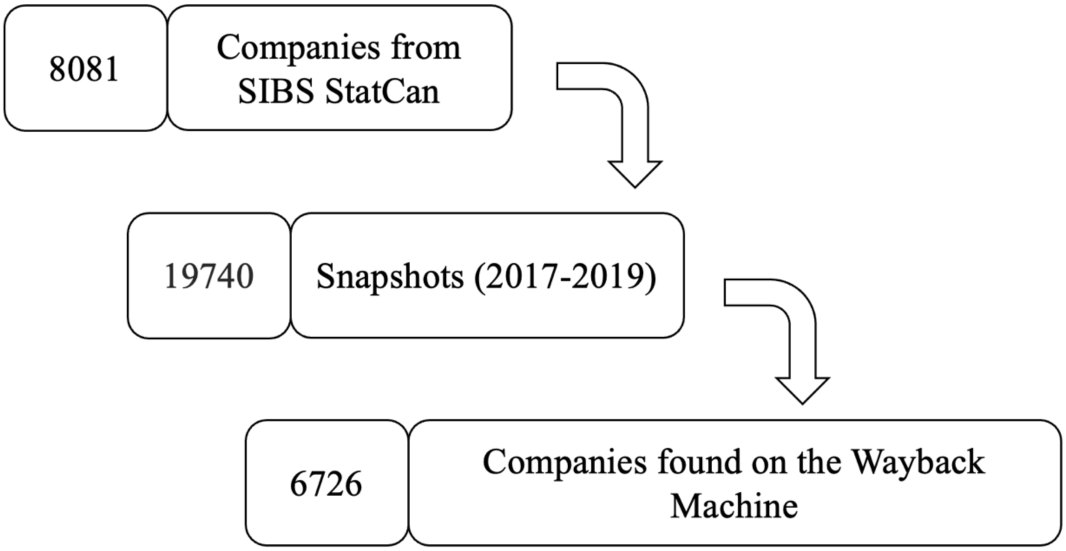
8081
Companies from SIBS StatCan
19740
Snapshots (2017-2019)
6726
Companies found on the Wayback Machine

Figure 6.1 Dataset Construction Process

6.2.2 Methods

Figure 6.2 presents the methodology used for this research. Following the approach established in our previous study (Article 2, see section 3.3.4), we applied the advanced RAG framework (Y. Gao et al., 2023) using the same LLM and the same set of hyperparameter, as these had already been validated for company website data and our current study uses the same type of data source (Table 6.1).

After creating the corpora for the research using the retrieved websites, we selected the questions from the survey that could be answered using the available information. Each SIBS question was evaluated to determine whether the required information could be extracted from company websites. Due to limited availability of relevant information on the corporate websites retrieved from the Wayback Machine, several question categories were excluded from our analysis. These included questions about expenditures on innovation activity, sales, number of employees in Canada and abroad, obstacles faced by the company, main markets and competitive landscape, and employee skills and training.[48] Instead, we focused on four key areas where website signals was more readily available: innovation for environmental benefit, cooperation for innovation, introduction of new or improved products or services, and government support programs for innovation-related activities. Given the nature of corporate website content and language, we adapted our analysis approach to align with the SIBS survey questions rather than forcing the website information into incompatible formats. For example, while SIBS asks specific collaboration questions, we broadened our search to capture general partnership activities mentioned on websites, while for questions about grants, environmental products, and new products or services we decide to format them to match the SIBS framework.

We then analyzed the question types to create ad hoc prompts for the LLM DeepSeek r-1 32B (DeepSeek-AI et al., 2025). As for Article 2, this step requires significant time and effort since the prompt wording affects the outcomes. We therefore performed extensive manual checks and revisions to ensure that the prompts (see Appendix for the final wording) effectively guide the model toward the desired structured results. Finally, we designed prompts with three main components: a binary field for yes/no responses, a multiple-choice field aligned with the corresponding SIBS question, and an “explanation” field where the model justifies its answers. The “explanation” field and the think output of the DeepSeek r-1 are used for the manual evaluation of the answer.

[48] These kinds of information are generally provided in metadata available through governments or firm financial repositories such as Bureau Van Dijk Orbis or Dunn and Bradstreet. In Canada, however, only public companies are obliged to divulge such information.

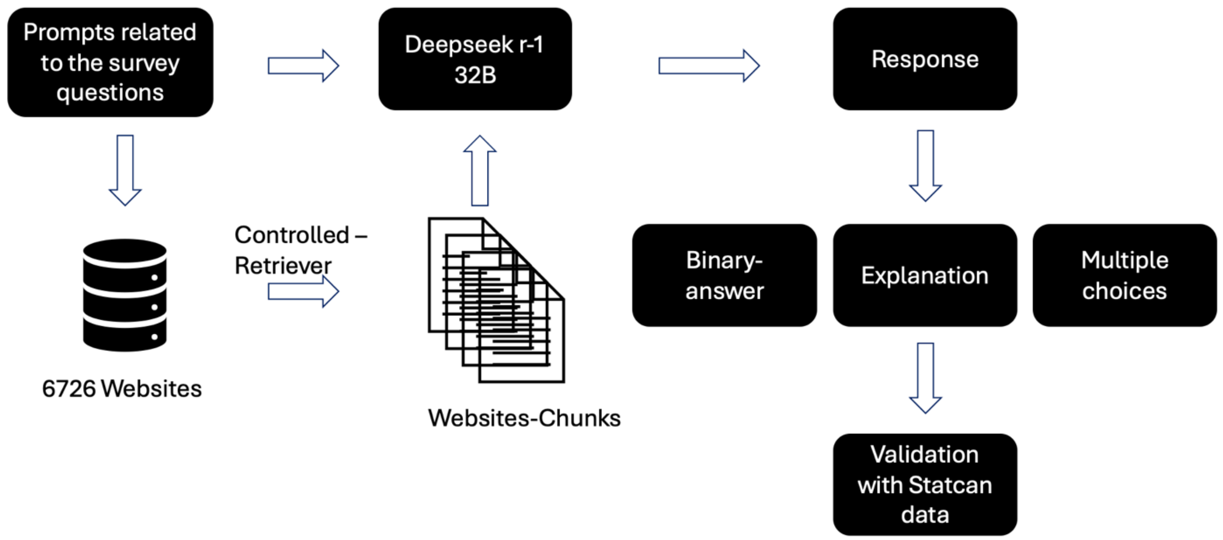

Figure 6.2 Pipeline RAG

Table 6.1: Hyperparameters used in the RAG

Parameters	Meaning	Values chosen
Chunk size	Maximum number of characters or tokens allowed in a single segment of text when breaking down documents.	1,024
Top k	number of most relevant chunks of text that the retriever will return based on their similarity to the user's query.	10
Temperature	This parameter controls the randomness or creativity of the model's output during the text generation process. Lower temperature (e.g., 0.1 or 0.2) makes the model more deterministic.	0
Embedding model	This model is responsible for the creation of the Document store.	BAAI LLM-Embedder
Context Windows	The amount of text measured by tokens that the LLM can process at once to generate a response.	3,900

6.3 Results

The following paragraph present the comparison between the weighted aggregate results of SIBS and the signals captured from the companies' websites. However, statistical testing suggests caution in interpreting these observed differences as definitive patterns. Thus, we present an indication of the actual results given that we have only a subset of the companies surveyed on SIBS.

6.3.1 New product introduction

Table 6.2 suggests a discrepancy between the innovation activities reported in the SIBS and what companies communicate on their public websites. While the SIBS data, relevant to many businesses, reveals that a majority of companies (52.7%) introduced a product or service innovation, a comprehensive website analysis found that only a small fraction (4.61%, representing 310 companies) made any mention of such activities. This divergence extends into the very nature of the innovation being discussed. Among the few companies that do mention the introduction of an innovative good or service on their websites, there is a clear emphasis on new products (38.39% of mentions) and a combination of both products and services (36.13%), with new services being the least common topic (25.48%). In contrast, the SIBS findings suggest potential differences in internal business priorities, indicating a stronger strategic focus on new service innovation (37.90%) over new product innovation (32.20%).

Table 6.2: Announced product or service innovations on the websites versus the percentage that reported such activities in the SIBS survey, including a breakdown by the type of innovation.

New product service	Companies' websites	SIBS
Product	38.39% (119)	32.20%
Both	36.13% (112)	-
Service	25.48% (79)	37.90%
Total	4.61% (310)	52.70%
N. obs.	6,726	10,407

Source: Survey of Innovation and Business Strategy (SIBS): Interactive Dashboard: https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2021020-eng.htm

6.3.2 Introduction of environmental product

Table 6.3 presents an observed disparity in environmental innovation adoption between survey data and website reporting. While nearly half (49.8%) of businesses in the broader SIBS survey introduced environmental innovations between 2017-2019, only 8.91% of companies analyzed through their websites reported similar activities.

The SIBS survey reveals substantial innovation impact: businesses cited significant resource efficiency benefits in production, over a third (35.8%) noted consumer benefits, and among those with consumer-related benefits, three in five (60.3%) stated their innovation was new to the market[49]. In contrast, website's analysis found much lower reported rates. However, among the small cohort that does report environmental innovations online, companies tend to be creators rather than adopters, with 43.24% introducing new products, 34.06% using existing solutions, and 22.04% doing both.

Table 6.3: Rate of reporting on environmental innovations on company websites with data from the SIBS survey regarding product or process innovations with environmental benefits, and details whether the web-disclosed innovations were newly introduced by the company or adaptations of existing solutions

Environmental innovation	Companies' websites	SIBS
Introduce	43.24% (259)	-
Used	34.06% (204)	-
Both	22.04% (132)	-
Not specified	0.67% (4)	-
Total	8.91% (599)	49.30%
N. obs.	6,726	10,407

Source: Survey of Innovation and Business Strategy (SIBS) 2019: Table: 27-10-0149-01 – Product or process innovations with environmental benefits, by industry and enterprise size: https://www150.statcan.gc.ca/t1/tbl1/en/cv.action?pid=2710014901

6.3.3 Government programs

Table 6.4 shows an apparent disparity between SIBS survey data and website analysis regarding public funding usage. While SIBS indicates that approximately 30% of Canadian businesses utilized government programs for innovation between 2017 and 2019, only 1.25% of companies in

[49] https://www150.statcan.gc.ca/n1/daily-quotidien/210426/dq210426a-eng.htm, 14/06/2025

our website analysis publicly cited any form of government support. This is very much in line with what article 2 highlighted.

The types of support mentioned also appear to differ between sources. SIBS found that 38.3% of companies used training and hiring programs for innovation activities. In contrast, among the small number of companies that did mention government support on their websites, grants and subsidies were most commonly cited (38.08%), followed by tax incentive programs (35.71%), public loans (17.86%), and government-funded training programs (8.33%).

Table 6.4: Utilization of government programs reported in the SIBS survey and the acknowledgment of such support on company websites, with a breakdown of the types of funding mentioned.

Funding type	Companies' websites	SIBS
1. Grants + subsidies	38.08% (32)	-
2. Loans	17.86% (15)	-
3. Tax incentive programs	35.71% (30)	-
4. Government-funded training	8.33% (7)	38.3%
Total	1.25% (84)	30.5%
N. obs.	6,726	10,407

Source: Survey of Innovation and Business Strategy (SIBS) 2019: report done by StatCan published at the following link: https://www150.statcan.gc.ca/n1/daily-quotidien/210426/dq210426a-eng.htm

6.3.4 Collaboration

An analysis of business partnerships reveals an apparent gap between collaborations reported in SIBS and those mentioned on companies' websites. Table 6.5 shows that while SIBS data indicates that 17.80% of businesses engage in collaborations, only 5.52% of companies (371 out of 6,726) mention such activities on their websites.

The datasets align on only one point: both identify clients or customers from the private sector as the most common partners, cited at nearly identical rates (32.61% on websites vs. 33.40% in SIBS). Beyond this similarity, the data appears to diverge dramatically. Websites appear to significantly underreport key strategic partnerships, particularly operational relationships. Suppliers are mentioned by only 18.60% of companies on websites versus 53.00% in SIBS, while parent or affiliated companies appear on 10.24% of websites compared to 43.00% in the survey. Knowledge-based collaborations seem to be even less visible online: consultants (3.23% vs. 16.90%) and universities (1.89% vs. 15.60%) are rarely mentioned on websites despite their substantial presence in SIBS data.

Table 6.5: Frequency and type of business cooperations mentioned on corporate websites against those reported in the SIBS survey.

Cooperation	Companies' websites	SIBS
Clients or customers from the private sector	32.61% (121)	33.40%
Suppliers of equipment, materials, components or software	18.60% (69)	53.00%
Other co-operation partners	11.32% (42)	9.30%
Parent, affiliated or subsidiary businesses	10.24% (38)	43.00%
Competitors or other businesses in the sector	8.09% (30)	18.80%
Clients or customers from the public sector	5.12% (19)	9.00%
Non-profit organizations	4.85% (18)	8.00%
Consultants and commercial laboratories	3.23% (12)	16.90%
Government, public or private research institutes	2.16% (8)	10.20%
Universities, colleges or other higher education institutions	1.89% (7)	15.60%
No specified	1.89% (7)	-
Total	5.52% (371)	17.80%
N. obs.	6,726	10,407

Source: Survey of Innovation and Business Strategy (SIBS) 2019: Interactive Dashboard: https://www150.statcan.gc.ca/n1/pub/71-607-x/71-607-x2021020-eng.htm

6.4 Discussion and Conclusion

Although the aggregate results are not statistically significant[50], the great difference in percentage suggests a disconnect between what companies report in confidential surveys and what they disclose on their public websites. The divergence could be explained by a confluence of three main factors: the inherent self-reporting biases of surveys, the strategic communication thresholds that govern corporate websites, and the differential consequences of providing false information. These elements demonstrate that websites and surveys capture fundamentally different dimensions of business activity. One reflects a curated public narrative while the other provides a comprehensive, confidential inventory.

The first factor driving this discrepancy is the self-reporting bias (Antolín-López et al., 2016) inherent in surveys, which is heavily influenced by the respondent's comprehension of key terms. The definition of "innovation," for instance, can be highly subjective. While Statistics Canada provides a formal definition from the Oslo (2005) Manual, a respondent from a traditional manufacturing firm might interpret it differently than a tech startup CEO, one might count a minor process improvement while the other only considers disruptive new products. This variability in understanding can lead to inconsistencies in the data, making direct comparisons between firms and sectors challenging.

The second and arguably most important factor is that a corporate website functions as a strategic marketing tool, not a neutral repository of information. The decision to publish content is an active, goal-oriented action designed to manage public perception. Consequently, the threshold for what is deemed worthy of mention is extremely high. For instance, it is possible that the digital signals for product innovations, in the form of minor or incremental changes, while technically innovations, are often omitted because they fail to serve a compelling marketing narrative.

[50] Such mean comparison tests cannot be performed to compare data reported by Statistics Canada and websites from a smaller sample of the firms that responded to SIBS. The descriptive statistics reported by the organisation are weighted by sampling weights, which are not available to the researcher. "The response values for sampled units were multiplied by a final weight in order to provide an estimate for the entire population. The final weight was calculated using a certain number of factors, such as the probability for a unit to be selected in the sample, and adjustment of the units that could not be contacted or that refused to respond. Using a statistical technique called calibration, the final set of weights is adjusted in such a way that the sample represents as closely as possible the entire population." (see the data source and methodology section of https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&Id=1260908#a2).

This explains why government training programs, a common form of support reported in SIBS, are rarely mentioned, while more marketable grants and subsidies appear more often. This principle of strategic curation by companies is perfectly illustrated by the collaboration data. The reporting of "Clients or customers" is the one area where websites and SIBS align. This apparent anomaly reinforces the argument: companies selectively publicize partnerships that serve as powerful public testimonials while keeping deeper, more operationally sensitive collaborations with suppliers or universities, which are core to their competitive strategy, internal and away from their public-facing narratives. Another interesting aspect relates to the web indicators for environmental products, which are the most populated category of website indicators. In this case, there is also a precise goal for reporting the use or introduction of environmentally sustainable products. Companies are ensuring their environmental efforts are seen by stakeholders with the aims of achieving their sustainability goals, diminishing pressure to "be green", and improving their image to potentially attract new customers.

The third factor involves how the consequences of providing false information differ vastly between the two contexts. When a company includes false statements on its website, it engages in direct public deception aimed at its most critical stakeholders. The risks are immediate and market-driven: loss of public trust, brand reputation damage, customer backlash, and severe legal penalties for false advertising. For publicly-traded firms, this can trigger sharp drops in stock prices. Conversely, misrepresentation on the confidential SIBS survey is directed at a statistical agency, making the harm indirect and systemic. While it is an offense under the Statistics Act, the penalty is a smaller fine of up to $500. The more significant, though less visible, consequence is the corruption of the national data pool used for policy-making and research purposes.

These three factors collectively hint why SIBS surveys and corporate websites seems to present different pictures of corporate innovation. According to our view, this study suggests that the primary value of website data lies not in its potential to substitute for survey metrics, but in its unique ability to reveal a firm's strategic communication.

CHAPTER 7 DISCUSSION AND CONCLUSION

7.1 Discussion

This thesis addresses critical yet fundamentally different measurement challenges within traditional indicators for innovation and sustainability. Sustainability faces a crisis of consensus due to a lack of standardized definitions. Thousands of unique metrics fragment the field, making cross-industry comparisons difficult. Furthermore, the reliability of these measures is questionable, as companies largely self-report their sustainability data without external auditing. Innovation confronts different but equally significant limitations. While established standards like patents and large-scale surveys exist, practical constraints hamper them: they are often too slow, costly, or lack the granular detail that modern policy and investment decisions need. Additionally, they fail to capture the increasingly vital environmental, and social values of modern enterprise.

In response to these distinct gaps, researchers have turned to company websites as a rich, real-time data source, though possible methodological shortcomings have hampered initial efforts, yielding mixed or weak results. Recognizing the potential of advanced Natural Language Processing (NLP) to overcome these hurdles, this research pursued the central question:

Do website signals represent valid and reliable measures of the underlying innovation and sustainability quality that organizations are attempting to communicate?

To systematically explore this question, the research advanced and tested a series of propositions designed to probe the validity and meaning of web-based signals from multiple angles. This inquiry was structured across two papers and the project described in Chapter 6, each addressing a distinct facet of corporate signaling.

Article 1 focused on environmental signals, beginning with construct validation. It first tested the direct correlation between web-based indicators and the objective, third-party B-Lab environmental indices (Proposition 1). Building on this, it further investigated how this relationship was moderated by key contextual factors, such as a firm’s location, size, and industrial sector (Propositions 2, 2m, 3, 3m, 4, and 4m), to understand the conditions under which signals are most credible.

Article 2 shifted the investigation to innovation-related signals within private funding dynamics. The paper explored whether signals on company websites could distinguish between funded and unfunded firms. Overall, this study investigated whether web-based signals align with literature-documented signals that attract capital, examining how firms strategically communicate quality to investors. Specifically, the study tested whether firms that have secured private financing more prominently showcase prior funding received, founder experience and key collaborations on their websites.

Finally, Chapter 6 compares website indicators of innovation and environmental efforts with the SIBS results, aiming to evaluate the convergence between these two measures of corporate performances. After examining the survey questions and available website information, the research focuses on the analysis of four primary areas. Our study, therefore, examines website content for information corresponding to these key topics: innovation for environmental benefit, cooperation for innovation, the introduction of new or improved products or services, and government support for innovation-related activities.

Taken together, these findings paint a nuanced picture that aligns with the differing disclosure pressures that the literature identifies. On the one hand, environmental practices face increasing stakeholder scrutiny, raising the reputational risks for firms engaging in "greenwashing" while creating competitive advantages for those who genuinely invest in sustainable innovation. In this high-scrutiny context, as Article 1 found, website content about sustainability is more likely to act as a mirror, credibly reflecting verifiable, third-party validated performance. This conclusion is further supported by the findings in Chapter 6, which show that indicators related to environmental products used in business processes or sold are the most frequently communicated, underscoring the strategic importance of this signal. On the other hand, the results suggest that firms tend to do not disclose forward-looking economic and strategic information. As Article 2 demonstrated, for the companies that decide to disclose information, the website functions more as a lens, projecting the qualities and capabilities, such as founder experience and funding momentum, that are most likely to attract capital. This divergence underscores that web signals are not monolithic; the specific strategic risks, stakeholder pressures, and corporate objectives of the domain being communicated shape their meaning and veracity. While Chapter 6 confirms this thematic divergence, it also reveals a quantitative discrepancy between the number of firms reporting an activity to SIBS and those communicating it on their websites. This suggests that a website signal, when present, is a reliable indicator of strategic intent highlighting the website's role as a curated corporate communications tool.

The main contribution of this research to Signalling Theory is the argument that the reliability of a digital signal is not inherent to the signal itself, but is contingent on the strategic context in which it is deployed. This thesis demonstrates that the pressures, risks, and goals of a specific domain (like sustainability vs. fundraising) fundamentally change how and what organizations communicate and, therefore, how their signals should be interpreted.

7.2 Conclusion

This thesis makes three distinct contributions: theoretical, methodological, and practical. The primary theoretical contribution advances Signalling Theory by addressing the central challenge of confirming whether a signal validly and reliably measures the signaler's underlying quality (Connelly et al., 2011). Signalling Theory provides the appropriate lens for assessing information asymmetry between firms and stakeholders, the core problem this research tackles. The findings validate corporate websites as a source for credible signals in two distinct domains. In sustainability, Article 1 confirmed that environmental signals on websites reflect tangible performance, validated against the B-Lab index. Extending into innovation finance, Article 2 demonstrated, to our knowledge for the first time, that signals of founder human capital and funding history credibly correlate with private capital acquisition. By validating signals across these critical corporate arenas, the thesis offers preliminary evidence addressing a core question within Signalling Theory.

Chapter 6 suggests another interesting finding: although a general discrepancy between a company's self-reported web-indicators and the formal SIBS indicators, there are promising findings related to environmental products and customer collaboration. The nature of these findings supports the conclusion that a corporate website primarily functions as a strategic tool. This implies that the decision to publish content is an active, goal-oriented action designed to influence public perception.

Concerning the methodological contribution this research makes significant methodological advances by introducing and validating novel NLP-based approaches. Article 1 introduces an efficient and generalizable framework using Zero-Shot Text Classification (ZSTC), enabling the creation of performance indicators without costly data annotation or pre-processing. The effectiveness of this approach is demonstrated by its ability to explain over 57% of the variance in the B-Lab environmental index, representing a significant improvement over previous methods. Building on this foundation, Article 2 further advances the methodological frontier by developing a novel pipeline that combines retrieval-augmented generation (RAG) with topic modeling. This integration improves indicator quality by leveraging language models' deeper contextual understanding of complex website text. Both methodologies share a key advantage in their ability to leverage timely website data, providing a valuable supplement to traditional, time-lagged survey methods.

Beyond theoretical and methodological contributions, the findings presented in this thesis have direct practical implications for multiple stakeholder groups, offering a toolkit for more agile and evidence-based decision-making. Government agencies and policymakers could leverage these methods to monitor entrepreneurial and sustainability trends in real-time, enabling much faster feedback loops on policy impact, such as Canada's 2030 Emissions Reduction Plan, than what is achievable through biennial surveys.

In the private sector, investors can gain distinct advantages through the ability to extract timelier signals, providing earlier and more nuanced views of venture dynamics by analyzing signals of founder background, collaboration intensity, and funding momentum. Similarly, managers can utilize these indicators to benchmark their public-facing communication and understand how stakeholders perceive their strategic priorities relative to competitors.

The research community also benefits significantly from this work, as it opens new avenues for studying innovation and sustainability dynamics at scale. These methods directly address well-documented limitations of traditional survey data while providing a validated path toward integrating web data with conventional measures, ultimately enriching our understanding of firm behavior across multiple domains.

7.3 Limitations

However, this approach introduces its own fundamental challenge: all data derived from websites is inherently voluntary and self-reported, raising critical questions of validity and bias that must be addressed for both innovation and sustainability assessments.

While this thesis presents significant advancements, it is crucial to acknowledge its limitations, which stem from a fundamental challenge: data derived from websites is inherently voluntary and self-reported, raising critical questions of validity and bias**.** A primary challenge flowing from this is the difficulty in distinguishing genuine underlying qualities from a firm's marketing sophistication in its website content. The reliance on self-reported data, even when validated, raises concerns about issues like "greenwashing" or performative "innovation-related signalling", making it difficult to systematically separate authentic signals from aspirational communication. The representativeness and external validity of the samples also pose limitations, as the B-Corps studied are inherently sustainability-focused and the 2020-2024 Canadian startup cohort reflects a specific temporal and geographic context, which restricts the direct generalizability of the findings and the performance of NLP models across different languages, sectors, or firm sizes. Different is the problem related to the project in Chapter 6. Not having access to the full data from the SIBS, the project focuses in comparing the results of the web-indicators with the aggregated SIBS data. This results in a difficult comparison, with the web-indicators not representative of the full sample.

Furthermore, the methodological sophistication that provides the thesis's strength also creates a practical barrier, as the implementation of advanced NLP pipelines requires significant technical expertise, potentially limiting adoption by intended end-users like policymakers or smaller investment firms that may lack the necessary technical competences.

7.4 Future research

The initial findings suggest that firms strategically manage their public innovation signals, but the underlying drivers and boundary conditions of this behavior remain underexplored. The next phase of this research should aims to develop a comprehensive model of corporate signaling in the digital age by: validating the 'mirror and lens' framework across diverse economic and cultural contexts, increasing the sample to make the signals results reliable, and integrating multi-modal data to capture a holistic view of a firm's strategic communication.

To validate the "mirror and lens" framework, future research should test it across different contexts. For instance, studies could investigate whether the "mirror" effect holds for other areas with high stakeholder scrutiny and third-party verification, such as workplace safety, or diversity and inclusion metrics. Concurrently, research should explore the nuances of the "lens" effect for innovation signals. This would involve examining the specific factors that determine which "lenses" firms choose to project, such as their stage of development (seed vs. growth stage), the competitive landscape, or the specific demands of their target investors, all of which likely influence their signaling strategy.

The reliability of the signals identified in this thesis is constrained by certain limitations in the three core contributions. On the one hand, Article 1 and Article 2 are two pilot studies tested with small, uniform samples. A large-scale international study will test whether the environmental and innovation signals maintain the same reliability in a larger analysis. This step is necessary before using the frameworks developed for the real-time analysis of companies.

On the other hand, Chapter 6 presents results that are only partially representative of the full project as the comparison with SIBS data was limited to aggregated results. Securing access to firm-level microdata from innovation surveys such as SIBS, would enable a direct, one-to-one comparison between what a firm reports in a confidential survey and what it publicly signals on its website. This would provide definitive insights into strategic disclosure. The sample characteristics will shed light on the several types of communications related to the size and sectorial differences, as aligned in Article 1 and Article 2.

Finally, an integration with other types of sources is necessary to have a full picture of the strategic communication of a company. Data such as published reports about sustainability, or other sources as Common Crawl[51] can improve the study by integrating the website and reducing the missing data from the Wayback machine. In this vein, the use of LLM agent that searches the web for news about companies could be a resource to leverage for the next projects.

[51] https://commoncrawl.org

REFERENCES

1. Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis*, *18*(4), 507–529. https://doi.org/10.1080/17421772.2023.2193222
2. Adams, R., Jeanrenaud, S., Bessant, J., Denyer, D., & Overy, P. (2016). Sustainability-oriented Innovation: A Systematic Review. *International Journal of Management Reviews*, *18*(2), 180–205. https://doi.org/10.1111/ijmr.12068
3. Aerts, K., & Schmidt, T. (2008). Two for the price of one?: Additionality effects of R&D subsidies: A comparison between Flanders and Germany. *Research Policy*, *37*(5), 806–822.
4. Afeltra, G., Alerasoul, S. A., & Strozzi, F. (2023). The evolution of sustainable innovation: From the past to the future. *European Journal of Innovation Management*, *26*(2), 386–421.
5. Agarwal, R., & Dhar, V. (2014). **Editorial** —Big Data, Data Science, and Analytics: The Opportunity and Challenge for IS Research. *Information Systems Research*, *25*(3), 443–448. https://doi.org/10.1287/isre.2014.0546
6. Aghion, P., Bond, S., Klemm, A., & Marinescu, I. (2004). Technology and financial structure: Are innovative firms different? *Journal of the European Economic Association*, *2*(2–3), 277–288.
7. Aguilar-Fernández, M. E., & Otegi-Olaso, J. R. (2018). Firm size and the business model for sustainable innovation. *Sustainability*, *10*(12), 4785.
8. Aguilera-Caracuel, J., & Ortiz-de-Mandojana, N. (2013). Green innovation and financial performance: An institutional approach. *Organization & Environment*, *26*(4), 365–385.

9. Ahi, P., & Searcy, C. (2015). An analysis of metrics used to measure performance in green and sustainable supply chains. *Journal of Cleaner Production*, *86*, 360–377.
10. Ahmad, W., & Zhang, Q. (2020). Green purchase intention: Effects of electronic service quality and customer green psychology. *Journal of Cleaner Production*, *267*, 122053. https://doi.org/10.1016/j.jclepro.2020.122053
11. Aimiuwu, E. E., & Bapna, S. (2011). *Measuring innovation using business intelligence dashboards*. https://aisel.aisnet.org/amcis2011_submissions/414/
12. Akerlof, G. A. (1970). 4. The market for 'lemons': Quality uncertainty and the market mechanism. *Market Failure or Success*, *66*. https://www.elgaronline.com/downloadpdf/edcollbook/1843760258.pdf#page=82
13. Akturan, U. (2018). How does greenwashing affect green branding equity and purchase intention? An empirical research. *Marketing Intelligence & Planning*, *36*(7), 809–824. https://doi.org/10.1108/MIP-12-2017-0339
14. Alemany, L., & Marti, J. (2005). Unbiased estimation of economic impact of venture capital backed firms. *Available at SSRN 673341*. https://papers.ssrn.com/Sol3/papers.cfm?abstract_id=673341
15. Alessandrini, P., Presbitero, A. F., & Zazzaro, A. (2010). Bank size or distance: What hampers innovation adoption by SMEs? *Journal of Economic Geography*, *10*(6), 845–881.
16. Allouche, J., & Laroche, P. (2005). A meta-analytical investigation of the relationship between corporate social and financial performance. *Revue de Gestion Des Ressources Humaines*, *57*, 18.

17. Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to 'webometrics.' *Journal of Documentation*, *53*(4), 404–426. https://doi.org/10.1108/EUM0000000007205
18. Anand, A., Argade, P., Barkemeyer, R., & Salignac, F. (2021). Trends and patterns in sustainable entrepreneurship research: A bibliometric review and research agenda. *Journal of Business Venturing*, *36*(3), 106092. https://doi.org/10.1016/j.jbusvent.2021.106092
19. Andersen, M. M. (2004). An innovation system approach to eco-innovation-Aligning policy rationales. *The Greening of Policies-Interlinkages and Policy Integration Conference*, 1–28. http://userpage.fu-berlin.de/ffu/akumwelt/bc2004/download/andersen_f.pdf
20. Anglin, A. H., Wolfe, M. T., Short, J. C., McKenny, A. F., & Pidduck, R. J. (2018). Narcissistic rhetoric and crowdfunding performance: A social role theory perspective. *Journal of Business Venturing*, *33*(6), 780–812.
21. Antolín-López, R., Delgado-Ceballos, J., & Montiel, I. (2016). Deconstructing corporate sustainability: A comparison of different stakeholder metrics. *Journal of Cleaner Production*, *136*, 5–17.
22. Antonelli, G. A., Leone, M. I., & Ricci, R. (2022). Exploring the Open COVID Pledge in the fight against COVID-19: A semantic analysis of the Manifesto, the pledgors and the featured patents. *R&D Management*, *52*(2), 255–272. https://doi.org/10.1111/radm.12493
23. Arbelaitz, O., Gurrutxaga, I., Lojo, A., Muguerza, J., Pérez, J. M., & Perona, I. (2013). A navigation-log based web mining application to profile the interests of users accessing the web of Bidasoa Turismo. *Proceedings from ENTER 2013 eTourism Conference. Innsbruck, Austria*.

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=32fbd95639c79563ae67bca4ba601ac9caa73ce6

24. Archibugi, D., & Planta, M. (1996). Measuring technological change through patents and innovation surveys. *Technovation*, *16*(9), 451–519.
25. Archibugi, D., & Sirilli, G. (2000). The direct measurement of technological innovation in business. *International Conference on Innovation and Enterprise: Statistics and Indicators*.
26. Arifi, D., Resch, B., Kinne, J., & Lenz, D. (2023). Innovation in hyperlink and social media networks: Comparing connection strategies of innovative companies in hyperlink and social media networks. *PloS One*, *18*(3), e0283372.
27. Arora, S. K., Li, Y., Youtie, J., & Shapira, P. (2020). Measuring dynamic capabilities in new ventures: Exploring strategic change in US green goods manufacturing using website data. *The Journal of Technology Transfer*, *45*(5), 1451–1480. https://doi.org/10.1007/s10961-019-09751-y
28. Arora, S. K., Youtie, J., Shapira, P., Gao, L., & Ma, T. (2013). Entry strategies in an emerging technology: A pilot web-based study of graphene firms. *Scientometrics*, *95*, 1189–1207.
29. Arrow, K. J. (1962). Economic welfare and the allocation of resources for invention. In R. R. Nelson (Ed.), *The rate and direction of inventive activity: Economic and social factors* (pp. 609–626). Princeton University Press. https://doi.org/10.1515/9781400879762-024
30. Arroyo, J., Corea, F., Jimenez-Diaz, G., & Recio-Garcia, J. A. (2019). Assessment of machine learning performance for decision support in venture capital investments. *Ieee Access*, *7*, 124233–124243.

31. Arthur, W. B. (1994). *Increasing returns and path dependence in the economy*. University of michigan Press. https://books.google.com/books?hl=it&lr=&id=nkc_DwAAQBAJ&oi=fnd&pg=PR7&dq=ARTHUR,+W.+B.+1994.+Increasing+Returns+and+Path+Dependence+in+the+Economy.+University+of+Michigan+Press.&ots=Hn2BcEhmeL&sig=_eny7EP9ebHBrk8kymb4wi6Dh3E
32. Arundel, A., O'Brien, K., & Torugsa, A. (2013). How firm managers understand innovation: Implications for the design of innovation surveys. In *Handbook of innovation indicators and measurement* (pp. 88–108). Edward Elgar Publishing. https://www.elgaronline.com/downloadpdf/edcollchap/edcoll/9780857933645/9780857933645.00012.pdf
33. Arzaghi, M., & Henderson, J. V. (2008). Networking off Madison Avenue. *Review of Economic Studies*, *75*(4), 1011–1038. https://doi.org/10.1111/j.1467-937X.2008.00499.x
34. Asheim, B. T., & Isaksen, A. (1997). Location, agglomeration and innovation: Towards regional innovation systems in Norway? *European Planning Studies*, *5*(3), 299–330. https://doi.org/10.1080/09654319708720402
35. Atuahene-Gima, K. (2004). STRATEGIC DECISION COMPREHENSIVENESS AND NEW PRODUCT DEVELOPMENT OUTCOMES IN NEW TECHNOLOGY VENTURES. *Academy of Management Journal*, *47*(4), 583–597. https://doi.org/10.2307/20159603
36. Auerswald, P. E., & Branscomb, L. M. (2003). Valleys of death and Darwinian seas: Financing the invention-to-innovation transition in the United States. *The Journal of Technology Transfer*, *28*(3/4), 227–239. https://doi.org/10.1023/A:1024980525678

37. Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity? *PloS One*, *16*(4), e0249583.

38. Bafera, J., & Kleinert, S. (2023). Signaling Theory in Entrepreneurship Research: A Systematic Review and Research Agenda. *Entrepreneurship Theory and Practice*, *47*(6), 2419–2464. https://doi.org/10.1177/10422587221138489

39. Balboa, M., & Martí, J. (2007). Factors that determine the reputation of private equity managers in developing markets. *Journal of Business Venturing*, *22*(4), 453–480.

40. Bansal, P., & DesJardine, M. R. (2014). Business sustainability: It is about time. *Strategic Organization*, *12*(1), 70–78. https://doi.org/10.1177/1476127013520265

41. Basuroy, S., Desai, K. K., & Talukdar, D. (2006). An Empirical Investigation of Signaling in the Motion Picture Industry. *Journal of Marketing Research*, *43*(2), 287–295. https://doi.org/10.1509/jmkr.43.2.287

42. Bate, A. F., Wachira, E. W., & Danka, S. (2023). The determinants of innovation performance: An income-based cross-country comparative analysis using the Global Innovation Index (GII). *Journal of Innovation and Entrepreneurship*, *12*(1), 20. https://doi.org/10.1186/s13731-023-00283-2

43. Batista-Navarro, R. T., Kontonatsios, G., Mihăilă, C., Thompson, P., Rak, R., Nawaz, R., Korkontzelos, I., & Ananiadou, S. (2013). Facilitating the Analysis of Discourse Phenomena in an Interoperable NLP Platform. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 7816, pp. 559–571). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37247-6_45

44. Baum, J. A., & Silverman, B. S. (2004). Picking winners or building them? Alliance, intellectual, and human capital as selection criteria in venture financing and performance of biotechnology startups. *Journal of Business Venturing*, *19*(3), 411–436.

45. Becheikh, N., Landry, R., & Amara, N. (2006). Lessons from innovation empirical studies in the manufacturing sector: A systematic review of the literature from 1993–2003. *Technovation*, *26*(5–6), 644–664.

46. Bellavitis, C., Filatotchev, I., Kamuriwo, D. S., & Vanacker, T. (2017). Entrepreneurial finance: New frontiers of research and practice: Editorial for the special issue *Embracing entrepreneurial funding innovations*. *Venture Capital*, *19*(1–2), 1–16. https://doi.org/10.1080/13691066.2016.1259733

47. Bellavitis, C., Kamuriwo, D. S., & Hommel, U. (2019). Mitigation of Moral Hazard and Adverse Selection in Venture Capital Financing: The Influence of the Country's Institutional Setting. *Journal of Small Business Management*, *57*(4), 1328–1349. https://doi.org/10.1111/jsbm.12391

48. Bellucci, A., Pennacchio, L., & Zazzaro, A. (2019). Public R&D subsidies: Collaborative versus individual place-based programs for SMEs. *Small Business Economics*, *52*(1), 213–240. https://doi.org/10.1007/s11187-018-0017-5

49. Bellucci, A., Pennacchio, L., & Zazzaro, A. (2023). Debt financing of SMEs: The certification role of R&D Subsidies. *International Review of Financial Analysis*, *90*, 102903.

50. Ben Arfi, W., Hikkerova, L., & Sahut, J.-M. (2018). External knowledge sources, green innovation and performance. *Technological Forecasting and Social Change*, *129*, 210–220. https://doi.org/10.1016/j.techfore.2017.09.017

51. Bergh, D. D., Connelly, B. L., Ketchen, D. J., & Shannon, L. M. (2014). Signalling Theory and Equilibrium in Strategic Management Research: An Assessment and a Research Agenda. *Journal of Management Studies*, *51*(8), 1334–1360. https://doi.org/10.1111/joms.12097
52. Berrone, P., Fosfuri, A., & Gelabert, L. (2017). Does greenwashing pay off? Understanding the relationship between environmental actions and environmental legitimacy. *Journal of Business Ethics*, *144*, 363–379.
53. Berrone, P., Fosfuri, A., Gelabert, L., & Gomez-Mejia, L. R. (2013). Necessity as the mother of 'green' inventions: Institutional pressures and environmental innovations. *Strategic Management Journal*, *34*(8), 891–909. https://doi.org/10.1002/smj.2041
54. Bertoni, F., & Tykvová, T. (2012). *Which form of venture capital is most supportive of innovation?* ZEW Discussion Papers. https://www.econstor.eu/handle/10419/56030
55. Beske-Janssen, P., Johnson, M. P., & Schaltegger, S. (2015). 20 years of performance measurement in sustainable supply chain management – what has been achieved? *Supply Chain Management: An International Journal*, *20*(6), 664–680. https://doi.org/10.1108/SCM-06-2015-0216
56. Bhattacharyya, J., & Subrahmanya, M. B. (2024). Determinants of a digital start-up's access to VC financing in India: A signaling theory perspective. *Technological Forecasting and Social Change*, *207*, 123631.
57. Björk, J., Frishammar, J., & Sundström, L. (2023). Measuring Innovation Effectively—Nine Critical Lessons. *Research-Technology Management*, *66*(2), 17–27. https://doi.org/10.1080/08956308.2022.2151232

58. Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, *55*(14), 1216–1227. https://doi.org/10.1002/asi.20077

59. Blasi, S., & Sedita, S. R. (2022). Mapping the emergence of a new organisational form: An exploration of the intellectual structure of the B Corp research. *Corporate Social Responsibility and Environmental Management*, *29*(1), 107–123. https://doi.org/10.1002/csr.2187

60. Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy*, *24*(2), 406–428.

61. Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival? *Online Information Review*, *42*(6), 956–970.

62. Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. https://doi.org/10.1145/2133806.2133826

63. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*(Jan), 993–1022.

64. Bloom, N., Van Reenen, J., & Williams, H. (2019). A toolkit of policies to promote innovation. *Journal of Economic Perspectives*, *33*(3), 163–184.

65. Boiral, O., & Henri, J.-F. (2017). Is Sustainability Performance Comparable? A Study of GRI Reports of Mining Organizations. *Business & Society*, *56*(2), 283–317. https://doi.org/10.1177/0007650315576134

66. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

67. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the Opportunities and Risks of Foundation Models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

68. Borah, P. S., Iqbal, S., & Akhtar, S. (2022). Linking social media usage and SME's sustainable performance: The role of digital leadership and innovation capabilities. *Technology in Society*, *68*, 101900.

69. Bossle, M. B., de Barcellos, M. D., Vieira, L. M., & Sauvée, L. (2016). The drivers for adoption of eco-innovation. *Journal of Cleaner Production*, *113*, 861–872.

70. Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2024). Scraping innovativeness from corporate websites: Empirical evidence on Italian manufacturing SMEs. *Technological Forecasting and Social Change*, *207*, 123597.

71. Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). *The snli corpus*. http://archive.nyu.edu/handle/2451/41728

72. Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

73. Brenner, T., & Broekel, T. (2011). Methodological Issues in Measuring Innovation Performance of Spatial Units. *Industry & Innovation*, *18*(1), 7–37. https://doi.org/10.1080/13662716.2010.528927

74. Brito-Ramos, S., Cortez, M. C., & Silva, F. (2024). Do Sustainability Signals Diverge? An Analysis of Labeling Schemes for Socially Responsible Investments. *Business & Society*, *63*(6), 1380–1425. https://doi.org/10.1177/00076503231204613

75. Brown, J. R., Martinsson, G., & Petersen, B. C. (2012). Do financing constraints matter for R&D? *European Economic Review*, *56*(8), 1512–1529.

76. Brown, S. L., & Eisenhardt, K. M. (1995). Product Development: Past Research, Present Findings, and Future Directions. *The Academy of Management Review*, *20*(2), 343. https://doi.org/10.2307/258850

77. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

78. Bubeck, S., Chadrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., & Lundberg, S. (2023). *Sparks of artificial general intelligence: Early experiments with gpt-4*. ArXiv.

79. Bulut, C., Nazli, M., Aydin, E., & Haque, A. U. (2021). The effect of environmental concern on conscious green consumption of post-millennials: The moderating role of greenwashing perceptions. *Young Consumers*, *22*(2), 306–319. https://doi.org/10.1108/YC-10-2020-1241

80. Burns, T., & Stalker, G. M. (1961). Mechanistic and organic systems. *London: Tavistock Publications*. https://books.google.com/books?hl=it&lr=&id=GMR-6gYRmxEC&oi=fnd&pg=PA24&dq=Burns,+T.,+%26+Stalker,+G.+M.+(1961).+Mechanistic+and+organic+systems.+Classics+of+Organizational+Theory,+209%E2%80%93214.&ots=fJqbX64AbE&sig=-hpcgdQOpdKVlVpXU9IMraExxP0

81. Busenitz, L. W., Fiet, J. O., & Moesel, D. D. (2005). Signaling in Venture Capitalist—New Venture Team Funding Decisions: Does it Indicate Long–Term Venture Outcomes? *Entrepreneurship Theory and Practice*, *29*(1), 1–12. https://doi.org/10.1111/j.1540-6520.2005.00066.x

82. Bush, V. (1945). *Science, the endless frontier: A report to the President*. US Government Printing Office. https://books.google.com/books?hl=it&lr=&id=JMTaAAAAMAAJ&oi=fnd&pg=PA8&dq=bush+Science:+The+endless+frontier.&ots=-Pk9JMu-pw&sig=fV-dT9fqlGfIUBoRDfzIb4InVBY

83. Busom ∗, I. (2000). An Empirical Evaluation of The Effects of R&D Subsidies. *Economics of Innovation and New Technology*, *9*(2), 111–148. https://doi.org/10.1080/10438590000000006

84. Buysse, K., & Verbeke, A. (2003). Proactive environmental strategies: A stakeholder management perspective. *Strategic Management Journal*, *24*(5), 453–470. https://doi.org/10.1002/smj.299

85. Cailou, J., & DeHai, L. (2022). Does venture capital stimulate the innovation of China's new energy enterprises? *Energy*, *244*, 122704.

86. Calabrese, A., Costa, R., Ghiron, N. L., Tiburzi, L., & Pedersen, E. R. G. (2021). How sustainable-orientated service innovation strategies are contributing to the sustainable development goals. *Technological Forecasting and Social Change*, *169*, 120816.

87. Callison, C. (2003). Media relations and the Internet: How Fortune 500 company web sites assist journalists in news gathering. *Public Relations Review*, *29*(1), 29–41.

88. Calvino, F., Samek, L., Squicciarini, M., & Morris, C. (2022). *Identifying and characterising AI adopters: A novel approach based on big data*. https://www.sipotra.it/wp-content/uploads/2023/01/Identifying-and-characterising-AI-adopters-A-novel-approach-based-on-big-data.pdf

89. Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu

(Eds.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, pp. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

90. Cantele, S., & Zardini, A. (2020). What drives small and medium enterprises towards sustainability? Role of interactions between pressures, barriers, and benefits. *Corporate Social Responsibility and Environmental Management*, *27*(1), 126–136. https://doi.org/10.1002/csr.1778
91. Cao, K., Gehman, J., & Grimes, M. G. (2017). Standing out and fitting in: Charting the emergence of Certified B Corporations by industry and region. In *Hybrid ventures* (pp. 1–38). Emerald Publishing Limited. https://www.emerald.com/insight/content/doi/10.1108/S1074-754020170000019001/full/html
92. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., & Erlingsson, U. (2021). Extracting training data from large language models. *30th USENIX Security Symposium (USENIX Security 21)*, 2633–2650. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting
93. Carlino, G., & Kerr, W. R. (2015). Agglomeration and Innovation. In *Handbook of Regional and Urban Economics* (Vol. 5, pp. 349–404). Elsevier. https://doi.org/10.1016/B978-0-444-59517-1.00006-4
94. Carpenter, C. R. (1969). 3 APPROACHES TO STUDIES OF THE NATURALISTIC COMMUNICATIVE BEHAVIOR IN NONHUMAN PRIMATES. In T. A. Sebeok & A. Ramsay (Eds.), *Approaches to Animal Communication*. DE GRUYTER MOUTON. https://doi.org/10.1515/9783110862850.40

95. Carree, M. A., & Thurik, A. R. (2010). The Impact of Entrepreneurship on Economic Growth. In Z. J. Acs & D. B. Audretsch (Eds.), *Handbook of Entrepreneurship Research* (pp. 557–594). Springer New York. https://doi.org/10.1007/978-1-4419-1191-9_20

96. Caselli, S., Gatti, S., & Perrini, F. (2009a). Are Venture Capitalists a Catalyst for Innovation? *European Financial Management*, *15*(1), 92–111. https://doi.org/10.1111/j.1468-036X.2008.00445.x

97. Caselli, S., Gatti, S., & Perrini, F. (2009b). Are Venture Capitalists a Catalyst for Innovation? *European Financial Management*, *15*(1), 92–111. https://doi.org/10.1111/j.1468-036X.2008.00445.x

98. Caselli, S., & Negri, G. (2021). Private equity and venture capital in Europe: Markets, techniques, and deals. Academic Press. https://books.google.com/books?hl=it&lr=&id=IOkGEAAAQBAJ&oi=fnd&pg=PP1&dq=Private+Equity+and+Venture+Capital+in+Europe:+Markets,+Techniques,+and+Deals&ots=v78AM_nBA6&sig=i8iXah53UUqNsLllxbOvqzaTGbk

99. Catalini, C. (2012). How Does Co-Location Affect the Rate and Direction of Innovative Activity? *Academy of Management Proceedings*, *2012*(1), 12888. https://doi.org/10.5465/AMBPP.2012.46

100. Caviggioli, F., Colombelli, A., Marco, A. D., & Paolucci, E. (2020). How venture capitalists evaluate young innovative company patent portfolios: Empirical evidence from Europe. *International Journal of Entrepreneurial Behavior &amp; Research*, *26*(4), 695–721. https://doi.org/10.1108/IJEBR-10-2018-0692

101. Centobelli, P., Cerchione, R., & Esposito, E. (2020). Pursuing supply chain sustainable development goals through the adoption of green practices and enabling technologies: A

cross-country analysis of LSPs. *Technological Forecasting and Social Change*, *153*, 119920.

102. Certo, S. T., Covin, J. G., Daily, C. M., & Dalton, D. R. (2001). Wealth and the effects of founder management among IPO-stage new ventures. *Strategic Management Journal*, *22*(6–7), 641–658. https://doi.org/10.1002/smj.182

103. Chandler, A. D. (1990). The enduring logic of industrial success. *Harvard Business Review*, *68*(2), 130–140.

104. Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Y., Yu, P. S., Yang, Q., & Xie, X. (2024). A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–45. https://doi.org/10.1145/3641289

105. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

106. Chen, H., & Chau, M. (2004). Web mining: Machine learning for web applications. *Annual Review of Information Science and Technology*, *38*(1), 289–329. https://doi.org/10.1002/aris.1440380107

107. Chen, J., & Ewens, M. (2021). *Venture capital and startup agglomeration*. National Bureau of Economic Research. https://www.nber.org/papers/w29211

108. Chen, J., Wang, X., Xu, R., Yuan, S., Zhang, Y., Shi, W., Xie, J., Li, S., Yang, R., Zhu, T., Chen, A., Li, N., Chen, L., Hu, C., Wu, S., Ren, S., Fu, Z., & Xiao, Y. (2024). *From Persona to Personalization: A Survey on Role-Playing Language Agents* (arXiv:2404.18231). arXiv. https://doi.org/10.48550/arXiv.2404.18231

109. Chen, S., Meng, W., & Lu, H. (2018). Patent as a Quality Signal in Entrepreneurial Finance: A Look Beneath the Surface. *Asia-Pacific Journal of Financial Studies*, *47*(2), 280–305. https://doi.org/10.1111/ajfs.12211

110. Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

111. Chen, Y.-S., & Chang, C.-H. (2013). Greenwash and Green Trust: The Mediation Effects of Green Consumer Confusion and Green Perceived Risk. *Journal of Business Ethics*, *114*(3), 489–500. https://doi.org/10.1007/s10551-012-1360-0

112. Chen, Y.-S., Huang, A.-F., Wang, T.-Y., & Chen, Y.-R. (2020). Greenwash and green purchase behaviour: The mediation of green brand image and green brand loyalty. *Total Quality Management & Business Excellence*, *31*(1–2), 194–209. https://doi.org/10.1080/14783363.2018.1426450

113. Chen, Y.-S., Lai, S.-B., & Wen, C.-T. (2006). The influence of green innovation performance on corporate advantage in Taiwan. *Journal of Business Ethics*, *67*, 331–339.

114. Chesbrough, H. (2006). *Open business models: How to thrive in the new innovation landscape*. Harvard Business Press. https://books.google.com/books?hl=it&lr=&id=MWPlLbULAmwC&oi=fnd&pg=PR9&dq=Open+business+models:+How+to+thrive+in+the+new+innovation+landscape&ots=BIuGY_ms7U&sig=v8jTXMyqRtaVdUgRyZqSusGRLBw

115. Chesbrough, H., & Bogers, M. (2014). Explicating open innovation: Clarifying an emerging paradigm for understanding innovation. *New Frontiers in Open Innovation. Oxford: Oxford University Press, Forthcoming*, 3–28.

116. Chesbrough, H. W. (2003). *Open innovation: The new imperative for creating and profiting from technology*. Harvard Business Press. https://books.google.com/books?hl=it&lr=&id=4hTRWStFhVgC&oi=fnd&pg=PR9&dq=chesbrough+2003+open+innovation&ots=XvVxTLy2AB&sig=ixyeZ64uUa_xrRewQCgyZmgEpQk

117. Ching, H. Y., & Gerab, F. (2017). Sustainability reports in Brazil through the lens of signaling, legitimacy and stakeholder theories. *Social Responsibility Journal*, *13*(1), 95–110.

118. Christensen, C. M. (1997). *The Innovator's Dilemma: When New Technologies Cause Great Firms to Fail*. Harvard Business School Press.

119. Christensen, H. B., Hail, L., & Leuz, C. (2021). Mandatory CSR and sustainability reporting: Economic analysis and literature review. *Review of Accounting Studies*, *26*(3), 1176–1248. https://doi.org/10.1007/s11142-021-09609-5

120. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/7017-deep-reinforcement-learning-from-human-preferences

121. Christmann, P. (2000). EFFECTS OF "BEST PRACTICES" OF ENVIRONMENTAL MANAGEMENT ON COST ADVANTAGE: THE ROLE OF COMPLEMENTARY ASSETS. *Academy of Management Journal*, *43*(4), 663–680. https://doi.org/10.2307/1556360

122. Chung, W., & Kalnins, A. (2001). Agglomeration effects and performance: A test of the Texas lodging industry. *Strategic Management Journal*, *22*(10), 969–988. https://doi.org/10.1002/smj.178

123. Cillo, V., Petruzzelli, A. M., Ardito, L., & Del Giudice, M. (2019). Understanding sustainable innovation: A systematic literature review. *Corporate Social Responsibility and Environmental Management*, *26*(5), 1012–1025.

124. Cirera, X., & Muzi, S. (2020). Measuring innovation using firm-level surveys: Evidence from developing countries☆. *Research Policy*, *49*(3), 103912. https://doi.org/10.1016/j.respol.2019.103912

125. Clark, T., & Charter, M. (2007). *Sustainable innovation: Key conclusions from sustainable innovation conferences 2003–2006 organised by the centre for sustainable design*. http://research.uca.ac.uk/id/eprint/694

126. Clarkson, M. B. E. (1995). A Stakeholder Framework for Analyzing and Evaluating Corporate Social Performance. *The Academy of Management Review*, *20*(1), 92. https://doi.org/10.2307/258888

127. Clough, D. R., Fang, T. P., Vissa, B., & Wu, A. (2019). Turning Lead into Gold: How Do Entrepreneurs Mobilize Resources to Exploit Opportunities? *Academy of Management Annals*, *13*(1), 240–271. https://doi.org/10.5465/annals.2016.0132

128. Cohen, B., Smith, B., & Mitchell, R. (2008). Toward a sustainable conceptualization of dependent variables in entrepreneurship research. *Business Strategy and the Environment*, *17*(2), 107–119. https://doi.org/10.1002/bse.505

129. Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). *Natural language processing (almost) from scratch*. https://www.jmlr.org/papers/volume12/collobert11a/collobert11a.pdf?source

130. Colombo, M. G., D’Adda, D., & Pirelli, L. H. (2016). The participation of new technology-based firms in EU-funded R&D partnerships: The role of venture capital. *Research Policy*, *45*(2), 361–375.

131. Colombo, M. G., Guerini, M., Hoisl, K., & Zeiner, N. M. (2023). The dark side of signals: Patents protecting radical inventions and venture capital investments. *Research Policy*, *52*(5), 104741.

132. Colombo, O. (2021). The use of signals in new-venture financing: A review and research agenda. *Journal of Management*, *47*(1), 237–259.

133. Connelly, B. L., Certo, S. T., Ireland, R. D., & Reutzel, C. R. (2011). Signaling theory: A review and assessment. *Journal of Management*, *37*(1), 39–67.

134. Cooley, R., Mobasher, B., & Srivastava, J. (1997). Web mining: Information and pattern discovery on the World Wide Web. *Proceedings Ninth IEEE International Conference on Tools with Artificial Intelligence*, 558–567. https://doi.org/10.1109/TAI.1997.632303

135. Coombs, R., Narandren, P., & Richards, A. (1996). A literature-based innovation output indicator. *Research Policy*, *25*(3), 403–413.

136. Corea, F., Bertinetti, G., & Cervellati, E. M. (2021). Hacking the venture industry: An Early-stage Startups Investment framework for data-driven investors. *Machine Learning with Applications*, *5*, 100062.

137. Cormier, D., & Magnan, M. (2015). The Economic Relevance of Environmental Disclosure and its Impact on Corporate Legitimacy: An Empirical Investigation. *Business Strategy and the Environment*, *24*(6), 431–450. https://doi.org/10.1002/bse.1829

138. Cowan, K., & Guzman, F. (2020). How CSR reputation, sustainability signals, and country-of-origin sustainability reputation contribute to corporate brand performance: An exploratory study. *Journal of Business Research*, *117*, 683–693.

139. Cropper, S. (2008). *The Oxford handbook of inter-organizational relations*. Oxford Handbooks Online. https://books.google.com/books?hl=it&lr=&id=BfKVxqq3JEwC&oi=fnd&pg=PR7&dq=Cropper,+S.,+Ebers,+M.,+%26+Huxham,+C.+(2008).+The+Oxford+handbook+of+inter-organizational+relations.+Oxford+Handbooks.&ots=Dw4OCaCz03&sig=Z-qUlJtNMIevSJweRtX3-P1R8ZE

140. Crosato, L., Domenech, J., & Liberati, C. (2024). Websites' data: A new asset for enhancing credit risk modeling. *Annals of Operations Research*, *342*(3), 1671–1686. https://doi.org/10.1007/s10479-023-05306-5

141. Cruciata, P., Pulizzotto, D., & Beaudry, C. (2024). First impressions on sustainable innovation matter: Using NLP to replicate B-lab environmental index by analyzing companies' homepages. *Technological Forecasting and Social Change*, *205*, 123455.

142. Czarnitzki, D. (2006). RESEARCH AND DEVELOPMENT IN SMALL AND MEDIUM-SIZED ENTERPRISES: THE ROLE OF FINANCIAL CONSTRAINTS AND PUBLIC FUNDING. *Scottish Journal of Political Economy*, *53*(3), 335–357. https://doi.org/10.1111/j.1467-9485.2006.00383.x

143. Czarnitzki, D., & Lopes-Bento, C. (2013). Value for money? New microeconometric evidence on public R&D grants in Flanders. *Research Policy*, *42*(1), 76–89.

144. Dahlander, L., & Gann, D. M. (2010). How open is innovation? *Research Policy*, *39*(6), 699–709.

145. Dahlke, J., Beck, M., Kinne, J., Lenz, D., Dehghan, R., Wörter, M., & Ebersberger, B. (2024). Epidemic effects in the diffusion of emerging digital technologies: Evidence from artificial intelligence adoption. *Research Policy*, *53*(2), 104917.

146. Daily, C. M., Certo, S. T., & Dalton, D. R. (2005). Investment bankers and IPO pricing: Does prospectus information matter? *Journal of Business Venturing*, *20*(1), 93–111.

147. Darnall, N., & Sides, S. (2008). Assessing the Performance of Voluntary Environmental Programs: Does Certification Matter? *Policy Studies Journal*, *36*(1), 95–117. https://doi.org/10.1111/j.1541-0072.2007.00255.x

148. David, P. A., Hall, B. H., & Toole, A. A. (2000). Is public R&D a complement or substitute for private R&D? A review of the econometric evidence. *Research Policy*, *29*(4–5), 497–529.

149. Davila, A., Foster, G., & Gupta, M. (2003). Venture capital financing and the growth of startup firms. *Journal of Business Venturing*, *18*(6), 689–708. https://doi.org/10.1016/S0883-9026(02)00127-1

150. de Azevedo Rezende, L., Bansi, A. C., Alves, M. F. R., & Galina, S. V. R. (2019). Take your time: Examining when green innovation affects financial performance in multinationals. *Journal of Cleaner Production*, *233*, 993–1003.

151. De Rassenfosse, G., & Fischer, T. (2016). Venture Debt Financing: Determinants of the Lending Decision. *Strategic Entrepreneurship Journal*, *10*(3), 235–256. https://doi.org/10.1002/sej.1220

152. Dechezleprêtre, A., Martin, R., & Bassi, S. (2019). Climate change policy, innovation and growth. In *Handbook on green growth* (pp. 217–239). Edward Elgar Publishing. https://www.elgaronline.com/edcollchap/edcoll/9781788110679/9781788110679.00018.xml

153. DeepSeek-AI, Guo, D., Yang, D., Zhang, H., Song, J., Zhang, R., Xu, R., Zhu, Q., Ma, S., Wang, P., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., … Zhang, Z. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (arXiv:2501.12948). arXiv. https://doi.org/10.48550/arXiv.2501.12948

154. Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, *41*(6), 391–407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9

155. Del Río, P., Carrillo-Hermosilla, J., & Könnölä, T. (2010). Policy Strategies to Promote Eco-Innovation: An Integrated Framework. *Journal of Industrial Ecology*, *14*(4), 541–557. https://doi.org/10.1111/j.1530-9290.2010.00259.x

156. Delmas, M., & Keller, A. (2005). Free riding in voluntary environmental programs: The case of the U.S. EPA WasteWise program. *Policy Sciences*, *38*(2–3), 91–106. https://doi.org/10.1007/s11077-005-6592-8

157. Dernis, H., Calvino, F., Moussiegt, L., Nawa, D., Samek, L., & Squicciarini, M. (2023). *Identifying artificial intelligence actors using online data* (OECD Science, Technology and Industry Working Papers 2023/01). OECD Publishing. https://doi.org/10.1787/1f5307e7-en

158. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint arXiv:1810.04805*.

159. Dewangan, V., & Godse, M. (2014). Towards a holistic enterprise innovation performance measurement system. *Technovation*, *34*(9), 536–545.

160. Dias, A., Rodrigues, L. L., Craig, R., & Neves, M. E. (2019). Corporate social responsibility disclosure in small and medium-sized entities and large companies. *Social Responsibility Journal*, *15*(2), 137–154.

161. Díaz-García, C., González-Moreno, Á., & Sáez-Martínez, F. J. (2015). Eco-innovation: Insights from a literature review. *Innovation*, *17*(1), 6–23. https://doi.org/10.1080/14479338.2015.1011060

162. DiVito, L., & Bohnsack, R. (2017). Entrepreneurial orientation and its effect on sustainability decision tradeoffs: The case of sustainable fashion firms. *Journal of Business Venturing*, *32*(5), 569–587. https://doi.org/10.1016/j.jbusvent.2017.05.002

163. Doblinger, C., Surana, K., & Anadon, L. D. (2019). Governments as partners: The role of alliances in U.S. cleantech startup innovation. *Research Policy*, *48*(6), 1458–1475. https://doi.org/10.1016/j.respol.2019.02.006

164. Doran, J., & Ryan, G. (2012). Regulation and firm perception, eco-innovation and firm performance. *European Journal of Innovation Management*.

165. Dörr, J. O., Kinne, J., Lenz, D., Licht, G., & Winker, P. (2022). An integrated data framework for policy guidance during the coronavirus pandemic: Towards real-time decision support for economic policymakers. *Plos One*, *17*(2), e0263898.

166. Du, L., Zhang, Z., & Feng, T. (2018). Linking green customer and supplier integration with green innovation performance: The role of internal integration. *Business Strategy and the Environment*, *27*(8), 1583–1595. https://doi.org/10.1002/bse.2223

167. Dushnitsky, G., & Lenox, M. J. (2006). When does corporate venture capital investment create firm value? *Journal of Business Venturing*, *21*(6), 753–772.

168. Dybvig, P. H., & Spatt, C. S. (1983). Adoption externalities as public goods. *Journal of Public Economics*, *20*(2), 231–247.

169. Dziallas, M., & Blind, K. (2019). Innovation indicators throughout the innovation process: An extensive literature analysis. *Technovation*, *80*, 3–29.

170. Eccles, R. G., Ioannou, I., & Serafeim, G. (2014). The impact of corporate sustainability on organizational processes and performance. *Management Science*, *60*(11), 2835–2857.

171. Edeh, J. N., Obodoechi, D. N., & Ramos-Hidalgo, E. (2020). Effects of innovation strategies on export performance: New empirical evidence from developing market firms. *Technological Forecasting and Social Change*, *158*, 120167.

172. El Bassiti, L., & Ajhoun, R. (2016). Towards Innovation Excellence: Why and How to Measure Innovation Performance? *2016 6th International Conference on Information and Communication Technology for The Muslim World (ICT4M)*, 99–104. https://ieeexplore.ieee.org/abstract/document/7814884/?casa_token=x8f8KCRs9fAAAAAA:dZCqQAbUaUD5um3UzJ4vFvOowoIkIHue-L9djWRR499tPLsh8JAopZOrRjgsYda5aAN9Pz-u

173. Elkington, J. (1997). The triple bottom line. *Environmental Management: Readings and Cases*, *2*, 49–66.

174. Ellwood, P., Williams, C., & Egan, J. (2022). Crossing the valley of death: Five underlying innovation processes. *Technovation*, *109*, 102162.

175. Erdin, C., & Çağlar, M. (2023). National innovation efficiency: A DEA-based measurement of OECD countries. *International Journal of Innovation Science*, *15*(3), 427–456.

176. Etzioni, O. (1996). The World-Wide Web: Quagmire or gold mine? *Communications of the ACM*, *39*(11), 65–68. https://doi.org/10.1145/240455.240473

177. Etzkowitz, H., & Leydesdorff, L. (2000). The dynamics of innovation: From National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, *29*(2), 109–123.

178. European Commission, D.-G. for R. and I. (2024). *European Innovation Scoreboard 2024: Main report*. https://research-and-innovation.ec.europa.eu/statistics/performance-indicators/european-innovation-scoreboard_en

179. Evans, S., Vladimirova, D., Holgado, M., Van Fossen, K., Yang, M., Silva, E. A., & Barlow, C. Y. (2017). Business Model Innovation for Sustainability: Towards a Unified Perspective for Creation of Sustainable Business Models. *Business Strategy and the Environment*, *26*(5), 597–608. https://doi.org/10.1002/bse.1939

180. Evanschitzky, H., Eisend, M., Calantone, R. J., & Jiang, Y. (2012). Success Factors of Product Innovation: An Updated Meta-Analysis. *Journal of Product Innovation Management*, *29*(S1), 21–37. https://doi.org/10.1111/j.1540-5885.2012.00964.x

181. Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from weblogs: A survey. *Data & Knowledge Engineering*, *53*(3), 225–241. https://doi.org/10.1016/j.datak.2004.08.001

182. Fernández-Vázquez, J.-S., & Sancho-Rodríguez, Á. (2020). Critical discourse analysis of climate change in IBEX 35 companies. *Technological Forecasting and Social Change*, *157*, 120063.

183. Fernhaber, S. A., Mcdougall-Covin, P. P., & Shepherd, D. A. (2009). International entrepreneurship: Leveraging internal and external knowledge sources. *Strategic Entrepreneurship Journal*, *3*(4), 297–320. https://doi.org/10.1002/sej.76

184. Fischer, D., Brettel, M., & Mauer, R. (2020). The Three Dimensions of Sustainability: A Delicate Balancing Act for Entrepreneurs Made More Complex by Stakeholder

Expectations. *Journal of Business Ethics*, *163*(1), 87–106. https://doi.org/10.1007/s10551-018-4012-1

185. Fischer, E., & Reuber, R. (2007). The Good, the Bad, and the Unfamiliar: The Challenges of Reputation Formation Facing New Firms. *Entrepreneurship Theory and Practice*, *31*(1), 53–75. https://doi.org/10.1111/j.1540-6520.2007.00163.x

186. Flammer, C. (2021). Corporate green bonds. *Journal of Financial Economics*, *142*(2), 499–516.

187. Forsman, H. (2013). Environmental Innovations as a Source of Competitive Advantage or Vice Versa? *Business Strategy and the Environment*, *22*(5), 306–320. https://doi.org/10.1002/bse.1742

188. Frank, C., Sink, C., Mynatt, L., Rogers, R., & Rappazzo, A. (1996). Surviving the "valley of death": A comparative analysis. *The Journal of Technology Transfer*, *21*(1–2), 61–69. https://doi.org/10.1007/BF02220308

189. Freeman, C., & Perez, C. (1988). Structural crises of adjustment: Business cycles. *Technical Change and Economic Theory. Londres: Pinter*. https://carlotaperez.org/wp-content/downloads/publications/theoretical-framework/StructuralCrisesOfAdjustment.pdf

190. Freeman, R. E. (1984). *Strategic management: A stakeholder approach* (2. [print.]). Pitman.

191. Füller, J., Matzler, K., & Hoppe, M. (2008). Brand Community Members as a Source of Innovation. *Journal of Product Innovation Management*, *25*(6), 608–619. https://doi.org/10.1111/j.1540-5885.2008.00325.x

192. Gaddy, B. E., Sivaram, V., Jones, T. B., & Wayman, L. (2017). Venture Capital and Cleantech: The wrong model for energy innovation. *Energy Policy*, *102*, 385–395. https://doi.org/10.1016/j.enpol.2016.12.035

193. Gallardo-Vázquez, D., Valdez-Juárez, L. E., & Castuera-Díaz, Á. M. (2019). Corporate social responsibility as an antecedent of innovation, reputation, performance, and competitive success: A multiple mediation analysis. *Sustainability*, *11*(20), 5614.

194. Gama, N., Silva, M. M. D., & Ataíde, J. (2007). Innovation Scorecard: A Balanced Scorecard for Measuring the Value Added by Innovation. In P. F. Cunha & P. G. Maropoulos (Eds.), *Digital Enterprise Technology* (pp. 417–424). Springer US. https://doi.org/10.1007/978-0-387-49864-5_49

195. Gao, H., Darroch, J., Mather, D., & MacGregor, A. (2008). Signaling Corporate Strategy in IPO Communication: A Study of Biotechnology IPOs on the NASDAQ. *Journal of Business Communication*, *45*(1), 3–30. https://doi.org/10.1177/0021943607309349

196. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, H., & Wang, H. (2023). Retrieval-augmented generation for large language models: A survey. *arXiv Preprint arXiv:2312.10997*, *2*. https://simg.baai.ac.cn/paperfile/25a43194-c74c-4cd3-b60f-0a1f27f8b8af.pdf

197. Gao, Y., Xiong, Y., Wang, M., & Wang, H. (2024). *Modular RAG: Transforming RAG Systems into LEGO-like Reconfigurable Frameworks* (arXiv:2407.21059). arXiv. https://doi.org/10.48550/arXiv.2407.21059

198. Garriga, E., & Melé, D. (2004). Corporate social responsibility theories: Mapping the territory. *Journal of Business Ethics*, *53*(1), 51–71.

199. Gault, F. (2018). Defining and measuring innovation in all sectors of the economy. *Research Policy*, *47*(3), 617–622. https://doi.org/10.1016/j.respol.2018.01.007

200. Gault, F., Arundel, A., & Kraemer-Mbula, E. (2023). *Handbook of innovation indicators and measurement*. Edward Elgar Publishing. https://books.google.com/books?hl=it&lr=&id=cQXXEAAAQBAJ&oi=fnd&pg=PR1&dq=Handbook+of+innovation+indicators+and+measurement)&ots=P-ie8PXvUM&sig=V49ksaZRmXQtbA0SipCyGJZ_xrw

201. Geroski, P. A., Van Reenen, J., & Walters, C. F. (1997). How persistently do firms innovate? *Research Policy*, *26*(1), 33–48.

202. Ghasemaghaei, M., & Calic, G. (2020). Assessing the impact of big data on firm innovation performance: Big data is not always better data. *Journal of Business Research*, *108*, 147–162.

203. Ghosh, S., & Nanda, R. (2010). Venture capital investment in the clean energy sector. *Harvard Business School Entrepreneurial Management Working Paper*, *11–020*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1669445

204. Goczol, J., & Scoubeau, C. (2003). Corporate communication and strategy in the field of projects. *Corporate Communications: An International Journal*, *8*(1), 60–66.

205. Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

206. Gompers, P. A., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2020). How do venture capitalists make decisions? *Journal of Financial Economics*, *135*(1), 169–190.

207. Gompers, P., Gornall, W., Kaplan, S. N., & Strebulaev, I. A. (2021). Venture capitalists and COVID-19. *Journal of Financial and Quantitative Analysis*, *56*(7), 2474–2499.

208. Gompers, P., & Lerner, J. (2001a). The venture capital revolution. *Journal of Economic Perspectives*, *15*(2), 145–168.

209. Gompers, P., & Lerner, J. (2001b). The venture capital revolution. *Journal of Economic Perspectives*, *15*(2), 145–168.

210. González, X., & Pazó, C. (2008). Do public subsidies stimulate private R&D spending? *Research Policy*, *37*(3), 371–389.

211. Grafen, A. (1990). Biological signals as handicaps. *Journal of Theoretical Biology*, *144*(4), 517–546.

212. Green, K., McMeekin, A., & Irwin, A. (1994). Technological trajectories and R&D for environmental innovation in UK firms. *Futures*, *26*(10), 1047–1059.

213. Greenacre, P., Gross, R., & Speirs, J. (2012). Innovation Theory: A review of the literature. *Imperial College Centre for Energy Policy and Technology, London*, 141–164.

214. Griliches, Z. (1991). The search for R&D spillovers. *National Bureau of Economic Research Working Paper Series*, *w3768*. https://www.nber.org/papers/w3768

215. Grilli, L., & Murtinu, S. (2014). Government, venture capital and the growth of European high-tech entrepreneurial firms. *Research Policy*, *43*(9), 1523–1543.

216. Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? *Advances in Neural Information Processing Systems*, *35*, 507–520.

217. Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

218. Grübler, A., Nakićenović, N., & Victor, D. G. (1999). Dynamics of energy technologies and global change. *Energy Policy*, *27*(5), 247–280.

219. Gubitta, P., Tognazzo, A., & Destro, F. (2016). Signaling in academic ventures: The role of technology transfer offices and university funds. *The Journal of Technology Transfer*, *41*(2), 368–393. https://doi.org/10.1007/s10961-015-9398-7

220. Guerreiro, J., & Pacheco, M. (2021). How Green Trust, Consumer Brand Engagement and Green Word-of-Mouth Mediate Purchasing Intentions. *Sustainability*, *13*(14), 7877. https://doi.org/10.3390/su13147877

221. Guilford, T., & Dawkins, M. S. (1995). What are conventional signals? *Animal Behaviour*, *49*(6), 1689–1695.

222. Gulati, R., & Higgins, M. C. (2003). Which ties matter when? The contingent effects of interorganizational partnerships on IPO success. *Strategic Management Journal*, *24*(2), 127–144. https://doi.org/10.1002/smj.287

223. Guo, D., Guo, Y., & Jiang, K. (2022). Government R&D support and firms' access to external financing: Funding effects, certification effects, or both? *Technovation*, *115*, 102469.

224. Guo, D., & Jiang, K. (2013). Venture capital investment and the performance of entrepreneurial firms: Evidence from China. *Journal of Corporate Finance*, *22*, 375–395.

225. Gupta, J., & Das, N. (2022). Multidimensional corporate social responsibility disclosure and financial performance: A meta-analytical review. *Corporate Social Responsibility and Environmental Management*, *29*(4), 731–748. https://doi.org/10.1002/csr.2237

226. Gupta, S., Ranjan, R., & Singh, S. N. (2024). *A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions* (arXiv:2410.12837). arXiv. https://doi.org/10.48550/arXiv.2410.12837

227. Guyader, H., Ottosson, M., & Witell, L. (2017). You can't buy what you can't see: Retailer practices to increase the green premium. *Journal of Retailing and Consumer Services*, *34*, 319–325.

228. Hair, J., Tatham, R., Anderson, R., & Black, W. (1998). *Multivariate data analysis Prentice-Hall London* (5th ed.). UK.

229. Hajikhani, A., Pukelis, L., Suominen, A., Ashouri, S., Schubert, T., Notten, A., & Cunningham, S. W. (2022). Connecting firm's web scraped textual content to body of science: Utilizing microsoft academic graph hierarchical topic modeling. *MethodsX*, *9*, 101650.

230. Hall, B. H., & Lerner, J. (2010). The financing of R&D and innovation. In *Handbook of the Economics of Innovation* (Vol. 1, pp. 609–639). Elsevier. https://www.sciencedirect.com/science/article/pii/S0169721810010142

231. Hall, B. H., & Rosenberg, N. (2010). *Handbook of the Economics of Innovation* (Vol. 1). Elsevier. https://books.google.com/books?hl=it&lr=&id=4nZTCD_zjN4C&oi=fnd&pg=PP1&dq=he+financing+of+R%26D+and+innovation.+In+B.+H.+Hall+%26+N.+Rosenberg+(Eds.),+Handbook+of+the+Economics+of+Innovation&ots=5YH4yHI6eu&sig=BYAdyWwz18wvTEvOAV6m-XFQFMk

232. Hameed, I., Hyder, Z., Imran, M., & Shafiq, K. (2021). Greenwash and green purchase behavior: An environmentally sustainable perspective. *Environment, Development and Sustainability*, *23*(9), 13113–13134. https://doi.org/10.1007/s10668-020-01202-1

233. Hassan, S.-U., & Haddawy, P. (2015). Analyzing knowledge flows of scientific literature through semantic links: A case study in the field of energy. *Scientometrics*, *103*(1), 33–46. https://doi.org/10.1007/s11192-015-1528-3

234. Hayami, Y., & Ruttan, V. W. (1971). *Agricultural development: An international perspective.* https://www.cabidigitallibrary.org/doi/full/10.5555/19721890134

235. He, H., Bai, Y., Garcia, E. A., & Li, S. (2008). ADASYN: Adaptive synthetic sampling approach for imbalanced learning. *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, 1322–1328. https://ieeexplore.ieee.org/abstract/document/4633969/?casa_token=E8mI5M6J71cAAAAA:iabjW4zQin2LjqWgcuAZFNrRgIrPmuh0fp38AZn9UfBR6np0HXbyvMBfwKx0FS5OyF8BMWpk

236. Heijs, J., Guerrero, A. J., & Huergo, E. (2022). Understanding the Heterogeneous Additionality of R&D Subsidy Programs of Different Government Levels. *Industry and Innovation*, *29*(4), 533–563. https://doi.org/10.1080/13662716.2021.1990024

237. Hekkert, M. P., & Negro, S. O. (2009). Functions of innovation systems as a framework to understand sustainable technological change: Empirical evidence for earlier claims. *Technological Forecasting and Social Change*, *76*(4), 584–594.

238. Hellmann, T., & Puri, M. (2002). Venture Capital and the Professionalization of Start-Up Firms: Empirical Evidence. *The Journal of Finance*, *57*(1), 169–197. https://doi.org/10.1111/1540-6261.00419

239. Hermundsdottir, F., & Aspelund, A. (2021a). Sustainability innovations and firm competitiveness: A review. *Journal of Cleaner Production*, *280*, 124715.

240. Hermundsdottir, F., & Aspelund, A. (2021b). Sustainability innovations and firm competitiveness: A review. *Journal of Cleaner Production*, *280*, 124715.

241. Héroux-Vaillancourt, M. (2023). *Exploring Corporate Websites to Assess Enterprise Innovation: Measuring and Validating the Signal from Textual Content* [Ph.\,D. thesis, Polytechnique Montréal]. https://publications.polymtl.ca/53430/

242. Héroux-Vaillancourt, M., Beaudry, C., & Rietsch, C. (2020). Using web content analysis to create innovation indicators—What do we really measure? *Quantitative Science Studies*, *1*(4), 1601–1637.

243. Hirukawa, M., & Ueda, M. (2011). VENTURE CAPITAL AND INNOVATION: WHICH IS FIRST? *Pacific Economic Review*, *16*(4), 421–465. https://doi.org/10.1111/j.1468-0106.2011.00557.x

244. Hoehn-Weiss, M. N., & Karim, S. (2014). Unpacking functional alliance portfolios: How signals of viability affect young firms' outcomes. *Strategic Management Journal*, *35*(9), 1364–1385.

245. Hoelscher, K. (2016). The evolution of the smart cities agenda in India. *International Area Studies Review*, *19*(1), 28–44. https://doi.org/10.1177/2233865916632089

246. Hoenig, D., & Henkel, J. (2015). Quality signals? The role of patents, alliances, and team experience in venture capital financing. *Research Policy*, *44*(5), 1049–1064.

247. Hojnik, J., & Ruzzier, M. (2016). The driving forces of process eco-innovation and its impact on performance: Insights from Slovenia. *Journal of Cleaner Production*, *133*, 812–825. https://doi.org/10.1016/j.jclepro.2016.06.002

248. Horbach, J. (2008). Determinants of environmental innovation—New evidence from German panel data sources. *Research Policy*, *37*(1), 163–173.

249. Horbach, J., Rammer, C., & Rennings, K. (2012). Determinants of eco-innovations by type of environmental impact—The role of regulatory push/pull, technology push and market pull. *Ecological Economics*, *78*, 112–122.

250. Hörisch, J., Freeman, R. E., & Schaltegger, S. (2014). Applying Stakeholder Theory in Sustainability Management: Links, Similarities, Dissimilarities, and a Conceptual

Framework. *Organization & Environment*, *27*(4), 328–346. https://doi.org/10.1177/1086026614535786

251. Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, *24*(6), 417.

252. Howell, S. T. (2017). Financing innovation: Evidence from R&D grants. *American Economic Review*, *107*(4), 1136–1164.

253. Howell, S. T., Lerner, J., Nanda, R., & Townsend, R. R. (2020). *How resilient is venture-backed innovation? Evidence from four decades of US patenting*. National Bureau of Economic Research. https://www.nber.org/papers/w27150

254. Hsu, D. H. (2007). Experienced entrepreneurial founders, organizational capital, and venture capital funding. *Research Policy*, *36*(5), 722–741.

255. Hud, M., & Hussinger, K. (2015). The impact of R&D subsidies during the crisis. *Research Policy*, *44*(10), 1844–1855.

256. Hult, G. T. M., Hurley, R. F., & Knight, G. A. (2004). Innovativeness: Its antecedents and impact on business performance. *Industrial Marketing Management*, *33*(5), 429–438.

257. 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.

258. Iñigo, E. A., & Albareda, L. (2016). Understanding sustainable innovation as a complex adaptive system: A systemic approach to the firm. *Journal of Cleaner Production*, *126*, 1–20. https://doi.org/10.1016/j.jclepro.2016.03.036

259. Islam, M., Fremeth, A., & Marcus, A. (2018). Signaling by early stage startups: US government research grants and venture capital funding. *Journal of Business Venturing*, *33*(1), 35–51.

260. Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017). An expert system for diabetes prediction using auto tuned multi-layer perceptron. *2017 Intelligent Systems Conference (IntelliSys)*, 722–728. https://ieeexplore.ieee.org/abstract/document/8324209/?casa_token=Uqjn4-L7b04AAAAA:eqv6Uht2AqKBqZ3lG-hXt7P6ekvDv5TzQ1YN1X55F0GoImgH-NXlPGHynt8CCfUSujYiiV3T

261. Jang, S., Kim, J., & Von Zedtwitz, M. (2017). The importance of spatial agglomeration in product innovation: A microgeography perspective. *Journal of Business Research*, *78*, 143–154. https://doi.org/10.1016/j.jbusres.2017.05.017

262. Janger, J., Schubert, T., Andries, P., Rammer, C., & Hoskens, M. (2017). The EU 2020 innovation indicator: A step forward in measuring innovation outputs and outcomes? *Research Policy*, *46*(1), 30–42.

263. Janney, J. J., & Folta, T. B. (2003). Signaling through private equity placements and its impact on the valuation of biotechnology firms. *Journal of Business Venturing*, *18*(3), 361–380.

264. Janney, J. J., & Folta, T. B. (2006). Moderating effects of investor experience on the signaling value of private equity placements. *Journal of Business Venturing*, *21*(1), 27–44.

265. Jansson, J., Marell, A., & Nordlund, A. (2011). Exploring consumer adoption of a high involvement eco-innovation using value-belief-norm theory. *Journal of Consumer Behaviour*, *10*(1), 51–60. https://doi.org/10.1002/cb.346

266. Japkowicz, N., & Stephen, S. (2002). The class imbalance problem: A systematic study1. *Intelligent Data Analysis*, *6*(5), 429–449. https://doi.org/10.3233/IDA-2002-6504

267. Jiang, C., Yin, C., Tang, Q., & Wang, Z. (2023). The value of official website information in the credit risk evaluation of SMEs. *Journal of Business Research*, *169*, 114290.

268. Jin, J., Zhu, Y., Dong, G., Zhang, Y., Yang, X., Zhang, C., Zhao, T., Yang, Z., Dou, Z., & Wen, J.-R. (2025). *FlashRAG: A Modular Toolkit for Efficient Retrieval-Augmented Generation Research*. https://doi.org/10.1145/3701716.3715313.

269. Jog, D., & Singhal, D. (2024). Greenwashing Understanding Among Indian Consumers and Its Impact on Their Green Consumption. *Global Business Review*, *25*(2), 491–511. https://doi.org/10.1177/0972150920962933

270. Jolink, A., & Niesten, E. (2021). Credibly reducing information asymmetry: Signaling on economic or environmental value by environmental alliances. *Long Range Planning*, *54*(4), 101996.

271. Jolliffe, I. T. (2002). Principal Component Analysis for Special Types of Data. In *Principal Component Analysis* (pp. 338–372). Springer-Verlag. https://doi.org/10.1007/0-387-22440-8_13

272. Jung Moon, S., & Hyun, K. D. (2014). Online Media Relations as an Information Subsidy: Quality of *Fortune 500* Companies’ Websites and Relationships to Media Salience. *Mass Communication and Society*, *17*(2), 258–273. https://doi.org/10.1080/15205436.2013.779716

273. Kahraman, A., & Kazançoğlu, İ. (2019). Understanding consumers’ purchase intentions toward natural-claimed products: A qualitative research in personal care products.

Business Strategy and the Environment, *28*(6), 1218–1233. https://doi.org/10.1002/bse.2312

274. Kaiser, H. F. (1974). An Index of Factorial Simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

275. Kammerer, D. (2009). The effects of customer benefit and regulation on environmental product innovation.: Empirical evidence from appliance manufacturers in Germany. *Ecological Economics*, *68*(8–9), 2285–2295.

276. Kao, C., & Wu, C. (1994). Tests of Dividend Signaling Using the Marsh-Merton Model: A Generalized Friction Approach. *The Journal of Business*, *67*(1), 45. https://doi.org/10.1086/296623

277. Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). Scaling laws for neural language models. *arXiv Preprint arXiv:2001.08361*. https://arxiv.org/pdf/2001.08361/1000

278. Kaplan, S. N., & Schoar, A. (2005). Private Equity Performance: Returns, Persistence, and Capital Flows. *The Journal of Finance*, *60*(4), 1791–1823. https://doi.org/10.1111/j.1540-6261.2005.00780.x

279. Karakaya, E., Hidalgo, A., & Nuur, C. (2014). Diffusion of eco-innovations: A review. *Renewable and Sustainable Energy Reviews*, *33*, 392–399. https://doi.org/10.1016/j.rser.2014.01.083

280. Kemp, R., & Pearson, P. (2007). Final report MEI project about measuring eco-innovation. *UM Merit, Maastricht*, *10*(2), 1–120.

281. Kerr, W. R., & Kominers, S. D. (2015). Agglomerative Forces and Cluster Shapes. *Review of Economics and Statistics*, *97*(4), 877–899. https://doi.org/10.1162/REST_a_00471

282. Ketata, I., Sofka, W., & Grimpe, C. (2015). The role of internal capabilities and firms' environment for sustainable innovation: Evidence for G ermany. *R&d Management*, *45*(1), 60–75.

283. Khanna, M., Deltas, G., & Harrington, D. R. (2009). Adoption of pollution prevention techniques: The role of management systems and regulatory pressures. *Environmental and Resource Economics*, *44*, 85–106.

284. Khattak, M. S. (2020). Does access to domestic finance and international finance contribute to sustainable development goals? Implications for policymakers. *Journal of Public Affairs*, *20*(2), e2024.

285. Kim, E.-H., & Lyon, T. P. (2015). Greenwash vs. Brownwash: Exaggeration and Undue Modesty in Corporate Sustainability Disclosure. *Organization Science*, *26*(3), 705–723. https://doi.org/10.1287/orsc.2014.0949

286. Kim, S., & Schifeling, T. (2016). Varied incumbent behaviors and mobilization for new organizational forms: The rise of triple-bottom line business amid both corporate social responsibility and irresponsibility. *Available at SSRN 2794335*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2794335

287. Kinne, J., & Axenbeck, J. (2018). Web mining of firm websites: A framework for web scraping and a pilot study for Germany. *ZEW Discussion Papers*, *18*. https://madoc.bib.uni-mannheim.de/46518/

288. Kinne, J., & Axenbeck, J. (2020a). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041.

289. Kinne, J., & Axenbeck, J. (2020b). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041. https://doi.org/10.1007/s11192-020-03726-9

290. Kinne, J., & Lenz, D. (2021a). Predicting innovative firms using web mining and deep learning. *PloS One*, *16*(4), e0249071.

291. Kinne, J., & Lenz, D. (2021b). Predicting innovative firms using web mining and deep learning. *PloS One*, *16*(4), e0249071.

292. Kiss, A. N., & Barr, P. S. (2017). New Product Development Strategy Implementation Duration and New Venture Performance: A Contingency-Based Perspective. *Journal of Management*, *43*(4), 1185–1210. https://doi.org/10.1177/0149206314549251

293. Kleinert, S., Bafera, J., Urbig, D., & Volkmann, C. K. (2022). Access Denied: How Equity Crowdfunding Platforms Use Quality Signals to Select New Ventures. *Entrepreneurship Theory and Practice*, *46*(6), 1626–1657. https://doi.org/10.1177/10422587211011945

294. Kleinknecht, A., Reijnen, J. O. N., & Smits, W. (1993). Collecting Literature-based Innovation Output Indicators. The Experience in the Netherlands. In A. Kleinknecht & D. Bain (Eds.), *New Concepts in Innovation Output Measurement* (pp. 42–84). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-22892-8_3

295. Kleinknecht, A., Van Montfort, K., & Brouwer, E. (2002). The Non-Trivial Choice between Innovation Indicators. *Economics of Innovation and New Technology*, *11*(2), 109–121. https://doi.org/10.1080/10438590210899

296. Klette, T. J., Møen, J., & Griliches, Z. (2000). Do subsidies to commercial R&D reduce market failures? Microeconometric evaluation studies. *Research Policy*, *29*(4–5), 471–495.

297. Kline, S. J., & Rosenberg, N. (1986). An Overview of Innovation. In R. Landau & N. Rosenberg (Eds.), *The Positive Sum Strategy: Harnessing Technology for Economic Growth* (pp. 275–305). National Academy Press.

298. Ko, E.-J., & McKelvie, A. (2018). Signaling for more money: The roles of founders' human capital and investor prominence in resource acquisition across different stages of firm development. *Journal of Business Venturing*, *33*(4), 438–454.

299. Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, *35*, 22199–22213.

300. Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM SIGKDD Explorations Newsletter*, *2*(1), 1–15. https://doi.org/10.1145/360402.360406

301. Krebs, J. R., & Dawkins, R. (1984). *Animal signals: Mind-reading and manipulation*. https://philpapers.org/rec/KREASM

302. Kumar, G. D., & Gosul, M. (2011). Web Mining Research and Future Directions. In D. C. Wyld, M. Wozniak, N. Chaki, N. Meghanathan, & D. Nagamalai (Eds.), *Advances in Network Security and Applications* (Vol. 196, pp. 489–496). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-22540-6_47

303. Kundu, S. (2012). An intelligent approach of web data mining. *International Journal on Computer Science and Engineering*, *4*(5), 919.

304. Kuznets, S. (1962). Inventive Activity: Problems of Definition and Measurement. In National Bureau Of Economic Research, *The Rate and Direction of Inventive Activity* (pp. 19–52). Princeton University Press. https://doi.org/10.1515/9781400879762-002

305. Lampel, J., & Shamsie, J. (2000). Critical Push: Strategies for Creating Momentum in the Motion Picture Industry. *Journal of Management*, *26*(2), 233–257. https://doi.org/10.1177/014920630002600204

306. Lampert, C. H., Nickisch, H., & Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. *2009 IEEE Conference on Computer Vision*

and Pattern Recognition, 951–958. https://ieeexplore.ieee.org/abstract/document/5206594/

307. Lanchimba, C., Welsh, D. H., Fadairo, M., & Silva, V.-L. D. (2021). The impact of franchisor signaling on entrepreneurship in emerging markets. *Journal of Business Research*, *131*, 337–348.

308. Last, F., Douzas, G., & Bacao, F. (2017). Oversampling for imbalanced learning based on k-means and smote. arXiv 2017. *arXiv Preprint arXiv:1711.00837*, *2*.

309. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444.

310. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, *36*(4), 1234–1240.

311. Lee, S., Yoon, B., & Park, Y. (2009). An approach to discovering new technology opportunities: Keyword-based patent map approach. *Technovation*, *29*(6–7), 481–497. https://doi.org/10.1016/j.technovation.2008.10.006

312. Leiponen, A., & Helfat, C. E. (2010). Innovation objectives, knowledge sources, and the benefits of breadth. *Strategic Management Journal*, *31*(2), 224–236. https://doi.org/10.1002/smj.807

313. Lester, R. H., Certo, S. T., Dalton, C. M., Dalton, D. R., & Cannella, A. A. (2006). Initial Public Offering Investor Valuations: An Examination of Top Management Team Prestige and Environmental Uncertainty. *Journal of Small Business Management*, *44*(1), 1–26. https://doi.org/10.1111/j.1540-627X.2006.00151.x

314. Levidow, L., & Papaioannou, T. (2018). Which inclusive innovation? Competing normative assumptions around social justice. *Innovation and Development*, *8*(2), 209–226. https://doi.org/10.1080/2157930X.2017.1351605

315. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv Preprint arXiv:1910.13461*.

316. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., & Rocktäschel, T. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

317. Li, D., Zheng, M., Cao, C., Chen, X., Ren, S., & Huang, M. (2017). The impact of legitimacy pressure and corporate profitability on green innovation: Evidence from China top 100. *Journal of Cleaner Production*, *141*, 41–49.

318. Li, H., Su, Y., Cai, D., Wang, Y., & Liu, L. (2022). *A Survey on Retrieval-Augmented Text Generation* (arXiv:2202.01110). arXiv. https://doi.org/10.48550/arXiv.2202.01110

319. Li, Y., Arora, S., Youtie, J., & Shapira, P. (2018). Using web mining to explore Triple Helix influences on growth in small and mid-size firms. *Technovation*, *76*, 3–14.

320. Libaers, D., Hicks, D., & Porter, A. L. (2016). A taxonomy of small firm technology commercialization. *Industrial and Corporate Change*, *25*(3), 371–405.

321. Liles, P. R. (1977). *Sustaining the venture capital firm*. Management Analysis Center.

322. Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

323. Liute, A., & De Giacomo, M. R. (2022). The environmental performance of UK-based B Corp companies: An analysis based on the triple bottom line approach. *Business Strategy and the Environment*, *31*(3), 810–827.

324. Liyanage, S., Greenfield, P. F., & Don, R. (1999). Towards a fourth generation R&D management model-research networks in knowledge management. *International Journal of Technology Management*, *18*(3/4), 372. https://doi.org/10.1504/IJTM.1999.002770

325. Lo Mele, V., Quas, A., Reichert, P., & Romito, S. (2024). Impact orientation and venture capital financing: The interplay of governmental, social impact and traditional venture capital. *Finance Research Letters*, *68*, 105987. https://doi.org/10.1016/j.frl.2024.105987

326. Lopez, B. (2020). Connecting business and sustainable development goals in Spain. *Marketing Intelligence & Planning*, *38*(5), 573–585.

327. Lorentzen, D. G. (2014). Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. *Scientometrics*, *99*(2), 409–445.

328. Lunardon, N., Menardi, G., & Torelli, N. (2014). *ROSE: A package for binary imbalanced learning*. https://digitalcommons.unl.edu/r-journal/418/

329. Maas, K., Schaltegger, S., & Crutzen, N. (2016). Integrating corporate sustainability assessment, management accounting, control, and reporting. *Journal of Cleaner Production*, *136*, 237–248.

330. Madria, S. K., Bhowmick, S. S., Ng, W.-K., & Lim, E. P. (1999). Research Issues in Web Data Mining. In M. Mohania & A. M. Tjoa (Eds.), *DataWarehousing and Knowledge Discovery* (Vol. 1676, pp. 303–312). Springer Berlin Heidelberg. https://doi.org/10.1007/3-540-48298-9_32

331. Magnusson, P., Westjohn, S. A., & Zdravkovic, S. (2011). “What? I thought Samsung was Japanese”: accurate or not, perceived country of origin matters. *International Marketing Review*, *28*(5), 454–472.

332. Maimon, O., & Rokach, L. (Eds.). (2005). *Data Mining and Knowledge Discovery Handbook*. Springer US. https://doi.org/10.1007/978-0-387-09823-4

333. Majumder, M. T. H., Akter, A., & Li, X. (2017). Corporate governance and corporate social disclosures: A meta-analytical review. *International Journal of Accounting & Information Management*, *25*(4), 434–458.

334. Makkonen, T., & Van Der Have, R. P. (2013). Benchmarking regional innovative performance: Composite measures and direct innovation counts. *Scientometrics*, *94*(1), 247–262. https://doi.org/10.1007/s11192-012-0753-2

335. Manget, J., Roche, C., & Münnich, F. (2009). Capturing the Green Advantage. *MIT Sloan Management Review*. https://sloanreview.mit.edu/projects/capturing-the-green-advantage/

336. Marino, M., Lhuillery, S., Parrotta, P., & Sala, D. (2016). Additionality or crowding-out? An overall evaluation of public R&D subsidy on private R&D expenditure. *Research Policy*, *45*(9), 1715–1730.

337. Markham, S. K. (2002). Moving Technologies from Lab To Market. *Research-Technology Management*, *45*(6), 31–42. https://doi.org/10.1080/08956308.2002.11671531

338. Mateut, S. (2018). Subsidies, financial constraints and firm innovative activities in emerging economies. *Small Business Economics*, *50*(1), 131–162. https://doi.org/10.1007/s11187-017-9877-3

339. Mavlanova, T., Benbunan-Fich, R., & Koufaris, M. (2012). Signaling theory and information asymmetry in online commerce. *Information & Management*, *49*(5), 240–247.

340. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). *On Faithfulness and Factuality in Abstractive Summarization* (arXiv:2005.00661). arXiv. https://doi.org/10.48550/arXiv.2005.00661

341. Mazzei, A. (2010). Promoting active communication behaviours through internal communication. *Corporate Communications: An International Journal*, *15*(3), 221–234.

342. Mazzucato, M. (2022). Financing the green new deal. *Nature Sustainability*, *5*(2), 93–94.

343. McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

344. McInnes, L., Healy, J., & Melville, J. (2020). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (arXiv:1802.03426). arXiv. https://doi.org/10.48550/arXiv.1802.03426

345. Menardi, G., & Torelli, N. (2014). Training and assessing classification rules with imbalanced data. *Data Mining and Knowledge Discovery*, *28*(1), 92–122. https://doi.org/10.1007/s10618-012-0295-5

346. Metrick, A., & Yasuda, A. (2011). Venture Capital and Other Private Equity: A Survey: Venture Capital and Other Private Equity. *European Financial Management*, *17*(4), 619–654. https://doi.org/10.1111/j.1468-036X.2011.00606.x

347. Metrick, A., & Yasuda, A. (2021). *Venture capital and the finance of innovation*. John Wiley & Sons. https://books.google.com/books?hl=it&lr=&id=gXcfEAAAQBAJ&oi=fnd&pg=PA7&d

q=Venture+Capital+and+the+Finance+of+Innovation&ots=LrkhC2IwML&sig=oBQf7O8d84Y0jfKKR1_FNzymcVo

348. Michie, J. (1998). Introduction. The Internationalisation of the Innovation Process. *International Journal of the Economics of Business*, *5*(3), 261–277. https://doi.org/10.1080/13571519884387

349. Mikolov, T., Yih, W., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 746–751. https://aclanthology.org/N13-1090.pdf

350. Miloud, T., Aspelund, A., & Cabrol, M. (2012). Startup valuation by venture capitalists: An empirical study. *Venture Capital*, *14*(2–3), 151–174. https://doi.org/10.1080/13691066.2012.667907

351. Miner, G. D., Elder, J., Fast, A., Hill, T., Nisbet, R., & Delen, D. (2012). *Practical text mining and statistical analysis for non-structured text data applications*. Academic Press. https://books.google.com/books?hl=it&lr=&id=SM94BMsy50gC&oi=fnd&pg=PP1&dq=ractical+text+mining+and+statistical+analysis+for+non-structured+text+data+applications.&ots=jdFXWkmGX8&sig=XpTprFYAiApAeywB7fXyoLkSdE0

352. Mirończuk, M. M., & Protasiewicz, J. (2020). Recognising innovative companies by using a diversified stacked generalisation method for website classification. *Applied Intelligence*, *50*(1), 42–60. https://doi.org/10.1007/s10489-019-01509-1

353. Mirtsch, M., Kinne, J., & Blind, K. (2020). Exploring the adoption of the international information security management system standard ISO/IEC 27001: A web mining-based analysis. *IEEE Transactions on Engineering Management*, *68*(1), 87–100.

354. Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a Theory of Stakeholder Identification and Salience: Defining the Principle of Who and What Really Counts. *The Academy of Management Review*, *22*(4), 853. https://doi.org/10.2307/259247

355. Montiel, I., Husted, B. W., & Christmann, P. (2012). Using private management standard certification to reduce information asymmetries in corrupt environments. *Strategic Management Journal*, *33*(9), 1103–1113. https://doi.org/10.1002/smj.1957

356. Moratis, L. (2018). Signalling responsibility? Applying signalling theory to the ISO 26000 standard for social responsibility. *Sustainability*, *10*(11), 4172.

357. Muñoz, P., & Cohen, B. (2018). Sustainable Entrepreneurship Research: Taking Stock and looking ahead. *Business Strategy and the Environment*, *27*(3), 300–322. https://doi.org/10.1002/bse.2000

358. Nanda, R., Younge, K., & Fleming, L. (2015). Innovation and entrepreneurship in renewable energy. *The Changing Frontier: Rethinking Science and Innovation Policy*, *199*. https://www.degruyter.com/document/doi/10.7208/9780226286860-009/pdf?licenseType=restricted

359. Nasraoui, O., Rojas, C., & Cardona, C. (2006). A framework for mining evolving trends in Web data streams using dynamic learning and retrospective validation. *Computer Networks*, *50*(10), 1488–1512. https://doi.org/10.1016/j.comnet.2005.10.021

360. Nathan, M., & Rosso, A. (2022). Innovative events: Product launches, innovation and firm performance. *Research Policy*, *51*(1), 104373.

361. Nekhili, M., Nagati, H., Chtioui, T., & Rebolledo, C. (2017). Corporate social responsibility disclosure and market value: Family versus nonfamily firms. *Journal of Business Research*, *77*, 41–52.

362. Nelson, R. R. (1959). The Simple Economics of Basic Scientific Research. *Journal of Political Economy*, *67*(3), 297–306. https://doi.org/10.1086/258177

363. Nelson, R. R., & Winter, S. G. (1982). *An evolutionary theory of economic change*. Harvard University Press.

364. Nemet, G. F., & Kammen, D. M. (2007). U.S. energy research and development: Declining investment, increasing need, and the feasibility of expansion. *Energy Policy*, *35*(1), 746–755. https://doi.org/10.1016/j.enpol.2005.12.012

365. Neville, C., & Lucey, B. M. (2022). Financing Irish high-tech SMEs: The analysis of capital structure. *International Review of Financial Analysis*, *83*, 102219.

366. Nguyen, H. M., Cooper, E. W., & Kamei, K. (2011). Borderline over-sampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms*, *3*(1), 4. https://doi.org/10.1504/IJKESDP.2011.039875

367. Nigam, N., Benetti, C., & Johan, S. A. (2020). Digital start-up access to venture capital financing: What signals quality? *Emerging Markets Review*, *45*, 100743. https://doi.org/10.1016/j.ememar.2020.100743

368. Norouzi, M., Mikolov, T., Bengio, S., Singer, Y., Shlens, J., Frome, A., Corrado, G. S., & Dean, J. (2014). *Zero-Shot Learning by Convex Combination of Semantic Embeddings* (arXiv:1312.5650). arXiv. https://doi.org/10.48550/arXiv.1312.5650

369. Nyilasy, G., Gangadharbatla, H., & Paladino, A. (2014). Perceived Greenwashing: The Interactive Effects of Green Advertising and Corporate Environmental Performance on

Consumer Reactions. *Journal of Business Ethics*, *125*(4), 693–707. https://doi.org/10.1007/s10551-013-1944-3

370. OECD. (1992). *OECD Proposed Guidelines for Collecting and Interpreting Technological Innovation Data: Oslo Manual*. OECD Publishing. https://doi.org/10.1787/87954fc6-en

371. OECD. (1997). *Proposed Guidelines for Collecting and Interpreting Technological Innovation Data: Oslo Manual* (2nd ed.). OECD Publishing. https://doi.org/10.1787/9789264192263-en

372. OECD. (2005). *Oslo Manual: Guidelines for Collecting and Interpreting Innovation Data* (3rd ed.). OECD Publishing.

373. OECD. (2010). *Eco-Innovation in Industry: Enabling Green Growth*. OECD. https://doi.org/10.1787/9789264077225-en

374. OECD. (2018). *Oslo Manual 2018: Guidelines for Collecting, Reporting and Using Data on Innovation, 4th Edition*. Organisation for Economic Co-operation and Development. https://www.oecd-ilibrary.org/science-and-technology/oslo-manual-2018_9789264304604-en

375. O’Neil, I., & Ucbasaran, D. (2016). Balancing “what matters to me” with “what matters to them”: Exploring the legitimation process of environmental entrepreneurs. *Journal of Business Venturing*, *31*(2), 133–152.

376. Ou, J.-C., Lee, C.-H., & Chen, M.-S. (2008). Efficient algorithms for incremental Web log mining with dynamic thresholds. *The VLDB Journal*, *17*(4), 827–845. https://doi.org/10.1007/s00778-006-0043-9

377. Ozmel, U., Robinson, D. T., & Stuart, T. E. (2013). Strategic alliances, venture capital, and exit decisions in early stage high-tech firms. *Journal of Financial Economics*, *107*(3), 655–670.

378. Packard, M. D., Clark, B. B., & Klein, P. G. (2017). Uncertainty Types and Transitions in the Entrepreneurial Process. *Organization Science*, *28*(5), 840–856. https://doi.org/10.1287/orsc.2017.1143

379. Paelman, V., Van Cauwenberge, P., & Vander Bauwhede, H. (2020). Effect of B Corp certification on short-term growth: European evidence. *Sustainability*, *12*(20), 8459.

380. Park, N. K., & Mezias, J. M. (2005). Before and after the technology sector crash: The effect of environmental munificence on stock market response to alliances of e-commerce firms. *Strategic Management Journal*, *26*(11), 987–1007. https://doi.org/10.1002/smj.489

381. Pearson, K. (1901). LIII. *On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, *2*(11), 559–572. https://doi.org/10.1080/14786440109462720

382. Peneder, M. (2010). The impact of venture capital on innovation behaviour and firm growth. *Venture Capital*, *12*(2), 83–107. https://doi.org/10.1080/13691061003643250

383. Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://aclanthology.org/D14-1162.pdf

384. Pirola-Merlo, A., & Mann, L. (2004). The relationship between individual creativity and team creativity: Aggregating across people and time. *Journal of Organizational Behavior*, *25*(2), 235–257. https://doi.org/10.1002/job.240

385. Pisano, G. P. (2015). You need an innovation strategy. *Harvard Business Review*, *93*(6), 44–54.

386. Piva, E., & Rossi-Lamastra, C. (2018). Human capital signals and entrepreneurs' success in equity crowdfunding. *Small Business Economics*, *51*(3), 667–686. https://doi.org/10.1007/s11187-017-9950-y

387. Ponta, L., Puliga, G., & Manzini, R. (2021). A measure of innovation performance: The Innovation Patent Index. *Management Decision*, *59*(13), 73–98. https://doi.org/10.1108/MD-05-2020-0545

388. Porter, M. E. (1990). *The Competitive Advantage of Nations*. Free Press.

389. Porter, M. E., & Linde, C. van der. (1995). Toward a new conception of the environment-competitiveness relationship. *Journal of Economic Perspectives*, *9*(4), 97–118.

390. Puri, M., & Zarutskie, R. (2012). On the Life Cycle Dynamics of Venture-Capital- and Non-Venture-Capital-Financed Firms. *The Journal of Finance*, *67*(6), 2247–2293. https://doi.org/10.1111/j.1540-6261.2012.01786.x

391. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.

392. Rae, J. W., Borgeaud, S., Cai, T., Millican, K., Hoffmann, J., Song, F., Aslanides, J., Henderson, S., Ring, R., Young, S., Rutherford, E., Hennigan, T., Menick, J., Cassirer, A., Powell, R., Driessche, G. van den, Hendricks, L. A., Rauh, M., Huang, P.-S., … Irving, G. (2022). *Scaling Language Models: Methods, Analysis & Insights from Training Gopher* (arXiv:2112.11446). arXiv. https://doi.org/10.48550/arXiv.2112.11446

393. Ragozzino, R., & Blevins, D. (2023). An Investigation of the Attention Effects of Venture Capitalist Backing on Entrepreneurial Firms. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4357858

394. Ramaswami, A., Dreher, G. F., Bretz, R., & Wiethoff, C. (2010). GENDER, MENTORING, AND CAREER SUCCESS: THE IMPORTANCE OF ORGANIZATIONAL CONTEXT. *Personnel Psychology*, *63*(2), 385–405. https://doi.org/10.1111/j.1744-6570.2010.01174.x

395. Rammer, C., & Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators-a literature review. *Technological Forecasting and Social Change*, *197*, 122874.

396. Rao, A. R., Qu, L., & Ruekert, R. W. (1999). Signaling Unobservable Product Quality through a Brand Ally. *Journal of Marketing Research*, *36*(2), 258–268. https://doi.org/10.1177/002224379903600209

397. Razaghzadeh Bidgoli, M., Raeesi Vanani, I., & Goodarzi, M. (2024). Predicting the success of startups using a machine learning approach. *Journal of Innovation and Entrepreneurship*, *13*(1), 80. https://doi.org/10.1186/s13731-024-00436-x

398. Reeb, D. M., & Zhao, W. (2020). Patents Do Not Measure Innovation Success. *Critical Finance Review*, *9*(1–2), 157–199. https://doi.org/10.1561/104.00000087

399. Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv Preprint arXiv:1908.10084*. https://arxiv.org/abs/1908.10084

400. Rennings, K. (2000). Redefining innovation—Eco-innovation research and the contribution from ecological economics. *Ecological Economics*, *32*(2), 319–332.

401. Reuber, A. R., & Fischer, E. (2011). International entrepreneurship in internet-enabled markets. *Journal of Business Venturing*, *26*(6), 660–679.

402. Reuer, J. J., Tong, T. W., & Wu, C.-W. (2012). A Signaling Theory of Acquisition Premiums: Evidence from IPO Targets. *Academy of Management Journal*, *55*(3), 667–683. https://doi.org/10.5465/amj.2010.0259

403. Reynolds, O., Sheehan, M., & Hilliard, R. (2018). Exploring strategic agency in sustainability-oriented entrepreneur legitimation. *International Journal of Entrepreneurial Behavior & Research*, *24*(2), 429–450.

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

405. Roma, P., Vasi, M., & Kolympiris, C. (2021). On the signaling effect of reward-based crowdfunding: (When) do later stage venture capitalists rely more on the crowd than their peers? *Research Policy*, *50*(6), 104267. https://doi.org/10.1016/j.respol.2021.104267

406. Romi, A., Cook, K. A., & Dixon-Fowler, H. R. (2018). The influence of social responsibility on employee productivity and sales growth: Evidence from certified B corps. *Sustainability Accounting, Management and Policy Journal*.

407. Rothaermel, F. T., & Deeds, D. L. (2004). Exploration and exploitation alliances in biotechnology: A system of new product development. *Strategic Management Journal*, *25*(3), 201–221. https://doi.org/10.1002/smj.376

408. Salton, G. (1983). Some research problems in automatic information retrieval. *ACM SIGIR Forum*, *17*(4), 252–263. https://doi.org/10.1145/1013230.511830

409. Samila, S., & Sorenson, O. (2011). Venture capital, entrepreneurship, and economic growth. *The Review of Economics and Statistics*, *93*(1), 338–349.

410. SanMiguel, P., Pérez-Bou, S., Sádaba, T., & Mir-Bernal, P. (2021). How to communicate sustainability: From the corporate Web to E-commerce. The case of the fashion industry. *Sustainability*, *13*(20), 11363.

411. Savage, G. T., Bunn, M. D., Gray, B., Xiao, Q., Wang, S., Wilson, E. J., & Williams, E. S. (2010). Stakeholder Collaboration: Implications for Stakeholder Theory and Practice. *Journal of Business Ethics*, *96*(S1), 21–26. https://doi.org/10.1007/s10551-011-0939-1

412. Savaneviciene, A., Venckuviene, V., & Girdauskiene, L. (2015). Venture capital a catalyst for start-ups to overcome the "Valley of death": Lithuanian case. *Procedia Economics and Finance*, *26*, 1052–1059.

413. Schaefer, H. M., & Ruxton, G. D. (2011). *Plant-animal communication*. OUP Oxford. https://books.google.com/books?hl=it&lr=&id=4AU8HZ3Ib1cC&oi=fnd&pg=PT26&dq=Schaefer+and+Ruxton+2011&ots=gK_pJA4jFR&sig=Cjz29dqWFGcuvRGg3mwa-LNE-WI

414. Schaltegger, S., Beckmann, M., & Hockerts, K. (2018). Collaborative entrepreneurship for sustainability. Creating solutions in light of the UN sustainable development goals. *International Journal of Entrepreneurial Venturing*, *10*(2), 131. https://doi.org/10.1504/IJEV.2018.092709

415. Schlange, L. E. (2006). Stakeholder Identification in Sustainability Entrepreneurship. *Greener Management International*, *2006*(55), 13–32. https://doi.org/10.9774/GLEAF.3062.2006.au.00004

416. Schumpeter, J. (1934). The theory of economic development Harvard University Press. *Cambridge, MA*.

417. Sen, S., & Bhattacharya, C. B. (2001). Does Doing Good Always Lead to Doing Better? Consumer Reactions to Corporate Social Responsibility. *Journal of Marketing Research*, *38*(2), 225–243. https://doi.org/10.1509/jmkr.38.2.225.18838

418. Sethi, S. P. (1975). Dimensions of Corporate Social Performance: An Analytical Framework. *California Management Review*, *17*(3), 58–64. https://doi.org/10.2307/41162149

419. Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC Medical Informatics and Decision Making*, *18*(1), 46. https://doi.org/10.1186/s12911-018-0639-1

420. Sharma, K., & Bhatnagar, V. (2011). Private and secure hyperlink navigability assessment in web mining information system. *International Journal on Computer Science and Engineering*, *3*(6), 2245–2250.

421. Sharma, S., & Henriques, I. (2005). Stakeholder influences on sustainability practices in the Canadian forest products industry. *Strategic Management Journal*, *26*(2), 159–180. https://doi.org/10.1002/smj.439

422. Shin, M., Bae, J., & Ozmel, U. (2025). Effect of venture capital investment horizon on new product development: Evidence from the medical device sector. *Journal of Business Venturing*, *40*(1), 106454.

423. Siano, A., Conte, F., Amabile, S., Vollero, A., & Piciocchi, P. (2016). Communicating sustainability: An operational model for evaluating corporate websites. *Sustainability*, *8*(9), 950.

424. Singh, A., Ehtesham, A., Kumar, S., & Khoei, T. T. (2025). *Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG* (arXiv:2501.09136). arXiv. https://doi.org/10.48550/arXiv.2501.09136

425. Skudiene, V., & Auruskeviciene, V. (2012). The contribution of corporate social responsibility to internal employee motivation. *Baltic Journal of Management*, *7*(1), 49–67.

426. Smith, D. F., & Florida, R. (2013). Venture capital's role in regional innovation systems: Historical perspective and recent evidence. In *Regional innovation, knowledge and global change* (pp. 205–227). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315066653-17/venture-capital-role-regional-innovation-systems-historical-perspective-recent-evidence-donald-smith-richard-florida

427. Smith, M. J., & Harper, D. G. (1995). Animal signals: Models and terminology. *Journal of Theoretical Biology*, *177*(3), 305–311.

428. Socher, R., Ganjoo, M., Manning, C. D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/2d6cc4b2d139a53512fb8cbb3086ae2e-Abstract.html

429. Solé Udina, E., Domingo-Perez, S., & Amat, O. (2022). Biotechnology firms, signals, and venture capital investment. *Intangible Capital*, *18*(3), 350. https://doi.org/10.3926/ic.1978

430. Solow, R. M. (1957). Technical change and the aggregate production function. *The Review of Economics and Statistics*, *39*(3), 312–320.

431. Spence, M. (1973). Job market signaling. *Quarterly Journal of Economics*, *87*, 354–374.

432. Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *ACM SIGKDD Explorations Newsletter*, *1*(2), 12–23. https://doi.org/10.1145/846183.846188

433. Steigenberger, N., Garz, M., & Cyron, T. (2024). Signaling theory in entrepreneurial fundraising and crowdfunding research. *Journal of Small Business Management*, 1–26. https://doi.org/10.1080/00472778.2024.2412710

434. Steigenberger, N., & Wilhelm, H. (2018). Extending Signaling Theory to Rhetorical Signals: Evidence from Crowdfunding. *Organization Science*, *29*(3), 529–546. https://doi.org/10.1287/orsc.2017.1195

435. Stern, N., & Valero, A. (2021). Innovation, growth and the transition to net-zero emissions. *Research Policy*, *50*(9), 104293.

436. Stevenson, R., Kier, A. S., & Taylor, S. G. (2021). Do policy makers take grants for granted? The efficacy of public sponsorship for innovative entrepreneurship. *Strategic Entrepreneurship Journal*, *15*(2), 231–253. https://doi.org/10.1002/sej.1376

437. Stock, G. N., Greis, N. P., & Fischer, W. A. (2002). Firm size and dynamic technological innovation. *Technovation*, *22*(9), 537–549.

438. Stuart, T. E., Hoang, H., & Hybels, R. C. (1999). Interorganizational Endorsements and the Performance of Entrepreneurial Ventures. *Administrative Science Quarterly*, *44*(2), 315–349. https://doi.org/10.2307/2666998

439. Su, W., Peng, M. W., Tan, W., & Cheung, Y.-L. (2016). The signaling effect of corporate social responsibility in emerging economies. *Journal of Business Ethics*, *134*, 479–491.

440. Sun, S., Johanis, M., & Rychtář, J. (2020). Costly signalling theory and dishonest signalling. *Theoretical Ecology*, *13*(1), 85–92. https://doi.org/10.1007/s12080-019-0429-0

441. Taques, F. H., López, M. G., Basso, L. F., & Areal, N. (2021). Indicators used to measure service innovation and manufacturing innovation. *Journal of Innovation & Knowledge*, *6*(1), 11–26.

442. Tariq, A., Badir, Y. F., Tariq, W., & Bhutta, U. S. (2017). Drivers and consequences of green product and process innovation: A systematic review, conceptual framework, and future outlook. *Technology in Society*, *51*, 8–23.

443. Terán-Yépez, E., Marín-Carrillo, G. M., del Pilar Casado-Belmonte, M., & de las Mercedes Capobianco-Uriarte, M. (2020). Sustainable entrepreneurship: Review of its evolution and new trends. *Journal of Cleaner Production*, *252*, 119742.

444. Terwiesch, C., & Xu, Y. (2008). Innovation Contests, Open Innovation, and Multiagent Problem Solving. *Management Science*, *54*(9), 1529–1543. https://doi.org/10.1287/mnsc.1080.0884

445. Testa, S., Massa, S., Martini, A., & Appio, F. P. (2020). Social media-based innovation: A review of trends and a research agenda. *Information & Management*, *57*(3), 103196. https://doi.org/10.1016/j.im.2019.103196

446. Tether, B. S. (2002). Who co-operates for innovation, and why: An empirical analysis. *Research Policy*, *31*(6), 947–967.

447. Thelwall, M. (2009). *Introduction to Webometrics: Quantitative Web Research for the Social Sciences*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02261-6

448. Thies, F., Wessel, M., & Benlian, A. (2018). Network effects on crowdfunding platforms: Exploring the implications of relaxing input control. *Information Systems Journal*, *28*(6), 1239–1262. https://doi.org/10.1111/isj.12194

449. Thompson, P., Nawaz, R., Korkontzelos, I., Black, W., McNaught, J., & Ananiadou, S. (2013). News search using discourse analytics. *2013 Digital Heritage International Congress (DigitalHeritage)*, 597–604. https://doi.org/10.1109/DigitalHeritage.2013.6743801

450. Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2017). Enriching news events with meta-knowledge information. *Language Resources and Evaluation*, *51*(2), 409–438. https://doi.org/10.1007/s10579-016-9344-9

451. Todorov, L., Shopova, M., Panteleeva, I. M., & Todorova, L. (2024). Innovation Metrics: A Critical Review. *Economies*, *12*(12), 327.

452. Topaler, B., & Adar, G. (2025). The role of signals in new venture financing in the context of an emerging market: A configurational approach. *International Journal of Emerging Markets*, *20*(1), 407–427. https://doi.org/10.1108/IJOEM-08-2022-1234

453. Urbański, M., & Ul Haque, A. (2020). Are you environmentally conscious enough to differentiate between greenwashed and sustainable items? A global consumers perspective. *Sustainability*, *12*(5), 1786.

454. Van den Heuvel, M., & Popp, D. (2023). The role of venture capital and governments in clean energy: Lessons from the first cleantech bubble. *Energy Economics*, *124*, 106877.

455. Vanacker, T., Forbes, D. P., Knockaert, M., & Manigart, S. (2020). Signal Strength, Media Attention, and Resource Mobilization: Evidence from New Private Equity Firms. *Academy of Management Journal*, *63*(4), 1082–1105. https://doi.org/10.5465/amj.2018.0356

456. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, \Lukasz, & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/7181-attention-is-all

457. Vayansky, I., & Kumar, S. A. (2020). A review of topic modeling methods. *Information Systems*, *94*, 101582.

458. Venkataraman, S. (2004). Regional transformation through technological entrepreneurship. *Journal of Business Venturing*, *19*(1), 153–167.

459. Virkkala, S., & Mariussen, Å. (2021). Networks of Innovation: Measuring Structure and Dynamics between and within Helices, Regions and Spatial Levels. Empirical Evidence from the Baltic Sea Region. *Triple Helix*, *8*(2), 282–328. https://doi.org/10.1163/21971927-bja10019

460. Vladimirov, Z., & Williams, A. M. (2018). Hotel innovations and performance–the mediating role of staff related innovations. *Tourism Management Perspectives*, *28*, 166–178.

461. Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, *104*, 106189. https://doi.org/10.1016/j.chb.2019.106189

462. Wang, H., Ma, B., & Bai, R. (2020). The spillover effect of greenwashing behaviours: An experimental approach. *Marketing Intelligence & Planning*, *38*(3), 283–295. https://doi.org/10.1108/MIP-01-2019-0006

463. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., & Zhou, M. (2020). Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, *33*, 5776–5788.

464. Warner, A. G., Fairbank, J. F., & Steensma, H. K. (2006). Managing Uncertainty in a Formal Standards-Based Industry: A Real Options Perspective on Acquisition Timing. *Journal of Management*, *32*(2), 279–298. https://doi.org/10.1177/0149206305280108

465. Wasikowski, M., & Chen, X. (2010). Combating the Small Sample Class Imbalance Problem Using Feature Selection. *IEEE Transactions on Knowledge and Data Engineering*, *22*(10), 1388–1400. https://doi.org/10.1109/TKDE.2009.187

466. WCED, S. W. S. (1987). World commission on environment and development. *Our Common Future*, *17*(1), 1–91.

467. Weale, A. (1992). *The new politics of pollution*. Manchester University Press. https://books.google.com/books?hl=it&lr=&id=99JRAQAAIAAJ&oi=fnd&pg=PR7&dq=weale+1992+triple+bottom+line&ots=hLEGGwKUrS&sig=WHlkV_gswckdQTeezJ1_u11xr-o

468. Wehnert, P., Baccarella, C. V., & Beckmann, M. (2019). In crowdfunding we trust? Investigating crowdfunding success as a signal for enhancing trust in sustainable product features. *Technological Forecasting and Social Change*, *141*, 128–137.

469. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent Abilities of Large Language Models* (arXiv:2206.07682). arXiv. https://doi.org/10.48550/arXiv.2206.07682

470. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837.

471. Wellalage, N. H., & Fernandez, V. (2019). Innovation and SME finance: Evidence from developing countries. *International Review of Financial Analysis*, *66*, 101370.

472. Wesley II, C. L., Kong, D. T., Lubojacky, C. J., Saxton, M. K., & Saxton, T. (2022). Will the startup succeed in your eyes? Venture evaluation of resource providers during entrepreneurs' informational signaling. *Journal of Business Venturing*, *37*(5), 106229.

473. Wheelwright, S. C., & Clark, K. B. (1992). *Revolutionizing product development: Quantum leaps in speed, efficiency, and quality*. Simon and Schuster. https://books.google.com/books?hl=it&lr=&id=s3Ivkou-BVcC&oi=fnd&pg=PR11&dq=Revolutionizing+Product+Development&ots=0cSeWVMmdI&sig=0zuppnuYjkh63AVj4eIWLRFtUXQ

474. Wu, T., Yang, S., & Tan, J. (2020). Impacts of government R&D subsidies on venture capital and renewable energy investment–an empirical study in China. *Resources Policy*, *68*, 101715.

475. Xian, Y., Schiele, B., & Akata, Z. (2017). Zero-shot learning-the good, the bad and the ugly. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4582–4591. http://openaccess.thecvf.com/content_cvpr_2017/html/Xian_Zero-Shot_Learning_-_CVPR_2017_paper.html

476. Yalabik, B., & Fairchild, R. J. (2011). Customer, regulatory, and competitive pressure as drivers of environmental innovation. *International Journal of Production Economics*, *131*(2), 519–527.

477. Ye, Z., Geng, Y., Chen, J., Chen, J., Xu, X., Zheng, S., Wang, F., Zhang, J., & Chen, H. (2020). Zero-shot text classification via reinforced self-training. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3014–3024. https://aclanthology.org/2020.acl-main.272/

478. Yigitcanlar, T., Sabatini-Marques, J., Kamruzzaman, Md., Camargo, F., Moreira da-Costa, E., Ioppolo, G., & Palandi, F. E. D. (2018). Impact of funding sources on innovation: Evidence from Brazilian software companies. *R&D Management*, *48*(4), 460–484. https://doi.org/10.1111/radm.12323

479. Yildiz, H., Tahali, S., & Trichina, E. (2023). The adoption of the green label by SMEs in the hotel sector: A leverage for reassuring their customers. *Journal of Enterprise Information Management*.

480. Yin, W., Hay, J., & Roth, D. (2019). Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv Preprint arXiv:1909.00161*.

481. Youtie, J., Hicks, D., Shapira, P., & Horsley, T. (2012). Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. *Technology Analysis & Strategic Management*, *24*(10), 981–995. https://doi.org/10.1080/09537325.2012.724163

482. Zahavi, A. (1975). Mate selection—A selection for a handicap. *Journal of Theoretical Biology*, *53*(1), 205–214.

483. Zahavi, A. (1977). The cost of honesty. *Journal of Theoretical Biology*, *67*(3), 603–605. https://doi.org/10.1016/0022-5193(77)90061-3

484. Żbikowski, K., & Antosiuk, P. (2021). A machine learning, bias-free approach for predicting business success using Crunchbase data. *Information Processing & Management*, *58*(4), 102555.

485. Zerbini, F. (2017). CSR initiatives as market signals: A review and research agenda. *Journal of Business Ethics*, *146*(1), 1–23.

486. Zhang, J., Zi, S., Shao, P., & Xiao, Y. (2020). The value of corporate social responsibility during the crisis: Chinese evidence. *Pacific-Basin Finance Journal*, *64*, 101432.

487. Zhang, L., Guo, Y., & Sun, G. (2019). How patent signals affect venture capital: The evidence of bio-pharmaceutical start-ups in China. *Technological Forecasting and Social Change*, *145*, 93–104. https://doi.org/10.1016/j.techfore.2019.05.013

488. Zhang, L., Li, D., Cao, C., & Huang, S. (2018). The influence of greenwashing perception on green purchasing intentions: The mediating role of green word-of-mouth and moderating role of green concern. *Journal of Cleaner Production*, *187*, 740–750.

489. Zhang, Q., & Segall, R. S. (2008). WEB MINING: A SURVEY OF CURRENT RESEARCH, TECHNIQUES, AND SOFTWARE. *International Journal of Information Technology & Decision Making*, *07*(04), 683–720. https://doi.org/10.1142/S0219622008003150

490. Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F., & Shi, S. (2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models* (arXiv:2309.01219). arXiv. https://doi.org/10.48550/arXiv.2309.01219

491. 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

492. Zhao, C., Jiang, Y., Qiu, Y., Zhang, H., & Yang, W.-Y. (2023). Differentiable Retrieval Augmentation via Generative Language Modeling for E-commerce Query Intent

Classification. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4445–4449. https://doi.org/10.1145/3583780.3615210

493. Zhou, Y., Park, S., Zhang, J. Z., & Ferreira, J. J. (2023). How do innovative internet tech startups attract venture capital financing? *Journal of Management & Organization*, 1–22. https://doi.org/10.1017/jmo.2023.39

494. Zhu, J., & Hua, W. (2017). Visualizing the knowledge domain of sustainable development research between 1987 and 2015: A bibliometric analysis. *Scientometrics*, *110*(2), 893–914.

495. Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2020). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, *109*(1), 43–76.

496. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2020). *Fine-Tuning Language Models from Human Preferences* (arXiv:1909.08593). arXiv. https://doi.org/10.48550/arXiv.1909.08593

497. Zúñiga-Vicente, J. Á., Alonso-Borrego, C., Forcadell, F. J., & Galán, J. I. (2014). ASSESSING THE EFFECT OF PUBLIC SUBSIDIES ON FIRM R&D INVESTMENT: A SURVEY. *Journal of Economic Surveys*, *28*(1), 36–67. https://doi.org/10.1111/j.1467-6419.2012.00738.x

.

APPENDICES

APPENDIX A SUPPLEMENTAL MATERIAL ARTICLE 1

Table A. 1 Variable description

Dependent variable			
BLabEnvIndex	Continuous	It assesses the company's environmental practices and commitment to sustainability*, with scores ranging from 0 for companies performing poorly to a maximum of 66.1*	
Control variables			
CommInd	Continuous	*It assesses the company's influence on the external communities where it operates, encompassing aspects such as diversity, economic contributions, civic participation, and the impact on the supply chain, with scores ranging from 6.5 for companies performing poorly to a maximum of 115.2*	*ln(CommInd+1)*
GovernInd	Continuous	*It assesses the company's overall mission, ethics, accountability and transparency, with scores ranging from 4.1 for companies performing poorly to a maximum of 24.3*	
d2015&2019	Dummy (omitted)	*Dummy variable with value 1 if the firm has done the assessment test in the year 2015 or 2019, and 0 otherwise.*	
d2016	Dummy	*Dummy variable with value 1 if the firm has done the assessment test in the year 2016, and 0 otherwise.*	
d2017	Dummy	*Dummy variable with value 1 if the firm has done the assessment test in the year 2017, and 0 otherwise.*	
d2018	Dummy	*Dummy variable with value 1 if the firm has done the assessment test in the year 2018, and 0 otherwise.*	
d2020	Dummy	*Dummy variable with value 1 if the firm has done the assessment test in the year 2020, and 0 otherwise.*	
d2021	Dummy	*Dummy variable with value 1 if the firm has done the assessment test in the year 2021, and 0 otherwise.*	
Web-based environment culture indicators (from the PCA analysis)			
FinMan	Continuous	*Financial and Management aspects related to the web-based environmental culture of the firm*	
WatLand	Continuous	*Web-based environmental culture indicator of the company's impact on water and land*	
EnergyEff	Continuous	*Web-based environmental culture indicator of the company's energy efficiency*	*ln(EnergyEff+11)*
ComImp	Continuous	*Web-based environmental culture indicator of the company impact on the community*	
Agri	Continuous	*Web-based environmental culture indicator of the company impact on agriculture practice and process*	*ln(Agri+11)*
ManTransSaf	Continuous	*Web-based environmental culture indicator of the company process put in place to reduce the impact on the environment of manufacturing and transportation process*	*ln((ManTransSaf+5)*10+1)*

Table A. 1 Variable description (continued)

Other independent variables		
dCanada	Dummy	Dummy variable that takes the value 1 if the firm is located in Canada, and 0 otherwise (it is an American company)
dmicro	Dummy (omitted)	Dummy variable with value 1 if the firm has 0 to 9 employees, and 0 otherwise.
dsmall	Dummy	Dummy variable with value 1 if the firm has 10 to 49 employees, and 0 otherwise.
dmedium	Dummy	Dummy variable with value 1 if the firm has 50 to 250 employees, and 0 otherwise.
dlarge	Dummy	Dummy variable that takes the value 1 if the firm has more than 250 employees, and 0 otherwise.
dAgricul	Dummy	Dummy variable that takes the value 1 if the firm in the Agriculture industry category, and 0 otherwise.
dBuild	Dummy	Dummy variable with value 1 if the firm in the Building industry category, and 0 otherwise.
dConsPdct	Dummy (omitted)	Dummy variable that takes the value 1 if the firm in Consumer Products & Services industry category, and 0 otherwise.
dEducationTr	Dummy	Dummy variable that takes the value 1 if the firm in the Education & Training Services industry category, and 0 otherwise.
dEnergyEnv	Dummy	Dummy variable that takes the value 1 if the firm operates in the Energy & Environmental Services industry category, and 0 otherwise.
dFinLegserv	Dummy	Dummy variable with value 1 if the firm operates in the Financial &Legal services industry category, and 0 otherwise.
dHealthHuman	Dummy	Dummy variable with value 1 if the firm operates in the Health & Human Services industry category, and 0 otherwise.
dMedRestHosp	Dummy	Dummy variable with value 1 if the firm operates in the Media, Restaurant, Hospitality & Travel industry category, and 0 otherwise.
dRetTransLog	Dummy	Dummy variable with value 1 if the firm operates in the Retail, Transportation & Logistics industry category, and 0 otherwise.
dBusinessPro	Dummy	Dummy variable with value 1 if the firm operates in the Business Products & Services industry category, and 0 otherwise.

Table A. 2 Descriptive statistics

Variables	**Mean**	**Median**	**Std. Dev.**	**Min**	**Max**	**Skewness**	**Kurtosis**
ln (*CommInd*+1)	3.308	3.245	0.429	2.015	4.755	0.2703664	2.76969
CommInd	29.061	24.650	14.012	6.500	115.200	1.396835	5.568
GovernInd	14.343	14.800	3.826	4.100	24.300	-0.308	2.790
d2016	0.022	0	0.145	0	1	6.584	44.355
d2017	0.159	0	0.366	0	1	1.863	4.472
d2018	0.260	0	0.439	0	1	1.095	2.199
d2020	0.198	0	0.399	0	1	1.517	3.301
d2021	0.054	0	0.226	0	1	3.948	16.590
d2015&2019 (omitted group)	0.307	0	0.461	0	1	0.839	1.703
BLabEnvIndex	16.841	12.100	12.587	0	66.100	0.985	3.253
dmicro (omitted group)	0.458	0	0.498	0	1	0.170	1.029
dsmall	0.332	0	0.471	0	1	0.716	1.512
dmedium	0.158	0	0.365	0	1	1.879	4.530
dlarge	0.053	0	0.224	0	1	3.984	16.87
dAgricul	0.027	0	0.162	0	1	5.833	35.028
dBuild	0.036	0	0.186	0	1	4.979	25.787
dConsumPdt (omitted group)	0.283	0	0.451	0	1	0.964	1.930
dEducationTr	0.032	0	0.177	0	1	5.279	28.867
dEnergyEnvir	0.048	0	0.213	0	1	4.242	18.994
dFinLegservic	0.128	0	0.334	0	1	2.228	5.964
dHealthHuman	0.037	0	0.189	0	1	4.910	25.112
dMedRestHosp	0.031	0	0.172	0	1	5.448	30.679
dRetTransLog	0.028	0	0.165	0	1	5.730	33.835
dBusinessPro	0.350	0	0.477	0	1	0.627	1.393
dCanada	0.176	0	0.381	0	1	1.705	3.905
FinMan	0	-0.328	1	-2.190	4.516	1.416	4.976
WatLand	0	-0.115	1	-3.062	5.536	0.753	4.415
EnergEff	0	-0.113	1	-4.776	6.992	1.217	9.368
ComImp	0	-0.025	1	-3.006	3.576	0.145	3.346
Agri	0	-0.164	1	-3.816	6.235	1.845	9.658
ManTransSaf	0	-0.154	1	-3.092	5.746	0.924	5.938
ln (*EnergEff*+11)	2.394	2.388	0.089	1.828	2.890	0.292	8.505
ln (*Agri*+11)	2.394	2.383	0.086	1.972	2.847	1.175	7.729
ln ((*ManTransSaf*+5) *10+1)	3.913	3.901	0.193	3	4.686	-0.130	4.806

Note: *Number of observations = 1,11*

Table A. 3 Correlation table

Variables		1	2	3	4	5	6	7	8	9	10	11	12	13
impact area environment	1	1												
CommInd	2	-0.034	1											
GovernInd	3	-0.170 *	-0.021	1										
d2016	4	0.005	0.034	0.032	1 *									
d2017	5	0.035	0.068 *	-0.117 *	-0.065 *	1								
d2018	6	0.009	0.158 *	-0.105 *	-0.088 *	-0.258 *	1							
d2020	7	0.002	-0.106 *	0.146 *	-0.074 *	-0.216 *	-0.294 *	1						
d2021	8	-0.016	0.012	0.102 *	-0.036	-0.104 *	-0.142 *	-0.119 *	1					
d2015&2019	9	-0.032	-0.129 *	0.006	-0.099 *	-0.289 *	-0.394 *	-0.330 *	-0.159 *	1				
dmicro	10	-0.135 *	0.309 *	-0.015	0.000	0.005	-0.009	-0.003	0.020	-0.002	1			
dsmall	11	0.072 *	-0.178 *	0.008	0.014	0.028	-0.008	-0.024	-0.025	0.014	-0.647 *	1		
dmedium	12	0.076 *	-0.127 *	0.003	-0.013	-0.026	0.036	0.002	-0.016	-0.003	-0.397 *	-0.305 *	1	
dlarge	13	0.025	-0.106 *	0.011	-0.008	-0.026	-0.022	0.054	0.032	-0.018	-0.218 *	-0.167 *	-0.103 *	1
dAgricul	14	0.225 *	-0.007	-0.078 *	0.013	-0.027	0.040	-0.041	-0.040	0.034	-0.042	0.071 *	-0.042	0.010
dBuild	15	0.192 *	-0.037	-0.047	0.005	-0.045	0.040	0.001	-0.004	-0.003	-0.032	0.059 *	-0.017	-0.024
dConsPdct	16	0.362 *	0.154 *	-0.120 *	-0.025	0.043	-0.017	0.014	0.036	-0.040	-0.075 *	-0.009	0.052	0.101 *
dEducationTr	17	-0.147 *	-0.083 *	0.057	0.043	0.031	-0.062 *	-0.002	0.001	0.022	-0.025	-0.010	0.046	0.002
dEnergyEnvir	18	0.348 *	-0.116 *	-0.090 *	-0.033	0.006	0.002	0.026	-0.054	0.007	-0.070 *	0.094 *	0.008	-0.053
dFinLegservic	19	-0.292 *	-0.111 *	0.095 *	0.017	-0.012	0.025	0.006	-0.020	-0.015	0.054	-0.018	-0.047	-0.007
dHealthHuman	20	-0.133 *	-0.003	0.030	-0.029	0.006	0.004	-0.038	0.038	0.015	-0.017	-0.047	0.046	0.060 *
dMedRestHosp	21	-0.051	0.042	0.049	0.046	0.008	0.026	0.003	-0.019	-0.039	0.015	-0.047	0.052	-0.019
dRetTransLog	22	0.058	0.057	-0.023	-0.025	-0.014	-0.013	-0.029	0.032	0.042	0.053	-0.050	-0.013	0.009
dBusinessPro	23	-0.339 *	-0.002	0.088 *	0.008	-0.021	-0.014	0.009	0.000	0.020	0.083 *	-0.012	-0.048	-0.081 *
dCanada	24	0.001	0.022	0.001	-0.020	-0.052	-0.026	0.032	0.089 *	0.001	0.027	0.007	-0.044	-0.004
FinMan	25	0.527 *	0.012	-0.112 *	-0.047	0.070 *	0.014	0.013	-0.016	-0.057	0.051	0.023	-0.033	-0.107 *
WatLand	26	-0.196 *	0.002	0.052	0.012	0.033	-0.056	0.001	0.038	0.004	0.108 *	-0.044	-0.086 *	-0.009
EnergEff	27	0.062 *	-0.104 *	-0.004	0.018	-0.024	-0.037	0.041	-0.014	0.020	0.012	0.037	-0.059	-0.010
ComImp	28	-0.057	0.031	-0.044	-0.028	-0.035	0.069 *	0.005	0.011	-0.039	-0.031	-0.004	0.041	0.012
Agri	29	0.156 *	0.111 *	-0.079 *	0.009	-0.028	-0.004	-0.028	0.021	0.037	0.025	0.012	-0.063 *	0.020
ManTransSaf	30	0.001	0.048	0.031	-0.075 *	-0.011	-0.075 *	0.072 *	0.057	0.015	-0.066 *	0.029	-0.005	0.095 *

Notes: *p≤0.05.

EnergyEff represents the transformation ln (EnergyEff+11)

Agri represents the transformation ln (Agri+11)

*ManTransSaf represents the transformation ln((ManTransSaf+5) *10+1)*

CommInd represents the transformation ln (CommInd+1)

Table A. 4 Correlation table (continued)

Variables		**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
dAgricul	**14**	1																
dBuild	**15**	-0.032	1															
dConsPdct	**16**	-0.105*	-0.121*	1														
dEducationTr	**17**	-0.031	-0.035	-0.115*	1													
dEnergyEnvir	**18**	-0.037	-0.043	-0.141*	-0.041	1												
dFinLegservic	**19**	-0.064*	-0.074*	-0.241*	-0.0701*	-0.086*	1											
dHealthHuman	**20**	-0.033	-0.038	-0.123*	-0.0359	-0.044	-0.075*	1										
dMedRestHosp	**21**	-0.030	-0.034	-0.112*	-0.0325	-0.040	-0.0681*	-0.035	1									
dRetTransLog	**22**	-0.028	-0.033	-0.107*	-0.031	-0.038	-0.0649*	-0.033	-0.030	1								
dBusinessPro	**23**	-0.122*	-0.142*	-0.461*	-0.1345*	-0.165*	-0.2813*	-0.144*	-0.131*	-0.125*	1							
dCanada	**24**	-0.062*	-0.026	-0.001	-0.0043	-0.026	-0.0421	0.010	-0.013	0.008	0.073*	1						
FinMan	**25**	0.051	0.016	0.233*	-0.0974*	0.295*	-0.1322*	-0.147*	0.011	0.055	-0.211*	0.042	1					
WatLand	**26**	-0.103*	0.062*	-0.210*	-0.0218	-0.178*	0.036	-0.048	0.025	-0.041	0.295*	0.079*	0	1				
EnergEff	**27**	-0.086*	0.194*	-0.208*	-0.0469	0.335*	0.0733*	-0.066*	-0.105*	-0.050	0.047	-0.014	-0.048	0.000	1			
ComImp	**28**	-0.035	0.079*	-0.116*	0.1672*	0.097*	0.0615*	0.098*	0.021	-0.023	-0.097*	-0.004	0.000	0.000	-0.003	1		
Agri	**29**	0.386*	0.007	0.144*	-0.075*	-0.188*	0.0008	-0.048	-0.103*	-0.002	-0.101*	-0.024	-0.038	0.010	0.035	0.002	1	
ManTransSaf	**30**	-0.062*	-0.073*	0.246*	-0.0743*	-0.171*	0.0094	-0.011	-0.091*	0.043	-0.063*	0.019	-0.051	0.003	0.026	0.008	-0.007	1

Notes: *p≤0.05.

EnergyEff represents the transformation ln (EnergyEff+11)

Agri represents the transformation ln (Agri+11)

*ManTransSaf represents the transformation ln((ManTransSaf+5) *10+1)*

CommInd represents the transformation ln (CommInd+1)

Table A. 5 Base regression and robustness check

Variables	Complete		Vce (Robust)		Tobit	
CommInd	-1.613	**	-1.613	**	-1.579	**
	(0.648)		(0.737)		(0.737)	
GovernInd	-0.050		-0.050		-0.050	
	(0.068)		(0.065)		(0.065)	
dsmall	0.885		0.885		0.936	
	(0.595)		(0.583)		(0.583)	
dmedium	3.276	***	3.276	***	3.353	***
	(0.756)		(0.760)		(0.755)	
dlarge	3.088	***	3.088	***	3.168	***
	(1.182)		(1.140)		(1.129)	
dCanada	0.794		0.794		0.770	
	(0.656)		(0.663)		(0.660)	
dAgricul	7.407	***	7.407	***	7.413	***
	(1.702)		(2.589)		(2.561)	
dBuild	7.430	***	7.430	***	7.474	***
	(1.495)		(2.097)		(2.087)	
dEducationTr	-12.385	***	-12.385	***	-12.635	***
	(1.561)		(1.171)		(1.227)	
dEnergyEnvir	9.457	***	9.457	***	9.430	***
	(1.504)		(1.818)		(1.800)	
dFinLegservic	-13.002	***	-13.002	***	-12.984	***
	(0.931)		(0.772)		(0.765)	
dHealthHuman	-10.176	***	-10.176	***	-10.384	***
	(1.438)		(1.177)		(1.210)	
dMedRestHosp	-7.773	***	-7.773	***	-7.767	***
	(1.545)		(1.895)		(1.873)	
dRetTransLog	-1.736		-1.736		-1.724	
	(1.552)		(1.565)		(1.548)	
dBusinessPro	-9.056	***	-9.056	***	-9.076	***
	(0.747)		(0.802)		(0.793)	
FinMan	4.144	***	4.144	***	4.159	***
	(0.287)		(0.392)		(0.389)	
WatLand	-0.664	**	-0.664	**	-0.657	**
	(0.271)		(0.324)		(0.320)	
EnergEff	2.515		2.515		2.645	
	(3.245)		(4.484)		(4.464)	
ComImp	-0.588	**	-0.588	**	-0.593	**
	(0.260)		(0.283)		(0.281)	
Agri	15.548	***	15.548	***	15.537	***
	(3.363)		(3.862)		(3.818)	
ManTransSaf	1.249		1.249		1.231	
	(1.384)		(1.603)		(1.587)	
DummyYears	yes		yes		yes	
Constant	-21.839	*	-21.839		-22.202	
	(11.614)		(14.758)		(14.633)	
Nb obs.	1,110		1,110		1,110	
F	59.083	***	61.566	***	62.570	***
R^2	0.587		0.587			
Adjusted R^2	0.577		0.577			
Pseudo R^2					0.112	
Kurtosis						
Durbin-Watson:						
d_l						

d_u	
4-du	
4-dl	
Breusch-Pagan	
Limits (left-censored)	0 (N= 7)
(right-censored)	66.1 (N=1)

Notes: ***p≤0.001, **p≤0.05, *p≤0.1. The Breusch-Pagan test is a χ^2 with 1 degree of freedom. “dl” and “du” are the lower and upper critical values of the Durbin-Watson test. Since ‘2.004’ falls between “du” and “4-du” there is no autocorrelation.

dMicro, very small firms, is the omitted firm size category, *dConsPdct*, Consumer products, is the omitted industry category.

DummyYears refers to the control variables for the assessment years. Compared to the omitted variables *d2015&d2019* only *d2016* is significant for all the regressions but Basic Reg

EnergyEff represents the transformation *ln (EnergyEff+11)*

Agri represents the transformation *ln (Agri+11)*

ManTransSaf represents the transformation *ln ((ManTransSaf+5) *10+1)*

CommInd represents the transformation *ln (CommInd+1)*

APPENDIX B SUPPLEMENT MATERIAL ARTICLE 2

Table B 1 Description variables.

Variables	Description
Dependent variable	
dFunded	Dummy equal to 1 if the start-up received funding between 2020 and 2024, and 0 otherwise
Independent variables	
Web based independent variables	
WebFounderExpAcademic	Dummy = 1 for web signal related to founder academic experience, and 0 otherwise
WebFounderExpIndustry	Dummy = 1 for web signal related to founder industry experience, and 0 otherwise
WebFounderExpBoth	Dummy = 1 for web signal related to founder industry & academic experience, and 0 otherwise
No WebFounderExp	Dummy =1 if website has no signal related to the founder experience, 0 otherwise
WebFundingFPTGov	Dummy = 1 for a web signal related to federal, provincial or territories (FPT) government funding, and 0 otherwise
WebFundingNonFPT	Dummy = 1 for a web signal related to the non-FPT funding, and 0 otherwise
WebFundingBoth	Dummy = 1 for a web signal related to both types of funding, and 0 otherwise
No WebFunding	Dummy =1 if website has no signal related to the funding received, and 0 otherwise
No WebCollab	Dummy =1 if website has no signal related to the collaboration, and 0 otherwise
Low WebCollab	Dummy =1 if website has no signal related to the collaboration intensity is equal to 1, 2, 3, and 0 otherwise
Medium WebCollab	Dummy =1 if website has no signal related to the collaboration intensity is equal to 4, 5, 6, 7, and 0 otherwise
High WebCollab	Dummy =1 if website has no signal related to the collaboration intensity is equal to 8, 9, 10, and 0 otherwise
Control variables	
[11] Agriculture, forestry, fishing & hunting	Dummy = 1 if the NAICS code belongs to 11, and 0 otherwise
[21] Mining, quarrying, & oil & gas extraction, [22] Utilities, [23] Construction	Dummy = 1 if the NAICS code belongs to 21 or 22 or 23, and 0 otherwise
[41] Wholesale trade, [44-45] Retail trade, [48-49] Transp. & warehousing	Dummy = 1 if the NAICS code belongs to 41 or 44-45 or 48-49, and 0 otherwise
[51] Information and cultural industries	Dummy = 1 if the NAICS code belongs to 51, and 0 otherwise
[52] Finance & insurance, [53] Real estate & rental & leasing	Dummy = 1 if the NAICS code belongs to 52 or 53, and 0 otherwise
[54] Professional, S&T services, [55] Mngt of firms	Dummy = 1 if the NAICS code belongs to 54 or 55, and 0 otherwise
[56] Admin. and support, waste management & remediation serv.	Dummy = 1 if the NAICS code belongs to 56, and 0 otherwise
[61] Educational services, [62] Health care and social assistance	Dummy = 1 if the NAICS code belongs to 61 or 62, and 0 otherwise

Table B 1 Description variables. (continued)

Variables	Description
[71] Arts, entertainment & recreation, [72] Accom. & food serv.	Dummy = 1 if the NAICS code belongs to 71 or 72, and 0 otherwise
[81] Other services (except public admin.), [92] Public admin.	Dummy = 1 if the NAICS code belongs to 81 or 92, and 0 otherwise
dNot High Confidence	Dummy = 1 if CrunchBase is moderate (very few) or low confidence, and 0 otherwise
dHigh Confidence	Dummy = 1 if CrunchBase is highly confident about the ‘growing status’, and 0 otherwise
dAtlantic Province	Dummy variable taking the value 1 if the firm has its headquarter in the Atlantic provinces, and 0 otherwise
dQuebec	Dummy variable taking the value 1 if the firm has its headquarter in Quebec, and 0 otherwise
dOntario	Dummy variable taking the value 1 if the firm has its headquarter in Ontario, and 0 otherwise
dPrairies	Dummy variable taking the value 1 if the firm has its headquarter in the Prairies regions, and 0 otherwise
dBritish Columbia	Dummy variable taking the value 1 if the firm has its headquarter in British Columbia, and 0 otherwise
dNo-Low Growth	Dummy = 1 if not growing or no information, and 0 otherwise
dMedium-High Growth	Dummy = 1 if the start-up is growing (CrunchBase valuation), and 0 otherwise
dExtremelySmall	Dummy variable taking the value 1 if the firm has 1 to 10 employees, and 0 otherwise.
dVerySmall	Dummy variable taking the value 1 if the firm has 11 to 50 employees, and 0 otherwise.
dSmall	Dummy variable taking the value 1 if the firm has 51 to 100 employees, and 0 otherwise.
dMedium	Dummy variable taking the value 1 if the firm has 101 to 250 employees, and 0 otherwise.
dLarge	Dummy variable taking the value 1 if the firm has more than 250 employees, and 0 otherwise.

Table B 2 Descriptive statistics dependent variables

Variables	Mean	Nb. obs.
[11] Agriculture, forestry, fishing & hunting	0.012	69
[21] Mining, quarrying, & oil & gas extraction, [22] Utilities, [23] Construction	0.062	351
[41] Wholesale trade, [44-45] Retail trade, [48-49] Transp. & warehousing	0.055	315
[51] Information and cultural industries	0.080	453
[52] Finance & insurance, [53] Real estate & rental & leasing	0.214	1,220
[54] Professional, S&T services, [55] Mngt of firms	0.113	646
[56] Admin. and support, waste management & remediation serv.	0.184	1,050
[61] Educational services, [62] Health care and social assistance	0.025	144
[71] Arts, entertainment & recreation, [72] Accom. & food serv.	0.125	712
[81] Other services (except public admin.), [92] Public admin.	0.048	276
dFunded	0.309	1761
dHConf	0.024	134
dEastern Canada	0.037	208
dQuebec	0.135	768
dOntario	0.486	2,771
dMan_Sask_Alb	0.143	816
dBrit_Col	0.199	1,133
dGrowth	0.068	389
dExtremelySmall	0.597	3,398
dVerySmall	0.305	1,738
dSmall	0.040	228
dSmall-Medium	0.017	98
dMedium-Large	0.012	70
WebFounderExpAcademic	0.009	56

Table B 2 Descriptive statistics dependent variables (continued)

WebFounderExpIndustry	0.049	282
WebFounderExpBoth	0.033	192
No WebFounderExp	0.907	5166
WebFundingFPTGov	0.008	44
WebFundingNonFPT	0.012	66
WebFundingBoth	0.002	13
No WebFunding	0.978	5573
No WebCollab	0.798	4544
Low WebCollab	0.079	448
Medium WebCollab	0.074	420
High WebCollab	0.05	284

Note: Number of observations = 5,696.

Table B 3 LLM’s Prompt to create collaboration score.

Goal	Prompt
Collaboration	Please evaluate the company’s website to determine if the website explicitly mentions any partnership or collaboration. Structure your response as a JSON object with two keys: ‘Response’ and ‘Explanation’. - In the ‘Response’, indicate one of the following: ‘Yes’ or ‘No’. - In the ‘Explanation’, provide precise and concise evidence from the website to support your answer. If there is no relevant information, state ‘No information available’. **Important**: Provide **only** the JSON object in the following format and **do not include any additional text**:
Score collaboration	“You are an evaluator of the answer provided on a scale from 1 to 10 according to the following criteria: # Score: 9-10 (Strong Alignment) # * The response explicitly mentions terms like: # * “Partner”, “Collaborate”, “Work together”, “Joint effort”. # * Examples: “Collaboration is at the heart of our mission.” # Score: 7-8 (Moderate Alignment) # * The response indirectly refers to collaboration or partnership through terms like: # * “Work with”, “Cooperate”, “Team up”. # * Examples: “Our approach involves cooperation with stakeholders.” # Score: 5-6 (Weak Alignment) # * The response implies a collaborative intent but lacks direct or strong language. Examples include: # * “Support”, “Engage”, “Join forces” # * Examples: “We engage with industry leaders to achieve our goals,” “Our work is supported by various partnerships” # Score: 3-4 (Minimal Alignment) # * The response contains generic language that could be interpreted as collaboration-related but lacks explicit references or intent. # * Examples: “We work with others to achieve success,” “Our success relies on external contributions.” # Score: 1-2 (No or Very Weak Alignment) # * The response does not reflect any clear aspect of collaboration/partnership. # * Examples: Responses with irrelevant information, vague statements, or no meaningful content related to collaboration. “The output should be formatted as a JSON object that conforms to the JSON schema below:\n” “- ‘score’: an integer from 1 to 3.\n” “- ‘explanation’: a string explaining the criteria that led to the score.\n\n” “\**Important**: Provide **only** the JSON object in the following format without any additional text:\n\n” “```json\n” “{\n” ‘ “score”: <integer>,\n’ ‘ “explanation”: “<string>“\n’ “}\n”

Table B 4 LLM’s Prompt to create fundings and founder experience.

Goal	Prompt
Fundings	Please carefully review the company’s website to identify any explicit mentions of the company receiving a grant, subvention, or financial support from external sources. Structure your findings strictly as a JSON object with three keys: ‘Response’, ‘Explanation’, and ‘Source’. - Under ‘Response’, explicitly state either ‘Yes’ or ‘No’. - Under ‘Explanation’, provide a precise quotation or a concise summary from the website as evidence supporting your response. If no relevant information is found, explicitly state: ‘No information available’. - Under ‘Source’, specify the entity providing the funding by choosing exactly one of these categories: ‘Government’, ‘Private organization’, ‘Both’, or ‘Not specified’. If no information about the funding source is provided, select ‘Not specified’. **Important Instructions:** - Return **only** the JSON object. - Do not include any additional commentary, context, or explanations outside the JSON structure. - Adhere strictly to the following format:
Founder Experience	Please evaluate the company’s website to determine if the website mentions any information about the founder work experience or academic background. Structure your response as a JSON object with three keys: ‘Response’, ‘Explanation’, ‘Type’. “Response”: Provide “Yes” if the website mentions such information, or “No” if it does not. “Explanation”: Clearly and concisely reference the specific evidence from the website supporting your answer. If no relevant information is found, state “No information available.” “Type”: Categorize the founder’s experience explicitly mentioned as one of the following: - “Academic” if only academic background is mentioned. - “Industry” if only industry - related experience is mentioned. - “Both” if both academic and industry experiences are mentioned. - “Not specified” if there is no relevant information.

Table B 5 LLM prompt template for topic interpretation, showing example for Topic 2.

Goal	Prompt

Topic interpretation	You are given the following text, derived from a BERTopic analysis. I want you to analyze the text and provide the answer required below, considering that you need to name the Topic. The text provided represents documents that have a probability greater than 0.8 of belonging to Topic 2. Structure your response as a JSON object with one key:
	“Title”: Provide a title for the topic in a maximum of 5 words.
	Here is the text to analyze:
	{t_2}
	Here are the keywords returned by BERTopic:
	topic2 = [
	seo: 0.017649626202491484,
	website: 0.014094006015671288,
	marketing: 0.012971587224502873,
	….…………………………………….
	search engine: 0.005019141036465765’]
	Important: Provide only the JSON object in the following format:
	```
	json
	{{
	“Title”: “<string of maximum 5 words>“,
	}}
	Further instruction:
	1. **Be aware that the chunks of text come from several websites of different companies**.
	2. Do not focus on a company but focus on the topic itself.
	““““
	```

Table B 6 Retrievers' prompts.

Topic	**Prompt**
Collaboration	Please evaluate the company's website to determine if the website explicitly mentions any partnership or collaboration.
Fundings	Please carefully review the company's website to identify any explicit mentions of the company receiving a grant, subvention, or financial support from external sources.
Founder experience	Please evaluate the company's website to determine if there is any information about the founder work experience or academic background.

APPENDIX C SUPPLEMENT MATERIAL CHAPTER 6

Table C 1 Prompts used to create web-indicators.

<table>
<tr><th>Topic</th><th>Prompts</th></tr>
<tr><td><b>New product service</b></td><td>Analyze the provided website content to determine if it explicitly describes the introduction new products or new services.
Instructions:
- In the “Response” field, answer strictly with either “Yes” or “No”.
- In the “Explanation” field, briefly cite specific evidence directly from the website to support your response, or explicitly state “No information available” if such descriptions are absent.
- In the “Type” field, clearly indicate whether the cited content pertains to a “Product”, a “Service”, “Both”, or “None”.
Respond exclusively using the following JSON format:
{
“Response”: “<Yes|No>“,
“Explanation”: “<Cite specific evidence or state ‘No information available’>“,
“Type”: “<Product|Service|Both|None>“
}
Guidelines:
- Include tangible goods and intellectual or knowledge-based products over which ownership can be established and transferred through market transactions, or products that have significantly changed in design.
- Exclude the simple resale of new goods purchased from other businesses and purely aesthetic modifications.
- Clearly differentiate whether the new products or services are introduced by the company itself or merely represent services provided to other businesses.</td></tr>
</table>

Table C 1 Prompts used to create web-indicators.

<table>
<tr><td><b>Environmental innovation</b></td><td>Analyze the provided website content to determine whether it explicitly introduces or uses products for environmental goals.
Instructions:
- In the “Response” field, answer strictly with either “Yes” or “No”.
- In the “Explanation” field, briefly cite specific evidence from the website to support your response, or state “No information available” if no relevant descriptions are found.
- In the “Type” field, indicate the nature of the environmental product or technology as follows:
   - “Introduce”: if the company itself produces a product or technology that improves environmental performance.
   - “Used”: if the company uses a technology or product to improve environmental performance.
   - “Both”: if both production and usage are described.
   - “Not specified”: if there is no clear indication regarding environmental products or performance.
Respond exclusively using the following JSON format:
{
  “Response”: “<Yes|No>“,
  “Explanation”: “<Cite specific evidence or state ‘No information available’>“,
  “Type”: “<Introduce|Used|Both|Not specified>“
}
Guidelines:
- Include tangible goods and intellectual or knowledge-based products that can be owned, transferred, or that significantly influence market transactions.
- Exclude simple resales of goods purchased from other businesses and purely aesthetic modifications.
- Clearly differentiate whether the products or services are introduced by the company itself or merely represent services provided to other businesses.</td></tr>
</table>

Table C 1 Prompts used to create web-indicators.

Funding type	You are tasked with evaluating a company’s website to determine whether it explicitly states the use of any of the following government programs:
	1. Grants
	2. Loans
	3. Tax incentive programs
	4. Government-funded training or support
	5. Subsidies
	Carefully analyze the website content and respond strictly in the following JSON format, without any additional text:
	{
	“Response”: “<Yes or No>“,
	“Explanation”: “<Provide specific and concise evidence from the website supporting your response. If no relevant information is found, state clearly ‘No information available.’>“,
	“Type”: “<Type of government program from the list above>“
	}
	Instructions:
	- In ‘Response’, clearly indicate “Yes” if the website explicitly mentions using any of the above-listed government programs, otherwise “No”.
	- In ‘Type’, specify the exact program from the provided list that was explicitly mentioned. If none are mentioned, state “None”.
	- In ‘Explanation’, reference explicit statements from the website as evidence or clearly indicate “No information available” if applicable.
	Return your answer in the provided JSON format and do not include additional text.”““

Table C 1 Prompts used to create web-indicators.

Cooperation	You are an expert analyst tasked with reviewing a company’s website content to determine if it explicitly mentions any partnerships or collaborations.
	Additionally, if explicit on the website mention the type of partner among the following:
	- Parent, affiliated or subsidiary businesses
	- Suppliers of equipment, materials, components or software
	- Clients or customers from the private sector
	- Clients or customers from the public sector
	- Competitors or other businesses in the sector
	- Consultants and commercial laboratories
	- Universities, colleges or other higher education institutions
	- Government, public or private research institutes
	- Non-profit organizations
	- Households or individuals
	- Other co-operation partners
	- No specified
	Carefully examine the provided website text and **respond strictly as the JSON format below**, with no additional text or explanations outside this format:
	{
	“Response”: “<Yes or No>“,
	“Explanation”: “<Concise, specific evidence from the website supporting your response. If no relevant information is found, state clearly ‘No information available.’>“,
	“Type”: “<Write only one of the type of companies propose above>“
	}
	Guidelines:
	- Only respond “Yes” if the website explicitly uses terms like “partnership”, “collaboration”, “partner”, “collaborate”, or clearly describes working jointly with another organization.
	- Do not infer or assume partnerships from vague statements; explicit mentions only.
	- Only write in “Type” one of the possibilities listed above.
	- Your explanation must reference specific wording or sections from the provided text.”””