International Journal of Advance Scientific Research
(ISSN – 2750-1396)
VOLUME 06 ISSUE 02 Pages: 175-183
OCLC – 1368736135

ISSN-2750-1396

Journal Website: http://sciencebring.com/index.php/ijasr

Copyright: Original content from this work may be used under the terms of the creative commons attributes 4.0 licence.

Research Article

Artificial Intelligence, Web-Derived Signals, and the Measurement of Firm-Level Innovation: A Multidisciplinary Analytical Framework

Submission Date: January 03, 2026, **Accepted Date:** February 05, 2026,
Published Date: February 28, 2026

Aarav Mehta
Department of Economics and Innovation Studies, University of Amsterdam, Netherlands

Abstract

The measurement and prediction of firm-level innovation have long posed significant challenges for researchers and policymakers due to the latent, multidimensional, and context-dependent nature of innovation processes. Recent advancements in artificial intelligence (AI), big data analytics, and web mining methodologies have opened new avenues for capturing real-time, scalable, and nuanced indicators of innovation. This study develops a comprehensive, multidisciplinary framework that integrates insights from innovation economics, information retrieval theory, financial economics, and machine learning to examine how corporate websites, textual data, and AI-driven analytical tools can be utilized to measure and predict firm-level innovation. Drawing upon a carefully curated set of academic references, the article synthesizes theoretical and empirical contributions related to web-based innovation indicators, diffusion theory, transformer-based language models, and financial disclosure analytics.

The methodology relies on a conceptual integration of web scraping techniques, natural language processing models such as transformer architectures, and retrieval-augmented generation approaches to extract innovation signals from corporate digital footprints. The results demonstrate that website characteristics-including linguistic complexity, technological signaling, and product-related disclosures-serve as strong proxies for innovation activity, particularly when augmented with AI-driven analysis. Furthermore, the study highlights the role of AI in enhancing both the production and measurement of innovation, showing that firms leveraging AI technologies exhibit higher growth rates and increased product innovation.

The discussion critically evaluates the limitations of web-derived indicators, including issues related to data bias, interpretability, and cross-sector comparability, while proposing future research directions involving multimodal data integration and dynamic innovation tracking. The findings contribute to the broader literature by offering a unified framework that bridges theoretical and methodological gaps, providing actionable insights for academics, investors, and policymakers seeking to understand innovation dynamics in the digital age.

KEYWORDS

Artificial intelligence, innovation measurement, web mining, corporate websites, natural language processing, firm growth, big data analytics.

INTRODUCTION

Innovation has long been recognized as a central driver of economic growth, competitive advantage, and technological progress. Classical and contemporary theories of innovation emphasize its multifaceted nature, encompassing both radical and incremental changes in products, processes, and organizational structures. Early empirical work demonstrated that firms adopt innovations at varying rates depending on organizational characteristics, market conditions, and technological opportunities (Dewar and Dutton, 1986). Complementing this perspective, diffusion theory highlights the importance of inter-firm and intra-firm dynamics in shaping the spread of new technologies (Battisti and Stoneman, 2003). Despite these foundational insights, the empirical measurement of innovation remains inherently complex due to its intangible and often unobservable nature.

Traditional indicators of innovation, such as patent counts, research and development (R&D) expenditures, and survey-based measures, suffer from several limitations. Patents capture only a subset of innovative activity, often excluding process innovations and non-patented knowledge. R&D expenditures, while informative, do not necessarily translate into successful innovation outcomes. Surveys, on the other hand, are subject to reporting biases and lack real-time responsiveness. Consequently, there is a growing need for alternative data sources and methodologies capable of capturing innovation more comprehensively and dynamically.

In recent years, the proliferation of digital technologies has transformed the informational landscape, providing unprecedented access to firm-level data through online platforms. Corporate websites, in particular, have emerged as rich repositories of information, reflecting firms' strategic priorities, technological capabilities, and market positioning. Research has demonstrated that specific website characteristics, such as the presence of product descriptions, technological terminology, and innovation-related language, can serve as proxies for innovation activity (Axenbeck and Breithaupt, 2021). Building on this insight, web mining techniques have been increasingly employed to study innovation patterns, offering scalable and cost-effective alternatives to traditional methods (Gök et al., 2015).

The integration of artificial intelligence into this domain further enhances the analytical potential of web-derived data. Transformer-based language models, such as those introduced in natural language processing research, enable the extraction of semantic meaning from large volumes of textual data with remarkable accuracy (Devlin et al., 2019). These models facilitate the identification of nuanced linguistic patterns associated with innovation, such as references to emerging technologies, collaborative initiatives, and product development strategies. Moreover, recent advancements in retrieval-augmented generation have improved the contextual understanding of language models by combining generative capabilities with external knowledge retrieval (Gao et al., 2024).

Parallel developments in financial economics have underscored the importance of information disclosure in shaping market perceptions and investment decisions. Studies on financial reporting indicate that textual disclosures, including earnings calls and regulatory filings, contain valuable signals about firm performance and strategic direction (Stice, 1991; Tailor and Kale, 2025). These insights align with signaling theory, which posits that firms communicate information to reduce asymmetries and influence stakeholder behavior (Spence, 1973). In the context of innovation, corporate websites can be viewed as signaling mechanisms through which firms convey their innovative capabilities to external audiences.

Despite these advances, significant gaps remain in the literature. Existing studies often focus on isolated aspects of innovation measurement, such as web indicators or AI applications, without integrating them into a cohesive analytical framework. Additionally, there is limited understanding of how AI-driven methodologies can enhance both the measurement and generation of innovation simultaneously. This study addresses these gaps by developing a comprehensive framework that synthesizes insights from multiple disciplines to analyze the role of web-derived data and AI in measuring firm-level innovation.

METHODOLOGY

The methodological approach adopted in this study is conceptual and integrative, combining theoretical insights and empirical findings from the provided references to construct a unified framework for measuring innovation using web-derived data and artificial intelligence. The methodology is structured around three core components: data acquisition through web mining, semantic analysis באמצעות natural language processing, and interpretive modeling using AI-driven techniques.

The first component involves the systematic collection of data from corporate websites. Web scraping techniques enable the extraction of textual and structural information, including product descriptions, technological references, organizational narratives, and visual elements. The use of big data methodologies allows researchers to process large volumes of such data efficiently, facilitating cross-sectional and longitudinal analyses (Blazquez and Domenech, 2018). Importantly, the selection of firms and industries must adhere to standardized classification systems, such as those defined by statistical frameworks, to ensure comparability across sectors.

The second component focuses on the transformation of raw textual data into meaningful indicators of innovation. Natural language processing plays a critical role in this process, enabling the identification of relevant keywords, phrases, and semantic patterns. Early contributions to information retrieval theory highlight the importance of term specificity in distinguishing meaningful content from noise (Sparck Jones, 1972). Building on this foundation, modern transformer-based models provide advanced capabilities for contextual understanding, allowing for the detection of subtle linguistic cues associated with innovation activities.

In particular, transformer architectures leverage bidirectional attention mechanisms to capture relationships between words and phrases within a given context (Devlin et al., 2019). This enables the identification of complex constructs such as technological sophistication, collaborative innovation, and market differentiation. Furthermore, the integration of retrieval-augmented generation enhances the robustness of these models by incorporating external knowledge sources, thereby improving the accuracy and relevance of the extracted information (Gao et al., 2024).

The third component involves the interpretation and validation of the extracted indicators. This requires the integration of theoretical frameworks from innovation economics and financial analysis. For instance, the relationship between innovation and firm growth can be examined using insights from recent studies demonstrating that AI adoption is associated with increased product innovation and market expansion (Babina et al., 2024). Similarly, the evaluation of innovation-related signals can be informed by performance measurement frameworks that account for downside risks and uncertainty (Sortino and Price, 1994).

An important aspect of the methodology is the consideration of firm heterogeneity. Small and medium-sized enterprises, as defined by standardized criteria, often exhibit different innovation dynamics compared to larger firms. Recent empirical evidence suggests that web-derived indicators are particularly effective in capturing innovation among smaller firms, where traditional data sources may be limited (Bottai et al., 2024). This highlights the need for tailored analytical approaches that account for variations in firm size, industry, and geographic context.

Finally, the methodology emphasizes the importance of triangulation. By combining multiple data sources and analytical techniques, researchers can enhance the reliability and validity of their findings. This includes the integration of web-based indicators with financial disclosures, market data, and survey-based measures, providing a more comprehensive understanding of innovation dynamics.

Results

The synthesis of the referenced studies reveals several key findings regarding the measurement and prediction of firm-level innovation using web-derived data and artificial intelligence. First, corporate websites emerge as highly informative sources of innovation-related information. Empirical analyses demonstrate that specific website characteristics, such as the presence of detailed product descriptions, references to advanced technologies, and the use of innovation-related language, are strongly correlated with firm-level innovation activity (Axenbeck and Breithaupt, 2021). These findings suggest that firms actively communicate their innovative capabilities through their online presence, making websites valuable proxies for innovation measurement.

Second, the application of web mining techniques significantly enhances the scalability and timeliness of innovation analysis. Unlike traditional methods, which often rely on lagged data, web-based approaches enable real-time monitoring of innovation trends. This is particularly relevant in rapidly evolving industries, where timely information is critical for decision-making. The ability to analyze large datasets also facilitates the identification of patterns and trends that may not be apparent in smaller samples (Gök et al., 2015).

Third, the integration of artificial intelligence into the analytical process yields substantial improvements in accuracy and depth of analysis. Transformer-based language models are capable of capturing complex semantic relationships within textual data, enabling the identification of nuanced indicators of innovation. For example, these models can distinguish between superficial mentions of technology and substantive evidence of innovation, such as descriptions of proprietary processes or novel product features. This level of granularity is essential for accurately assessing innovation activity.

Fourth, the results highlight the dual role of artificial intelligence as both a tool for measuring innovation and a driver of innovation itself. Firms that adopt AI technologies tend to exhibit higher levels of product innovation and growth, suggesting a positive feedback loop between AI adoption and innovation performance (Babina et al., 2024). This finding underscores the importance of considering AI not only as an analytical tool but also as a strategic asset that influences firm behavior.

Fifth, the analysis reveals significant heterogeneity across firms and industries. Small and medium-sized enterprises, in particular, benefit from web-based indicators, as these provide insights into innovation activities that may not be captured by traditional metrics. Studies focusing on manufacturing SMEs demonstrate that web scraping can effectively identify innovative firms, even in the absence of formal R&D reporting (Bottai et al., 2024).

Finally, the results emphasize the importance of contextual factors in interpreting web-derived indicators. The effectiveness of these indicators depends on factors such as industry characteristics, market conditions, and regulatory environments. For instance, firms operating in highly regulated industries may be more cautious in disclosing innovation-related information, potentially limiting the usefulness of web-based measures.

Discussion

The findings of this study have significant implications for both theory and practice. From a theoretical perspective, the integration of web-derived data and artificial intelligence into innovation measurement represents a paradigm shift in how researchers conceptualize and operationalize innovation. Traditional approaches, which rely on static and often incomplete data sources, are increasingly being supplemented by dynamic, data-driven methodologies that capture the complexity and fluidity of innovation processes.

One of the key contributions of this study is the development of a unified framework that bridges multiple disciplines, including innovation economics, information retrieval, and machine learning. This interdisciplinary approach enables a more comprehensive understanding of innovation, addressing the limitations of single-method studies. By combining insights from different fields, the framework provides a holistic view of innovation that accounts for both quantitative and qualitative dimensions.

However, several challenges and limitations must be acknowledged. One of the primary concerns is the potential for bias in web-derived data. Not all firms maintain comprehensive or up-to-date websites, and the information presented may be influenced by strategic considerations. This raises questions about the reliability and representativeness of web-based indicators. Additionally, differences in language, cultural context, and industry norms can affect the interpretation of textual data, complicating cross-country and cross-sector comparisons.

Another limitation relates to the interpretability of AI-driven models. While transformer-based models offer powerful analytical capabilities, their complexity can make it difficult to understand how specific conclusions are derived. This lack of transparency poses challenges for researchers and practitioners seeking to validate and explain their findings. Addressing this issue requires the development of explainable AI techniques that enhance the interpretability of model outputs.

The ethical implications of web mining and AI-driven analysis also warrant careful consideration. The collection and use of online data raise concerns about privacy, data ownership, and consent. Ensuring that research practices adhere to ethical standards is essential for maintaining trust and legitimacy in the field.

Looking ahead, several avenues for future research emerge. One promising direction is the integration of multimodal data, including text, images, and audio, to capture a more comprehensive picture of innovation. Advances in multimodal machine learning, as demonstrated in recent studies on financial disclosures, suggest that combining different types of data can enhance predictive accuracy and provide deeper insights (Tailor and Kale, 2025).

Another important area for future research is the development of dynamic models that track innovation over time. By analyzing changes in website content and other digital signals, researchers can gain insights into the evolution of innovation strategies and the factors that drive success. This longitudinal perspective is particularly valuable for understanding the impact of external shocks, such as technological disruptions or economic crises.

Finally, there is a need to explore the policy implications of web-based innovation measurement. Policymakers can leverage these methodologies to monitor innovation activity, identify emerging trends, and design targeted interventions. This is especially relevant in the context of regional development, where disparities in innovation capacity can have significant economic consequences.

Conclusion

This study provides a comprehensive analysis of the role of artificial intelligence and web-derived data in measuring and predicting firm-level innovation. By synthesizing insights from a diverse set of academic references, the research develops a unified framework that integrates web mining, natural language processing, and AI-driven analysis. The findings demonstrate that corporate websites serve as valuable sources of innovation-related information, particularly when analyzed باستخدام advanced computational techniques.

The study highlights the transformative potential of artificial intelligence in both measuring and driving innovation, emphasizing the need for interdisciplinary approaches that combine theoretical and methodological perspectives. While challenges related to data quality, model interpretability, and ethical considerations remain, the opportunities offered by these emerging technologies are substantial.

In conclusion, the integration of AI and web-based methodologies represents a significant advancement in the field of innovation studies, providing new tools and insights for researchers, practitioners, and policymakers. As digital technologies continue to evolve, the ability to capture and analyze innovation in real time will become increasingly important, shaping the future of economic research and strategic decision-making.

References

1. Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites-Which website characteristics predict firm-level innovation activity? PLoS ONE, 16(4), e0249583. https://doi.org/10.1371/journal.pone.0249583
2. Babina, T., Fedyk, A., He, A., & Hodson, J. (2024). Artificial intelligence, firm growth, and product innovation. Journal of Financial Economics, 151, 103745. https://doi.org/10.1016/j.jfineco.2023.103745
3. Battisti, G., & Stoneman, P. (2003). Inter- and intra-firm effects in the diffusion of new process technology. Research Policy, 32(9), 1641–1655. https://doi.org/10.1016/S0048-7333(03)00055-6
4. Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. Technological Forecasting and Social Change, 130, 99–113. https://doi.org/10.1016/j.techfore.2017.07.027
5. 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. https://doi.org/10.1016/j.techfore.2024.123597
6. Bouschery, S.-G., Blazevic, V., & Piller, F.-T. (2023). Augmenting human innovation teams with artificial intelligence: Exploring transformer-based language models. Journal of Product Innovation Management. https://doi.org/10.1111/jpim.12656
7. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. https://doi.org/10.48550/arXiv.1810.04805
8. Dewar, R.-D., & Dutton, J.-E. (1986). The adoption of radical and incremental innovations: An empirical analysis. Management Science, 32(11), 1422–1433. https://doi.org/10.1287/mnsc.32.11.1422
9. European Commission. (2003). Commission recommendation of 6 May 2003 concerning the definition of micro, small and medium-sized enterprises. Official Journal of the European Union, L, 124, 36–41.
10. European Commission. (2008). NACE Rev. 2: Statistical classification of economic activities in the European Community. Publications Office.
11. European Commission. (2024). Regions in the European Union: Nomenclature of territorial units for statistics (NUTS). Publications Office. https://doi.org/10.2785/714519
12. Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., & Wang, H. (2024). Retrieval-augmented generation for large language models: A survey. https://doi.org/10.48550/arXiv.2312.10997

13. Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. Scientometrics, 102, 653–671. https://doi.org/10.1007/s11192-014-1434-0

14. Singh, M., & Biwas, A. (2023). AI stocks rally in latest Wall Street craze sparked by ChatGPT. Reuters.

15. Somefun, K., Perchet, R., Yin, C., & Leote de Carvalho, R. (2023). Allocating to thematic investments. Financial Analysts Journal, 79, 18–36.

16. Sortino, F. A., & Price, L. N. (1994). Performance measurement in a downside risk framework. Journal of Investing, 3, 59–64.

17. Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 28, 11–21.

18. Spence, M. (1973). Job market signaling. Quarterly Journal of Economics, 87, 355–374.

19. Stice, E. K. (1991). The market reaction to 10-K and 10-Q filings and to subsequent earnings announcements. Accounting Review, 66, 42–55.

20. Priyank Tailor, & Anjali Kale. (2025). Multimodal Sentiment Analysis of Earnings Calls and SEC Filings: A Deep Learning Approach to Financial Disclosures. Utilitas Mathematica, 122(1), 3163–3168. Retrieved from https://utilitasmathematica.com/index.php/Index/article/view/2676