Textual Alchemy: Predicting Company Innovation by Deciphering Unstructured Website Content in Time and Space

Korneliusz Pylak,
Lublin University of Technology, POLAND

Keywords
Innovation Dynamics
Unstructured Website Text
Temporal Analysis
Spatial Dynamics
Natural Language Processing (NLP)

JEL codes
O31 - Innovation and Invention: Processes and Incentives
O32 - Management of Technological Innovation and R&D
C55 - Large Data Sets: Modeling and Analysis
L86 - Information and Internet Services; Computer Software
R11 - Regional Economic Activity: Growth, Development, Environmental Issues, and Changes

Extended abstract

To estimate innovation and critical firm performance indicators, traditional methodologies have relied mainly on established secondary data sources, including patents, academic publications, R&D projects and administrative records (Abbasiharofteh et al., 2022; Cillo et al., 2019; Nasirov, 2020; Simensen & Abbasiharofteh, 2022). However, the landscape of innovation geography research has experienced a transformative shift, driven by advancements in computational power and language modelling, in tandem with the abundance of textual data available from diverse sources such as job postings, patent documents, web texts, and trademark data. This digital revolution has unlocked novel possibilities for exploring regional economic development, labour market dynamics, and the geographies of knowledge production and relationships (Aweisi et al., 2021; Cetera et al., 2022; Gök et al., 2015; Skhvediani et al., 2022).

The integration of textual data analytics has pioneered an approach to the study of the geography of innovation, including the exploration of digital footprints of corporate linkages, social media data and digitised historical newspaper archives (Abbasiharofteh et al., 2023; Ashouri et al., 2022; Daas & van der Doef, 2020; Gök et al., 2015; Kinne & Axenbeck, 2020). Currently, the use of unstructured textual data is gaining momentum, providing researchers with innovative avenues to comprehend and interpret the intricate connections within innovation geography. Nonetheless, while these methodologies exhibit promise in understanding innovative activity, our approach stands out by seamlessly bridging contextual, temporal, and spatial aspects of innovation. Venturing into this uncharted territory, our research makes a significant and pioneering contribution to the existing literature on two fronts, offering a holistic approach that anticipates and understands innovative activities within companies.

Primarily, our methodology extends the examination of unstructured website text beyond the static representations typically utilized by existing methods, employing advanced techniques such as web scraping, social network analysis, and natural language processing (Abbasiharofteh et al., 2023; Ashouri et al., 2022; Kinne & Axenbeck, 2020; Skhvediani et al., 2022). Instead of relying solely on company websites, we introduce a temporal dimension, foreseeing a company's innovations by monitoring changes in the textual content it publishes over time. While previous studies have explored text-based predictions of innovation in analyst reports (Bellstam et al., 2021), our innovative approach pioneers the use of unstructured website text as an early indicator of forthcoming innovations. This temporal perspective allows us to trace the evolution of innovative activities, offering a more comprehensive understanding of the innovation process within a single company.

Secondly, we acknowledge the close association between innovation capability and a company's ability to integrate existing knowledge and resources over time (Audretsch & Belitski, 2022; Bruno et al., 2022; Tomizawa et al., 2020). The spatial dimension is crucial, where physical proximity and inter-firm relationships play crucial roles in facilitating learning and catalyzing innovation (Alam et al., 2022; Bailey et al., 2018; Obschonka et al., 2023; Singh et al., 2022). To address this spatial aspect, our approach transcends isolated predictions by considering the colocation of innovative entities in proximity. By capturing the dynamic interaction between innovative firms in their spatial context, we offer a more refined

understanding of the geography of knowledge production and relationships, revealing how spatial dynamics shape innovation.

To meet our objectives of studying the dynamics of innovation in Polish companies and identifying early indicators preceding the official launch of innovations, we use an extensive methodology. We use the WebArchive snapshot database, which includes more than ten thousand business entities in Poland that filed patents between 2001 and 2023. Each company is accurately geolocalised, placed in a detailed socio-economic context, incorporating factors such as organisational structure, economic and technological diversity and local knowledge complexity. To comprehensively explore the causal links between measures of innovation over time, we adopt patent data extracted from patent databases and textual representations of innovations (new product launches). In the area of text mining, our methodology uses advanced topic modelling tools such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), Correlated Topics Models (CTM), and word embeddings like GloVe. In addition, we exploit the transformative capabilities of natural language processing (NLP), integrating cutting-edge Transformer models into our analysis. This multifaceted approach enables us to decode unstructured website text, anticipating and understanding the intricate connections within innovation geography, and offering a holistic view of innovative activities in Polish companies.

Bibliography

Abbasiharofteh, M., Castaldi, C., & Petralia, S. (2022). *From patents to trademarks: Towards a concordance map*. European Patent Office.

Abbasiharofteh, M., Krüger, M., Kinne, J., Lenz, D., & Resch, B. (2023). The digital layer: Alternative data for regional and innovation studies. *Spatial Economic Analysis*, 1–23.

Alam, M. A., Rooney, D., & Taylor, M. (2022). From ego-systems to open innovation ecosystems: A process model of inter-firm openness. *Journal of Product Innovation Management*, *39*(2), 177–201.

Ashouri, S., Suominen, A., Hajikhani, A., Pukelis, L., Schubert, T., Türkeli, S., Van Beers, C., & Cunningham, S. (2022). Indicators on firm level innovation activities from web scraped data. *Data in Brief*, *42*, 108246.

Audretsch, D. B., & Belitski, M. (2022). The knowledge spillover of innovation. *Industrial and Corporate Change*, *31*(6), 1329–1357.

Aweisi, A., Arora, D., Emby, R., Rehman, M., Tanev, G., & Tanev, S. (2021). Using web text analytics to categorize the business focus of innovative digital health companies. *Technology Innovation Management Review*, *11*(7/8).

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, *32*(3), 259–280.

Bellstam, G., Bhagat, S., & Cookson, J. A. (2021). A Text-Based Analysis of Corporate Innovation. *Management Science*, *67*(7), 4004–4031.

Bruno, R. L., Crescenzi, R., Estrin, S., & Petralia, S. (2022). Multinationals, innovation, and institutional context: IPR protection and distance effects. *Journal of International Business Studies*, *53*(9), 1945–1970.

Cetera, W., Gogołek, W., Żołnierski, A., & Jaruga, D. (2022). Potential for the use of large unstructured data resources by public innovation support institutions. *Journal of Big Data*, *9*(1), 46.

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.

Daas, P. J., & van der Doef, S. (2020). Detecting innovative companies via their website. *Statistical Journal of the IAOS*, *36*(4), 1239–1251.

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671.

Kinne, J., & Axenbeck, J. (2020). Web mining for innovation ecosystem mapping: A framework and a large-scale pilot study. *Scientometrics*, *125*(3), 2011–2041.

Nasirov, S. (2020). Trademark value indicators: Evidence from the trademark protection lifecycle in the U.S. pharmaceutical industry. *Research Policy*, *49*(4), 103929.

Obschonka, M., Tavassoli, S., Rentfrow, P. J., Potter, J., & Gosling, S. D. (2023). Innovation and inter-city knowledge spillovers: Social, geographical, and technological connectedness and psychological openness. *Research Policy*, *52*(8), 104849.

Simensen, E. O., & Abbasiharofteh, M. (2022). Sectoral patterns of collaborative tie formation: Investigating geographic, cognitive, and technological dimensions. *Industrial and Corporate Change*, *31*(5), 1223–1258.

Singh, A., Chhetri, P., & Padhye, R. (2022). Modelling inter-firm competitive rivalry in a port logistics cluster: A case study of Melbourne, Australia. *The International Journal of Logistics Management*, *33*(2), 455–476.

Skhvediani, A., Sosnovskikh, S., Rudskaia, I., & Kudryavtseva, T. (2022). Identification and comparative analysis of the skills structure of the data analyst profession in Russia. *Journal of Education for Business*, *97*(5), 295–304.

Tomizawa, A., Zhao, L., Bassellier, G., & Ahlstrom, D. (2020). Economic growth, innovation, institutions, and the Great Enrichment. *Asia Pacific Journal of Management*, *37*(1), 7–31.