CARMA 2025

7th Int. Conf. on Advanced Research Methods and Analytics
2-4 July 2025, Rome, Italy
DOI: https://doi.org/10.4995/CARMA2025.2025.20544

Detecting Potential Fake Innovators Using Web Scraping: The Case of Italian Startups

Ilaria Nigro, Agapito Emanuele Santangelo, Michele Modina
Department Of Economics, University of Molise, Italy

How to cite: Modina M., Nigro, I.; Santangelo, A. E. 2025. Detecting potential fake innovators using web scraping: the case for Italian startup. In: 7th International Conference on Advanced Research Methods and Analytics (CARMA 2025). Rome, 2-4 July 2025. https://doi.org/10.4995/CARMA2025.2025.20544

Abstract

Measuring innovation at the firm level is a complex challenge due to the lack of universally accepted metrics. Traditional measures such as patent counts and R&D expenditure often fail to capture the dynamic, intangible aspects of innovation, especially for new ventures. In this context, some companies may exploit innovation policy programs by claiming "innovative" status to gain regulatory benefits despite limited actual innovation. This paper proposes an alternative web data-driven approach to identify such potential "fake innovators." We leverage the structure and content of corporate websites, specifically the adoption of modern HTML5 web features, as a proxy for a firm's digital and technological sophistication. Using data on Italian small and medium-sized enterprises (SMEs), we show that this proxy helps distinguish truly innovative startups from those that appear to claim the status without substantive technological development. Our findings indicate that firms with higher digital sophistication (e.g., extensive use of HTML5 functionalities) are indeed more likely to reflect genuine innovative activity. In contrast, many officially designated innovative startups with low web sophistication exhibit characteristics suggesting they may not be as innovative as claimed. This approach can provide policymakers with a cost-effective tool to flag potentially misrepresented innovation status, complementing traditional innovation metrics.

Keywords: *innovation measurement; web scraping; policy evaluation.*

1. Introduction

Innovation is widely recognized as a key driver of economic growth, yet measuring a firm's innovativeness remains challenging. There is no single agreed-upon metric for innovation performance (Autio et al., 2014). Common indicators like the number of patents or R&D spending capture only certain facets of innovation and may overlook emerging or digital forms of innovation. This is particularly true for startups and small firms, which may innovate in ways not immediately reflected in formal R&D or patent data.

Policymakers have attempted to identify and support innovative companies through special designations and incentive programs. In Italy, for example, Law No. 221/2012 established the status of "innovative startup," offering tax reliefs and other benefits to new enterprises meeting certain criteria (Cian, 2015; Ricci, 2016). This policy intends to foster entrepreneurship and technological advancement among SMEs. However, because firms self-declare and only need to satisfy relatively broad requirements, there is a risk that some businesses obtain the innovative startup label primarily to access benefits, without truly engaging in innovative activities. We refer to these firms as "fake innovators" – companies that officially qualify as innovative startups but lack substantive innovation output or capabilities. Identifying these cases is important, as they can dilute the effectiveness of innovation policies and misallocate public support away from genuinely innovative ventures.

One promising avenue for detecting potential fake innovators is to analyze a firm's digital footprint. In today's business environment, a corporate website is not just a marketing tool but often a reflection of a company's technological adoption and innovation mindset. Firms on the cutting edge tend to maintain modern, feature-rich websites to showcase their products, services, or technological prowess.In this study, we apply a web scraping approach to systematically measure the presence of HTML5 features on Italian companies' websites and integrate this with firm-level data.

Our goal is to determine whether low digital sophistication aligns with firms that might be "innovative" in name only. We analyze a sample of Italian SMEs across various sectors (with a focus on manufacturing) in the year 2025, including both firms officially registered as innovative startups and those without such status. The analysis involves two main steps: first, an exploratory clustering of firms based on their digital and economic attributes to identify distinct profiles (and potentially isolate a group corresponding to low-tech "fake innovators"); second, a regression analysis to compare the determinants of innovative status among firms with differing levels of digital sophistication. Through this approach, we illustrate how web scraping and digital indicators can complement traditional metrics in assessing and verifying innovation claims.

2. Data and Methodology

2.1. Data and Sample

The study examines firm-level data for Italian SMEs active in 2025, combining official records with web-scraped information. We focus on small and medium-sized enterprises in the manufacturing sector and related industries, among which a subset are classified as innovative startups under the Italian law. The sample includes both companies that have obtained the innovative startup status and those that have not, allowing comparison between "innovative" and ordinary firms. For each firm, we obtained basic profile information such as annual sales revenue and number of employees (from financial statements or business registries), which serve as measures of firm size and performance.

2.2. Web Scraping and HTML5 Proxy

We identified the official website for each firm in the sample and performed web scraping to retrieve the homepage HTML code. From the HTML code, we measured the usage of HTML5-specific tags and features as an indicator of the website's technological sophistication. In particular, we looked for the presence of elements introduced or made prominent by the HTML5 standard (released in the early 2010s) – for example, semantic structure tags like <header>, <section>, <article>; multimedia and interactive content tags such as <video>, <audio>, <canvas>; and modern form or input elements. The total count of distinct HTML5 tags used on a site (or the frequency of their use) was recorded for each firm. A higher count of HTML5 elements suggests the firm has a more complex, modern website (likely reflecting greater digital capabilities), while a very low count (or none) suggests a basic, possibly outdated site. This HTML5 usage metric is admittedly a rough proxy for digital innovation, but it provides a quantitative gauge of how readily a firm adopts new web technologies.

2.3. Clustering Analysis

To explore patterns in the data without imposing strong assumptions, we first conducted an unsupervised cluster analysis on the firms. We used k-means clustering with the following standardized variables: the firm's HTML5 tag count (as defined above) to represent digital sophistication, the logarithm of the firm's sales revenue, and the logarithm of its number of employees (the latter two representing firm size and resources). Prior to clustering, all three variables were z-score standardized to give them equal weight despite different scales. We applied the "elbow method" on the within-cluster sum of squares to determine an appropriate number of clusters. The elbow plot indicated that a solution with three clusters was optimal, capturing distinct groupings in the data. After determining k = 3, we ran the k-means algorithm (with multiple random starts to ensure a stable solution) to assign each firm to one of three clusters. The characteristics of these clusters were then interpreted to understand the profiles of firms they represent:

- **Cluster 1** – firms with high HTML5 usage, regardless of size. This cluster consisted of companies that extensively use modern web features on their site, indicating technologically advanced firms. These firms ranged from small to larger SMEs, but shared a strong digital orientation.
- **Cluster 2** – firms with minimal HTML5 usage and relatively small size. This group was dominated by smaller businesses that had little to no presence of modern web elements, suggesting very limited digital sophistication.
- **Cluster 3** – firms of moderate size with intermediate HTML5 usage. Companies in this cluster had a middling level of web technology adoption and typically mid-range sales and employee numbers, indicating a moderate level of digital engagement.

This clustering helped reveal an interesting pattern: one cluster (Cluster 2) clearly captured the profile of small firms with poor digital presence. We suspected that many of the so-called fake innovators – innovative startups that are technologically lagging – would fall into this cluster. To verify this, we introduced a classification of firms based on their innovation status and web sophistication.

2.4. Defining "Fake Innovators"

We created an indicator to flag potential fake innovators among the innovative startups. Specifically, within the subset of firms officially recognized as innovative startups, we compared each firm's HTML5 tag count to the median HTML5 usage of that subset. Those innovative startups whose HTML5 usage was at or below the median were labeled as "potential fake innovators," reflecting relatively low digital sophistication for an allegedly innovative firm. Conversely, innovative startups with above-median HTML5 usage were labeled "true innovators," as their robust web presence aligns with expectations of genuine technological innovation.

2.5. Regression Analysis

In the second part of our methodology, we employed regression analysis to further investigate how the determinants of being an "innovative startup" might differ depending on a firm's digital sophistication. The dependent variable in our models is a binary indicator of whether the firm has the official innovative startup status (1 if yes, 0 if no). We estimated several probit regression models on this outcome. Rather than a single pooled model, we ran separate probit models on sub-samples of firms categorized by their HTML5 usage level. This approach allows us to see if certain factors are more or less important predictors of innovative status for firms with advanced websites versus those with basic websites.

We divided firms into four groups for this analysis:

1. No HTML5 usage – firms whose websites did not contain any of the identified HTML5 elements.
2. Low HTML5 usage – firms with some HTML5 elements but below a certain low threshold (e.g., below the median usage among all firms).
3. Medium HTML5 usage – firms with a moderate number of HTML5 elements (around the median level).
4. High HTML5 usage – firms with extensive use of HTML5 features (above the median or a high threshold).

For each group, we estimated a probit model of the form:

PP(startup_innovativa=1)=Φ(β0+β1·log_sales+β2·log_employees)P(startup_innovativa = 1) = \Phi(\beta_0 + \beta_1 \cdot log_sales + \beta_2 \cdot log_employees)

where Φ is the cumulative normal distribution. The key independent variables included are the logarithm of sales and the logarithm of employees, representing firm size/performance, along with a constant term. These variables were chosen because they capture fundamental differences in firm scale and are available for all firms; we wanted to see if the profile of firms obtaining innovative status (in terms of size) differs across digital sophistication levels. By comparing the coefficients β_1 and β_2 across the sub-sample models, we can infer how the relationship between firm characteristics and the likelihood of claiming innovative status changes with the firm's digital/technological sophistication.

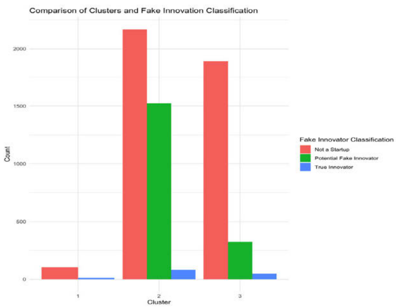

Figure 1. Comparison of clusters and fake innovation classification. Source: our elaboration (2025).

3. Results

3.1. Cluster Analysis Results

The k-means clustering yielded three clearly distinguishable clusters of firms as described above. Most notably, Cluster 2 (the small, low-HTML5 group) emerged as a concentration of firms that fit the profile of potential fake innovators. Indeed, when mapping the innovative startup firms onto these clusters, we observe that a large fraction of the startups with low digital sophistication fell into Cluster 2. In other words, many of the companies that had claimed the innovative startup title but had basic websites were grouped together by the algorithm, separate from those with more advanced digital profiles.

This finding is illustrated by the distribution of innovation categories across clusters. For example, among the innovative startups in our sample, those labeled as potential fake innovators were predominantly found in Cluster 2. Meanwhile, innovative startups deemed true innovators were mostly in Cluster 1 or 3, which have higher HTML5 usage. The Chi-square test mentioned earlier confirmed that cluster membership and our fake innovator classification are significantly associated ($p < 0.01$), supporting the idea that the cluster of low-tech firms corresponds to the group of interest (fake innovators).

Apart from size and web sophistication, we also looked at the sectoral composition of these clusters. We found that Cluster 2 firms (low digital sophistication) spanned multiple industries, but interestingly, the IT & Software sector showed one of the highest proportions of potential fake innovators within that sector. This might seem counterintuitive – one would expect IT firms to be digitally advanced – but it suggests that even in technology-oriented industries, there are instances of firms taking on the innovative startup label without demonstrating commensurate digital innovation on their websites. It underlines that the issue of nominal innovators is not confined to traditionally low-tech sectors.

3.2. Regression Analysis Results

The probit regression analyses shed further light on how the profile of firms with innovative status differs across levels of digital sophistication. Table 1 (omitted here for brevity) summarized the estimated coefficients for the key variables (log sales and log employees) in each sub-sample. Here we highlight the main findings:

- **Sales revenue:** A consistent result across all HTML5 usage groups is that the coefficient on log(sales) is negative and statistically significant, indicating that firms with lower sales (smaller revenue) are more likely to be registered as innovative startups. This negative relationship holds for companies with no HTML5 usage, low usage, medium usage, and high usage alike. However, the magnitude of the effect is strongest for the low and medium HTML5 groups. Smaller firms (in revenue terms) thus seem more inclined or able to obtain innovative status, which is in line with the policy's intent to target new, small businesses, but it also might indicate that some lower-performing firms could be using the status opportunistically.
- **Number of employees:** The effect of firm size measured by employees is more nuanced. In the no-HTML5 group and the high-HTML5 group, the coefficient on log(employees) is negative and significant (for example, in our estimates around -0.16

for no-usage firms and -0.19 for high-usage firms). For the no-HTML5 group, the finding suggests that the few firms with no modern web features that still manage to be labeled innovative are typically very small (perhaps newly founded micro-enterprises that have not developed their web presence). In contrast, in the low and medium HTML5 usage groups, the coefficient on employees is statistically not significant.

Table 1. Regressions results. Source: Authors' elaboration.

	Dependent variable:			
	innovative_startup			
	(1)	(2)	(3)	(4)
log_sales	-0.106***	-0.446***	-0.422***	-0.261***
	(0.009)	(0.150)	(0.137)	(0.044)
log_employees	-0.157***	0.146	-0.038	-0.187**
	(0.028)	(0.356)	(0.236)	(0.094)
Constant	0.046	1.363**	2.067***	1.261***
	(0.030)	(0.600)	(0.674)	(0.197)
Observations	5,598	41	79	439
Log Likelihood	-3,301.230	-10.714	-28.486	-159.695
Akaike Inf. Crit.	6,608.461	27.429	62.973	325.390
Note:	*p<0.1; **p<0.05; ***p<0.01			

Overall, these regression results indicate that across all types of firms, the innovative startup designation skews toward smaller firms (both in revenue and often in headcount), which is expected since the program is aimed at new, small companies. Importantly, however, the fact that this pattern holds even for the group with minimal digital sophistication underscores the concern: many of the innovative startups that lack a strong digital footprint are very small firms, possibly startups in name but with limited resources and technology. These could be precisely the cases where policy support may not translate into actual innovation outcomes.

4. Conclusion

This paper presented a novel application of web data analysis to the problem of identifying "fake innovators"—firms that claim an innovation-oriented status for incentives but exhibit little evidence of innovation activity. By using the adoption of HTML5 web technologies as a proxy for digital sophistication, we were able to differentiate firms with a genuine innovative orientation from those that might be merely nominally innovative. Our case study on Italian startups provides empirical support that this proxy correlates with meaningful differences: innovative startups with very low HTML5 usage tended to be smaller, lower-performing firms clustered together, aligning with the notion that they may not be driving substantial innovation.The key contribution of this work is to show that unconventional digital indicators, like website structure and features, can complement traditional innovation metrics in policy evaluation. For policymakers, integrating such digital proxies could enhance the screening and monitoring of beneficiaries of innovation programs. For instance, agencies could periodically assess the web presence of firms receiving innovation subsidies as one signal of progress or stagnation. If a company consistently shows poor digital engagement alongside other red flags, it might prompt a review to ensure compliance with the spirit of the program.

References

Autio, E., Kenney, M., Mustar, P., Siegel, D., & Wright, M. (2014). Entrepreneurial innovation: The importance of context. Research Policy, 43(7), 1097–1108.

Axenbeck, J., & Breithaupt, P. (2021). Innovation indicators based on firm websites – which website characteristics predict firm-level innovation activity? PLoS ONE, 16, e0251872.

Blazquez, D., & Domenech, J. (2018). Big data sources and methods for social and economic analyses. Technological Forecasting and Social Change, 130, 99–113.

Bottai, C., Crosato, L., Domenech, J., Guerzoni, M., & Liberati, C. (2022). Unconventional data for policy: Using big data to detect Italian innovative SMEs. In Proceedings of the 2022 ACM Conference on Information Technology for Social Good (pp. 338–344). New York, NY: ACM.

Cian, M. (2015). Società startup innovative e PMI innovative. Giurisprudenza Commerciale, 6, 961–986.

Crosato, L., Domenech, J., & Liberati, C. (2021). Predicting SME default: Are their websites informative? Economics Letters, 204, 109888.

Ricci, R. (2016). La disciplina delle start-up innovative: obiettivi, strumenti e prime evidenze. Economia e Politica Industriale, 43(4), 105–125.