5th International Conference on Advanced Research Methods and Analytics (CARMA2023)
Universidad de Sevilla, Sevilla, 2023
DOI: http://dx.doi.org/10.4995/CARMA2023.2023.16466

Can websites reveal a firm's innovativeness? Empirical evidence on Italian manufacturing SMEs

Carlo Bottai[1], Lisa Crosato[2], Josep Domenech[3], Marco Guerzoni[1], Caterina Liberati[1]
[1]Department of Economics Management and Statistics, University of Milano-Bicocca, Italy, [2]Department of Economics, Ca' Foscari University of Venice, Italy, [3]Department of Economics, Universitat Politècnica de València, Spain.

Abstract

Research in innovation usually builds on conventional data such as balance sheets, surveys, patents, or product catalogs. This paper intends to explore unconventional data, specifically web-scraped data, as an information source for innovation studies, proposing a careful procedure to establish the veracity of the linkage between web-based data and firm-level information retrieved from conventional sources. The study regards a sample of Italian manufacturing small and medium enterprises active in 2016, comprehending both innovative and non-innovative firms. It is based on HTML tags, whilst most of the previous literature worked on the web-pages text and related semantics. Our paper provides evidence that the way HTML language is applied to build a corporate website unveils the capabilities of the owner firm, helping to distinguish innovative from non-innovative SMEs.

Keywords: *innovation, SMEs, unconventional data, HTML code, web-scraping*

1. Introduction

A firm's innovativeness surely ingrains its practices and the shared knowledge of its employees, but is not always observable. This complicates the assessment of the presence and intensity of innovative activity, although the innovation economics literature has made significant progress in measuring this phenomenon using balance sheets, surveys, patents, and product catalogs. None of these conventional sources, however, can completely capture such latent features, in particular as far as small and medium-sized enterprises (SMEs) are concerned (OECD [1963; 1992]). Furthermore, the derived innovation policy indicators are not always updated enough to describe the current situation. Balance sheets, for instance, are available at the end of each year and released by data providers with further delay.

This study suggests that SMEs' corporate websites, as outputs of a firm's activity, can represent an additional data source to build indicators of firms' innovative character. Firms typically shape their website as virtual showcases to sell products and share information related to their business. This makes the content of corporate websites highly connected to the economic activity of the firms [Domènech et al., 2012]. Moreover, websites are publicly accessible and regularly updated, so to appear as good candidates to solve some of the limitations of the currently available, conventional, sources. Accordingly, part of literature started scraping websites for research purposes [e.g. Blázquez et al., 2018; Crosato et al., 2021], and to use corporate websites to analyze firms' innovative activity [Libaers et al., 2016; Gök et al., 2015; Héroux-Vaillancourt et al., 2020; Daas and van der Doef, 2020; Kinne and Axenbeck, 2020; Axenbeck and Breithaupt, 2021; Kinne and Lenz, 2021, Ashouri et al., 2022].

Our paper adds to this literature but shifts the focus from the semantic analysis of web-pages text to the HTML code structure of webpages. The HTML code employed to build a corporate website stems from a blending of the company's needs and skills with those of the programmers [Brinck et al., 2001], so that it is sensible to suppose it unveils latent features such as high skills and creativity linked to the innovativeness of a firm. Innovative SMEs are supposed to be oriented towards new products development and commercialization, so we may expect that they want their websites well indexed by search engines and social networks. They employ high-skill workers, so they are keener to adopt new technologies, which should emerge from the HTML structure. Moreover, equipping a website with e-commerce, customer engagements, and user monitoring is easier through particular HTML programming styles. Finally, from a researcher's point of view, the analysis of a much more structured language like HTML is easier and computationally less expensive when compared to the analysis of natural languages applied in previous works.

2. Data Description

Our dataset merges conventional and unconventional data sources, where by *conventional* we refer to data resulting from a traditional design, i.e. originally collected by reference institutions for administrative purposes, but available in the standard matrix format and ready to be used for research purposes. Conventional data sources (Orbis and Aida databases by Bureau van Dijk) were the starting point to build the sample and were essential to divide the sample in innovative and non-innovative firms. Our unconventional source of data is the Wayback Machine of the Internet Archive (https://web.archive.org/). Table 1 summarizes the type of information retrieved from the different sources.

Table 1: Framework for dataset building

Data source and type	**Sample Units**	**Collected variables**
Orbis-BvD (conventional)	Italian Manufacturing SMEs, Active in 2016 with reported website N= 77,993	Sample selection variables - 'status' (active, bankrupt, in liquidation, etc.) - number of employees - total assets - turnover - website URL Company's details: - tax identification number (codice fiscale) - business name - business address (street name, number, and postcode) - telephone number Additional Stratification variables: - industrial sector (NACE) - geographical location (NUTS 2)
Aida-BvD (conventional)	Italian Manufacturing SMEs retrieved from Orbis	Label of innovative SME, as defined by the *Italian Startup Act*
Wayback Machine (unconventional)	Italian Manufacturing SMEs retrieved from Orbis with: - Website present in Wayback machine - Website ownership checked by our matching algorithm N= 43,335	HTML tags used to structure the website's front-page

Italian manufacturing SMEs, active in 2016, were retrieved from Orbis, using the standard definition of Eurostat based on the three firm size variables reported in Table 1. Firms with recorded websites were 77,993 on a total of 116,389. Since Orbis does not classify firms in terms of their innovativeness, we have resorted to its companion dataset Aida, which focuses on Italian firms. Here we have exploited the list of 'innovative' SMEs collected by the Italian Chamber of Commerce's Business Register in compliance with the Italian Startup Act (221/2012 law). Note that the Italian Startup Act has a few advantages with respect to other indicators of firms' innovativeness [Guerzoni et al., 2021]. It concentrates on SMEs who must focus on novel products and it does not base the classification on a single innovation measure, so included firms possess at least one among the usual innovation proxies.

The collection of the web-based information required a previous rigorous screening process to assess the attribution of the website reported in Orbis for each firm, to be reasonably sure that we are observing the true website of the company of interest. This is a fundamental aspect of our sampling design that, to the best of our knowledge, was not taken care for but in a few works (such as Barcaroli et al. [2016]). To this end, we have accessed the 2016 archived version of each firm's website URL, as reported in Orbis, on the Wayback Machine and, in each of the reported websites, we have searched for the company's details collected from Orbis, both in the front-page of the website and in any of the web-pages reachable from a hyperlink contained in the front-page.

At the end of this process we were left with 43,335 SMEs, after removing firms whose 2016 was not archived in the Wayback Machine (about 14%) and firms whose website was not confirmed as their own by our algorithm (about 30%, including websites in which the details of the firms were not available).

Our unconventional, web-based, information was finally scraped from the SMEs verified corporate website, with a procedure similar to the one described in Blázquez et al. [2018] and Crosato et al. [2021]: we have accessed the websites front-page and collected any HTML tag used to structure the page. We have kept only the HTML tag used more than three times in the whole document corpus, thus obtaining a final set composed of 711 HTML tags and six aggregate web-based statistics.

3. Can SMEs websites unveil the innovative character of its owner?

In our sample, SMEs labeled as "innovative" in the Aida dataset were only 178. In order to assure a fair comparison between the two groups SMEs, we structured 100 samples of 680 non-innovative firms stratifying by firm size, region and industry to match the smallest sample of innovative firms.

To spot the difference in the HTML structure of innovative and non-innovative corporate websites, each of the collected descriptive statistics and HTML tags were then compared on each of the one hundred non-innovative samples against the group of innovative SMEs.

Results suggest that corporate websites of innovative SMEs appear to be bigger either when measured by the HTML code underlying their front-page, and by the embedded text as measured by the variables reported in Table 2. The variables *text_size* and *gztext_size* represent the amount of text used on the page, but the latter does not take into account repetitions; *html_size* measures the size of the HTML code of the web-page analyzed; *href_number* and *img_number* represent the number of hyperlinks and images present on the page, respectively. Finally, *linkhref_number* counts the number of external resources used by the page. The hypothesis of the two samples being drawn from the same distribution is strongly rejected by both the Kolmogorov-Smirnov (KS) and the Mann-Whitney tests: almost all the considered features show median p-values (Table 2) under the 5% significance level. We can clearly see that the size distributions of the HTML and the text shift to larger values for innovative firms with respect to controls (Figure 1).

Table 2: P-value of Kolmogorov–Smirnov (KS) and Mann–Whitney (MW) tests comparing the distribution of web-based variables (innovative SMEs VS one hundred samples of control firms, median pvalue)

Variable	KS-test	MW-test
text_size	0.003	0.001
gztext_size	0.002	0.001
html_size	0.001	0.000
href_number	0.001	0.000
img_number	0.083	0.012
linkhref_number	0.000	0.000

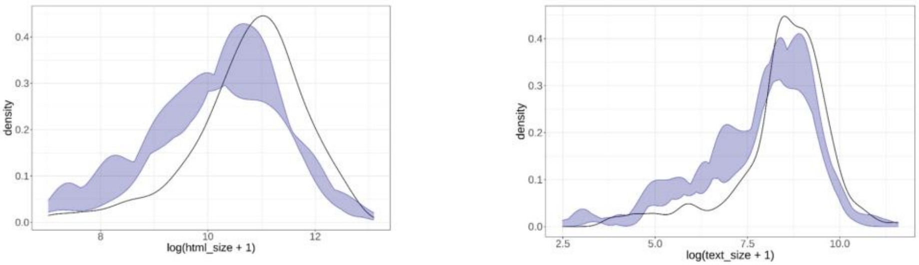

Figure 1: Density distributions of two descriptive statistics about the web-pages for innovative (solid line) and non-innovative (lilac fan on one hundred samples) SMEs.

As about the HTML tags used in the source code of the front-pages, several of them seem to discriminate between innovative SMEs and control firms. In this case, we have added the χ2 test for independence between belonging to the innovative group and presence of the feature, since for a few tags the KS test and the MW test disagreed. In Table 3 we have grouped the tags according to whether they discriminate in conformity with two out of three tests (moderate discriminating power) or with all of the three (highly discriminating power). We reject the null of similarity or independence if the median p-value of the test repeated over the 100 samples is smaller than 5%.

Table 3: HTML tags grouped according to moderate (2 out of three test) or high (3 out of 3 tests) discriminating power.

Number of test rejecting the hypothesis of similarity/independence	**TAGS**
Two out of three	<a>, <div>, <embed>, <link>, <meta>, <nav>, <object>, <p>, <param>, <script>, <section>, <style>
Three out of three	<footer>, <header>, <h>, <i>, <li>, <span>, <table>, <td>, <tr>, <ul>

Among the first group of tags, a few of them are of the kind essential for websites building: <div>, <a>, <p>, independently of the degree of innovativeness of the firm. On the other way round, highly discriminating tags include <footer> or <header>.

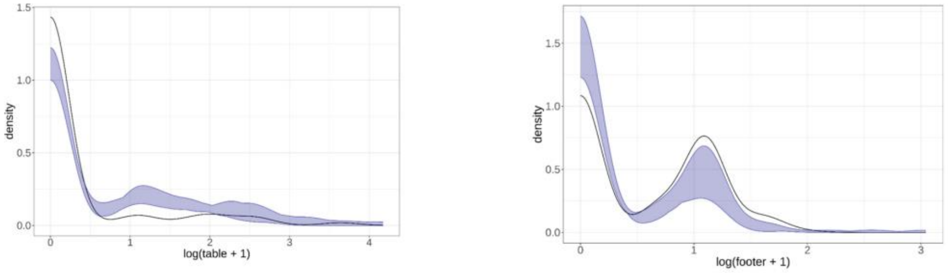

Figure 2: Density distributions of the number of times the indicated tag is used in a web-page for innovative (solid line) and non-innovative (lilac fan on one hundred samples) SMEs.

The distribution of occurrences of the former tag is represented in the right panel of figure 2 and shows that <footer> is more frequent in the front-pages of the websites of innovative SMEs. Tags like <table> (Figure 2, left panel) or <embed> are instead less used by innovative SMEs. Tables are nowadays deprecated, in favor of alternatives tailored for mobile devices.

The tag <embed>, mostly used to include Adobe Flash content in web-pages, is less present in the innovative SMEs web-pages since Adobe Flash was gradually set aside since the early 2010s. These examples confirm that corporate websites of the innovative SMEs rely more on the modern HTML (HTML5) with respect to those of comparable non-innovative firms.

4. Conclusions

In this paper, we have proposed and explored the use of unconventional data scraped from corporate websites as a complementary source of information for identifying innovative SMEs. Our results point out some of the characteristics shaping either group of firms: we found bigger websites and more updated HTLM language for the innovative SMEs group. These findings, although preliminary, confirm the underlying hypothesis that the HTML code of corporate websites and its characteristics represent observable proxies high skills and ingeniousness, characterizing the ability of an SME to embrace innovation. We thus provide the first contribution trying to translate the HTML code of corporate websites into data for identifying innovative firms and derive innovation policy indicators, relatively inexpensive to build and easy to be constantly updated. Ongoing research pursues an unsupervised learning approach to understand whether a natural grouping of the HTML tags emerges from the data.

Acknowledgments

We thank the Italian Ministry of University and Research (MUR) for sponsoring this work under the ‘Departments of Excellence 2018-2022’ funding schema, and the DEMS Data Science Lab of the University of Milano–Bicocca for computational resources. Josep Domenech acknowledges that this research was partially funded by MCIN/AEI/10.13039/501100011033 under grant PID2019-107765RB-I00.

References

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

Axenbeck, J. and P. Breithaupt (2021). “Innovation indicators based on firm websites—Which website characteristics predict firm-level innovation activity?”. PLOS ONE, 16(4):1–23.

Barcaroli, G., M. Scannapieco, and S. Donato (2016). “On the use of Internet as a data source for official statistics: A strategy for identifying enterprises on the Web”. Rivista Italiana di Economia Demografia e Statistica, 70(4):25–41.

Blázquez, D., J. Domènech, and A. Debón (2018). “Do corporate websites’ changes reflect firms’ survival?” Online Information Review, 42(6):956–970.

Brinck, T., D. Gergle, and S. D. Wood (2001). Usability for the Web: Designing Web Sites that Work. Elsevier.

Crosato, L., J. Domènech, and C. Liberati (2021). "Predicting SME's default: Are their websites informative?" Economics Letters, 204:109888.

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

Domènech, J., B. de la Ossa, A. Pont, J. A. Gil, M. Martinez, and A. Rubio (2012). "An intelligent system for retrieving economic information from corporate websites". In IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 573–578.

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

Guerzoni, M., C. R. Nava, and M. Nuccio (2021). "Start-ups survival through a crisis. Combining machine learning with econometrics to measure innovation". Economics of Innovation and New Technology, 30(5):468–493.

Héroux-Vaillancourt, M., C. Beaudry, and C. Rietsch (2020). "Using web content analysis to create innovation indicators—What do we really measure?". Quantitative Science Studies, 1(4):1601–1637.

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

Kinne, J. and D. Lenz (2021). "Predicting innovative firms using web mining and deep learning". PLOS ONE, 16(4):1–18.

Libaers, D., D. Hicks, and A. L. Porter (2016). "A taxonomy of small firm technology commercialization". Industrial and Corporate Change, 25(3):371–405.

OECD (1963). Frascati Manual: The Proposed Standard Practice for Surveys of Research and Experimental Development. OECD Publishing.

OECD (1992). Oslo Manual: OECD Proposed Guidelines for Collecting and Interpreting Technological Innovation Data. OECD Publishing.