CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.22442

Coding Styles and Financial Distress: Evidence from HTML Structure of Italian SMEs' Websites

Carlo Bottai[1], Lisa Crosato[2], Caterina Liberati[3]

[1]Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy, [2]Department of Economics, Ca' Foscari University of Venice, Italy, [3]WebSight Observatory, University of Milano-Bicocca, Italy (https://websight.unimib.it/).

How to cite: Bottai, C.; Crosato, L.; Liberati, C. (2026). Coding Styles and Financial Distress: Evidence from HTML Structure of Italian SMEs' Websites. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 26-28 June 2026. https://doi.org/10.4995/CARMA2026.2026.22442

Abstract

This study investigates whether the websites of financially distressed small and medium-sized enterprises (SMEs) exhibit distinctive latent coding styles, drawing on a sample of Italian firms operating in the wholesale and retail trade sector (NACE Rev. 2, Section G). Web content was systematically collected and analysed through a computational coding approach based on Multiple Correspondence Analysis, enabling the identification of latent structural configurations embedded in firms' online presence. Website alignment with each identified coding style was then quantified through the Adherence measure (Bottai et al., 2024), a normalised indicator capturing the degree to which each website conforms to a given structural profile. Results confirm the existence of distinct coding style groups, differentiated by structural complexity and level of interactivity, and suggest a systematic association between these profiles and financial distress.

Keywords: *HTML Tag; Corporate Websites; Coding Styles; SMEs; Financial Distress.*

1. Introduction

Corporate websites have become increasingly recognised as valuable sources of information about firm behaviour, capabilities, and strategies. A growing body of literature has explored the potential of web-scraped data to construct firm-level indicators, particularly in the domain of innovation and R&D activity — from early small-sample studies (Gök et al., 2015) to large-scale efforts (Kinne and Axenbeck, 2020; Ashouri et al., 2022; Guzman and Li, 2023), as systematically reviewed by Rammer and Es-Sadki (2023). Web-based data have proven especially valuable in overcoming well-known limitations of conventional sources: balance sheet data are typically released with a significant time lag, suffer from opacity issues that are

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

more acute for smaller firms, and may be unreliable for newly established businesses. However, the use of these alternative sources remains constrained by substantial methodological challenges. Textual information retrieved from websites is inherently unstructured, requiring appropriate analytical frameworks to extract meaningful signals, and must be handled through technical solutions capable of managing data sources that grow and change over time (Daas and van der Doef, 2020). Studies applying web data to financial distress prediction remain comparatively scarce: notable exceptions include Putra et al. (2020), who construct web-based features from social media for distress prediction, Stevenson et al. (2021), who analyse firm-level textual content, and Rozo et al. (2023), who demonstrate that online browsing behaviour can improve default prediction models. A common limitation of these contributions is their reliance on textual content (precisely the most volatile component of a website) which raises concerns about the stability of derived indicators over time. An alternative approach, proposed by Crosato et al. (2021, 2024), shifts the focus to the underlying HTML structure of corporate websites, which reflects cumulative technical choices rather than explicitly communicated information, and is inherently more stable. Building on this strand of research, and addressing the gap in the literature on web-based early warning systems, the present study investigates whether SMEs experiencing financial distress, defined as firms exiting the market within the following year, share distinctive latent HTML coding styles, and whether such styles differ systematically from those observed among financially sound firms.

2. Building Coding Styles and measuring website digital alignment

The sample includes Italian SMEs operating in the wholesale and retail trade sector (NACE Rev. 2, Section G) retrieved from the AIDA–Orbis for Italy database by Moody's (formerly Bureau van Dijk). To ensure the availability and validity of web-based data, only firms with a website recorded in AIDA and whose online presence could be verified and retrieved from the Wayback Machine (WBM) were retained, yielding a final sample of 37,012 firms, of which 486 are classified as financially distressed and 36,526 as sound. Distress is defined as market exit within the following year according to the definition adopted in the literature (Pederzoli et al., 2013; Crosato et al., 2023). Since the study focuses on firms' digital capabilities and their willingness to invest in an online presence, websites hosted on social media platforms (e.g., Facebook, Instagram) or generic web-building services (e.g., WordPress.com) were excluded, to retain only company-driven digital initiatives.

For each firm in the sample, the HTML source code of the corporate website was retrieved and parsed, yielding a binary summary matrix recording the presence or absence of each HTML tag across all observed websites. Individual tags, however, are numerous and potentially uninformative in isolation, as they exhibit strong interdependencies — tags are routinely co-deployed in HTML code to render specific features such as embedding images or videos, or

Article in Press

creating tables. Analysing them separately would therefore produce results of limited interpretability and would be susceptible to dimensionality and collinearity issues.

To address this, Multiple Correspondence Analysis (MCA) was applied to the website-by-tag binary matrix, with the aim of summarising the information contained in the full set of categorical indicators by projecting it into a low-dimensional Euclidean space (Benzécri et al., 1973). The first three principal coordinates of the HTML tag categories obtained from MCA were retained (71% explained inertia), and the column principal coordinates were used as input for a hierarchical clustering algorithm based on Ward's linkage criterion and Euclidean distance. This tandem approach stabilises clustering results by identifying groups within a reduced, noise-filtered factorial space (Lebart et al., 1984). The resulting dendrogram is displayed in Figure 1. The optimal number of clusters was determined by jointly considering four internal validation criteria — the Pseudo-F index (Calinski and Harabasz, 1974), the Friedman-Rubin index (Friedman and Rubin, 1967), the C-index (Hubert and Levin, 1967), and the Silhouette coefficient (Rousseeuw, 1987) — converging on a solution of 11 clusters, each representing a distinct latent HTML coding style.

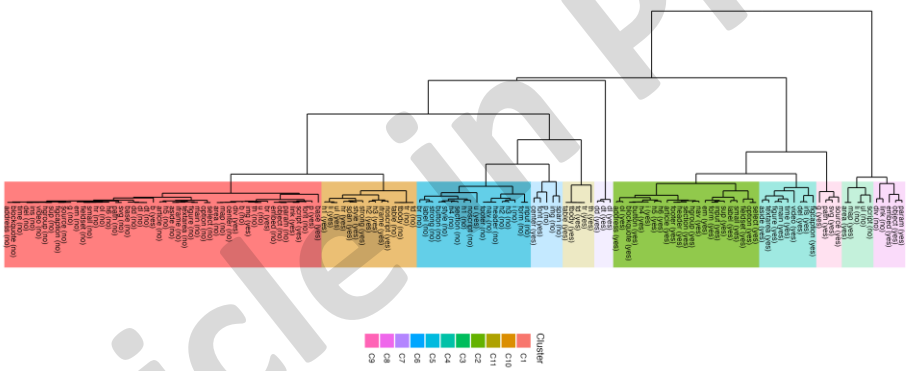


Figure 1. Dendrogram of the hierarchical clustering.

The 11 clusters identify a spectrum of coding styles, each characterised by a specific combination of present and absent tags drawn from the full HTML corpus[1] under analysis. For the sake of brevity, a detailed taxonomy of coding styles is omitted; however, the identified profiles span from legacy-driven configurations, characterized by rudimentary tag sets, to advanced architectural patterns aligned with modern web standards and high interactivity. Once the coding styles were identified, reducing the original tag matrix along the column dimension,

[1] The HTML corpus is a finite collection of machine-readable markup data structured as a sequence of tokens (tags). It represents the raw dataset used for empirical analysis.

Article in Press

the subsequent challenge was to associate each website with the resulting profiles in a meaningful and quantifiable way. To this end, we employed the Adherence measure (Bottai et al., 2024), defined as the proportion of a given cluster's tags that are present in a specific website. This yields a normalised indicator, bounded between 0 and 1, capturing the degree to which each website conforms to each of the 11 coding styles. Values approaching 1 indicate near-complete alignment with a given style, while lower values signal partial or negligible adherence.

3. Future Research

The present study represents a first step towards understanding the relationship between the structural configuration of corporate websites and firm-level financial vulnerability. Several promising directions remain open for further investigation. A natural extension of this work would be to test whether the adherence measures derived from the identified coding styles are significantly associated with financial distress, both when considered in isolation and when jointly modelled alongside traditional balance sheet indicators capturing firm solvency and leverage. Should the adherence measures retain statistical significance in the presence of financial covariates, as we expect, this would provide evidence that HTML coding styles carry informational content beyond what is already reflected in accounting data, representing an additional and complementary signal for explaining financial distress.

References

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.

Benzécri, J.P. (1973). *L'Analyse des Données* vol 2. Dunod.

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

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-Theory and Methods,* 3(1), 1-27.

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

Crosato, L., Liberati, C., & Repetto, M. (2023). Lost in a black-box? Interpretable machine learning for assessing Italian SMEs default. *Applied Stochastic Models in Business and Industry,* 39(6), 829–846.

Crosato, L., Domenech, J., & Liberati, C. (2024). Websites' data: a new asset for enhancing credit risk modeling. *Annals of Operations Research,* 342,1671–1686

Friedman, H. P., & Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association*, 62(320), 1159-1178.

Article in Press

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

Guzman, J., Li, A. (2023). Measuring founding strategy. *Management Science,* 69(1),101-118.

Hubert, L. J., & Levin, J. R. (1976). A general statistical framework for assessing categorical clustering in free recall. *Psychological Bulletin*, 83(6), 1072–1080.

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

Rammer, C., & Es-Sadki, N. (2023). Using big data for generating firm-level innovation indicators: A literature review. *Technological Forecasting and Social Change,* 197, 121876.

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

Lebart, L., Morineau, A., & Warwick, K.M. (1984) *Multivariate Descriptive Statistical Analysis: Correspondence Analysis and Related Techniques for Large Matrices*. John Wiley & Sons, New York

Pederzoli, C., Thoma, G., Torricelli, C. (2013). Modelling credit risk for innovative SMEs: the role of innovation measures. *Journal of Financial Services Research,* 44,111–129.

Putra, S. G. P., Joshi, B., Redi, J., & Bozzon, A. (2020) A credit scoring model for SMEs based on social media data. *In International conference on web engineering* (pp. 113-129). Cham: Springer International Publishing.

Rozo, B. J. G., Crook, J., & Andreeva, G. (2023) The role of web browsing in credit risk prediction. *Decision Support Systems* 169, 113879.

Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Stevenson, M., Mues, C., & Bravo, C. (2021). The value of text for small business default prediction: A deep learning approach. *European Journal of Operational Research,* 295(2), 758–771.

Article in Press