CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21942

Using website text to improve counterfactual construction in firm-level policy evaluation

Eugeni Gil-Ocana[1], Pablo de Pedraza[1], Ana Garcia-Bernabeu[1]
[1]eSMART Research Center, Universitat Politècnica de València, Spain.

How to cite: Gil-Ocana, E.; de Pedraza, P.; Garcia-Bernabeu, A. (2026). Using website text to improve counterfactual construction in firm-level policy evaluation In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.21942

Abstract

This paper investigates whether textual information from firm websites can improve the construction of counterfactuals in the evaluation of firm-level public subsidy programs. We integrate financial data, administrative subsidy records, and semantic embeddings derived from website text into a matching-based causal inference framework. Using the Spanish Kit Digital program as an empirical illustration, we compare matching specifications based on financial covariates, textual embeddings, and their combination. The results suggest that incorporating textual information can capture strategic firm characteristics not observed in financial data, improving the comparability of treated and control firms.

Keywords: *matching; text-data; causal evaluation; vim*

1. Introduction

The rapid expansion of digital technologies has transformed both economic activity and public governance. In the public sector, increasing digitalization has expanded the availability of administrative microdata on public support schemes to firms, creating new opportunities to evaluate policy interventions at the firm level (Srhoj et al. 2021). Credible evaluation typically requires linking subsidy records with firm-level financial information and constructing datasets that allow researchers to compare treated and untreated firms before and after the intervention. Access to pre-treatment characteristics is particularly important to account for non-random selection into treatment and to construct valid counterfactuals. Despite these advances, existing evaluation datasets remain incomplete along dimensions that are central to both policy targeting and firm behavior. Many firm characteristics, particularly those related to activities, products, or areas of specialization, are rarely observed in standard accounting or administrative datasets, which primarily capture financial outcomes. As a result, firms that appear comparable in terms

Article in Press

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

of size, sector, or financial performance may nonetheless differ in underlying characteristics that influence their behavior and their response to policy interventions.

A large share of this missing information is embedded in unstructured textual sources, including firm websites (Blazquez et al., 2018; Crosato et al., 2021). Corporate websites often contain diverse information that can reflect otherwise unobserved aspects of firm behavior and strategy. However, because these sources are unstructured, they are rarely incorporated into quantitative evaluation frameworks. Ignoring these qualitative dimensions may lead to counterfactual constructions that are financially comparable but strategically misaligned, thereby undermining the validity of causal estimates.

Recent advances in natural language processing provide new tools to address this limitation. Semantic embeddings allow textual information to be represented as numerical vectors that capture meaningful similarities between documents. These representations can potentially serve as proxies for otherwise unobserved firm characteristics and can therefore enrich the set of pre-treatment covariates used in causal evaluation. This paper therefore explores whether textual information extracted from firm websites can serve as proxies for otherwise unobserved firm characteristics and improve the construction of credible counterfactuals in firm-level policy evaluation. More specifically, we investigate whether semantic representations of website content can be incorporated into matching procedures to better capture strategic dimensions of firms that are not observable in traditional financial datasets. By expanding the set of observable firm characteristics, textual data may help reduce hidden bias arising from unobserved heterogeneity in firm-level policy evaluation.

This paper makes three main contributions. First, it shows how unstructured textual information from firm websites can be used as proxies for otherwise unobserved strategic firm characteristics in causal evaluation settings. Second, it proposes a structured data pipeline that integrates administrative subsidy records, firm-level financial data, and website-derived textual information into a unified evaluation-ready dataset. Third, it demonstrates how semantic embeddings derived from website content can be incorporated into matching procedures to improve the construction of counterfactual firms.

The proposed framework is illustrated through an empirical application combining firm-level financial data for Spanish companies with textual information extracted from company websites and administrative records of public subsidies. While the empirical results should be interpreted as a preliminary illustration, they demonstrate how textual information can be systematically integrated into causal evaluation pipelines.

Article in Press

2. Conceptual framework: incorporating web pages into causal matching

Matching-based evaluation methods rely on the conditional independence assumption (CIA), which requires that treated and control units be comparable across all relevant pre-treatment characteristics (Imbens & Rubin, 2015). In firm-level policy evaluations, researchers typically match firms using structural variables such as firm size, revenue, employment, or sector classification (Criscuolo et al., 2019; Srhoj et al., 2021). While these covariates capture important financial scale metrics, they often fail to reflect other latent organizational characteristics—such as strategic focus, management quality, or digital maturity—which are inherently correlated with both policy participation and firm outcomes. To bridge this gap, we adopt the framework proposed by Gentzkow, Kelly & Taddy (2019), who posit that textual data should be treated as a primary feature in econometric modeling rather than a mere descriptive accessory. As they emphasize, when textual content is structured into high-dimensional representations, it reveals rich, latent information about the economic agent's behavior. Recent empirical research supports this, showing that website text provides meaningful signals of innovation activities (Kinne & Lenz, 2021), sectoral dynamics (Fernández-Navarro, forthcoming), or export orientation (Blazquez & Domenech, 2018).

By transforming firm website content into semantic vectors, we effectively "uncover" dimensions that traditional financial datasets omit. Building on this insight, the approach proposed in this paper incorporates these semantic representations into the set of pre-treatment covariates used for matching. This methodology is not restricted to any specific matching algorithm; rather, it enriches the underlying covariate space used for distance-based measures. By capturing dimensions of firm similarity such as strategic intent or operational complexity that remain hidden in standard datasets, the method strengthens the plausibility of the unconfoundedness assumption, thereby enhancing the credibility of counterfactual selection in matching-based causal inference.

2.1. Text representation

We extract textual content from firm websites, segmenting it into manageable chunks (Pastor-Merino et al., 2025; Domenech et al., forthcoming). To preserve semantic integrity, we avoid traditional linguistic preprocessing, opting instead for high-dimensional vector representations generated with the OpenAI *text-embedding-3-small* model. Each fragment is mapped to a $d = 1536$ dimensional vector.

To aggregate multi-chunk documents into a single firm-level representation, we employ mean pooling, computing the unweighted average of the segment embeddings:

$$v_{firm} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad (1)$$

Article in Press

where $x_i$denotes the embedding vector of the i-th text segment, with each segment corresponding to a fragment of the firm's website content, and $N$is the total number of segments. Each $x_i$is a high-dimensional vector capturing the semantic meaning of the underlying text.

Mean pooling is a common aggregation strategy for deriving fixed-length representations from variable-length sequences of contextual embeddings and has been shown to outperform alternatives such as [CLS]-token or max pooling on semantic similarity tasks (Reimers & Gurevych, 2019). In our setting, mean pooling offers two practical advantages, it requires no additional learnable parameters, and it ensures that all segments of a website contribute equally to the firm-level representation.

2.2. Matching framework

To construct valid counterfactuals, we adopt the Variable Importance Matching (VIM) framework (Lanners et al., 2023). Unlike propensity score methods, which risk "black-box" dimensionality reduction (King & Nielsen, 2019), VIM directly addresses high-dimensional covariate spaces by weighting variables based on their predictive power regarding the outcome Y.

We estimate a Lasso model to predict Y using both financial covariates and embedding dimensions. The Lasso solves the following optimization problem:

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2N} \sum_{i=1}^{N} (Y_i - X_i^{\top} \beta)^2 + \lambda \sum_{j=1}^{K} | \beta_j | \right\} \quad (2)$$

where $X_i$is the vector of covariates for firm i, $\beta_j$are the coefficients to be estimated, $K$is the total number of covariates, and $\lambda$is a regularization parameter controlling the sparsity of the solution. The L_1-penalty shrinks less informative coefficients toward zero, effectively selecting only the covariates most predictive of Y.

The absolute values of the estimated coefficients $| \beta_j |$are then used as weights in a weighted Manhattan distance metric:

$$d(a, b) = \sum_{j=1}^{K} | \beta_j | \cdot | x_{aj} - x_{bj} | \qquad (3)$$

where $x_{aj}$and x_{bj}denote the values of covariate $j$for firms $a$and b, respectively. This metric ensures that the matching procedure prioritizes dimensions—whether financial or semantic—that are empirically relevant to the outcome, providing an auditable and robust alternative to Mahalanobis distance-based approaches.

Article in Press

2.3. Quality and diagnostics

To evaluate the added value of textual embeddings, we compare three specifications: (i) Baseline (financial data only), (ii) Embeddings (textual data only), and (iii) Hybrid (integrated features). We assess covariate balance across specifications using Standardized Mean Differences (SMD):

$$SMD = \frac{| \bar{x}_t - \bar{x}_c |}{\sqrt{\frac{s_t^2 + s_c^2}{2}}} \qquad (4)$$

where $\bar{x}_t$ and $\bar{x}_c$ denote the mean of covariate x in the treated and control groups, respectively, and s_t^2 and s_c^2 are the corresponding sample variances. The SMD measures the standardized difference between treated and control units for each covariate, providing a scale-free assessment of covariate balance.

An $SMD < 0.1$is commonly used as a benchmark for acceptable balance, indicating that treated and control groups are sufficiently similar along the observed covariates. Achieving this threshold demonstrates that the VIM framework effectively reduces selection bias by leveraging high-dimensional semantic signals in addition to traditional financial observables.

3. Empirical application: digitalization subsidies

We illustrate our methodology by analyzing the Kit Digital program, a Spanish public initiative designed to enhance firm-level digital presence. Our dataset integrates firm-level financial data from SABI (Bureau van Dijk, 2025), website textual content, and official records from the Base de Datos Nacional de Subvenciones (BDNS). Treatment status $D_i = 1$ is assigned to firms that received the Kit Digital subsidy in 2023, while untreated firms constitute the control pool. Given that the subsidy specifically targets digital maturity, assessing the firm's initial digital status via website content is crucial for causal identification. We utilize website text collected in 2022 to maintain temporal precedence.

The final analytical sample comprises 17,769 firms (2,552 treated, 15,217 controls). We implement the Variable Importance Matching (VIM) framework by estimating an L_1-regularized model (Lasso) to predict firm sales revenue in 2024. This procedure assigns higher weights to covariates that are most predictive of firm scale, effectively filtering out non-informative variables in high-dimensional settings.

Table 1 reports the predictive relevance of our covariates. Given the high dimensionality of the textual embeddings, weights are aggregated by category, showing the absolute sum of coefficients $| \beta_j |$and their relative contribution to the total model importance.

Article in Press

Table 1. Predictive relevance of financial observables and textual embeddings. Own source

Variable Category	Specification: Baseline	Specification: Embeddings	Specification: Hybrid (VIM)
Financial Observables (Aggregated)	1.19 (100%)	-	1.17 (72.34%)
Textual Embeddings (Aggregated)	-	1.30 (100%)	0.45 (27.66%)
R^2	0.916	0.355	0.920
N Variables Selected	6	625	186

Table 1 shows that financial covariates remain the dominant predictors, but the Hybrid specification indicates that textual embeddings provide meaningful, non-redundant information, accounting for 27.66% of total model importance. The selection of 186 embedding dimensions by the L_1 penalty suggests that the latent semantic space captures aspects of firm identity not fully reflected in balance sheet ratios. The slightly higher predictive fit of the Hybrid model ($R^2 = 0.920$) provides evidence that unstructured textual data complements financial covariates in explaining firm performance.

Using the Lasso-derived weights, we compute a weighted Manhattan distance to match each treated firm with its nearest control (1:1 matching with replacement). This ensures that the matching procedure prioritizes covariates that are most predictive of the outcome variable, combining both financial and textual information. Covariate balance is evaluated using Standardized Mean Differences (SMD)

Table 2. Covariate balance diagnostics SMD. Own source

Specification	Mean SMD	Max SMD
Financial Only	0.015	0.042
Embeddings Only	0.079	0.185
Hybrid (VIM)	0.097	0.210

As shown in Table 2, all specifications achieve mean SMD values below the recommended threshold of 0.1. While the Financial-only model exhibits the lowest SMD, this reflects the smaller dimensionality of its covariate space. Incorporating textual embeddings increases the number of covariates and introduces finer granularity, requiring the model to satisfy more stringent similarity criteria. Although some individual textual variables exhibit higher residual imbalance due to the high dimensionality and diversity of semantic features, the Hybrid specification successfully conditions the counterfactuals on both traditional financial characteristics and the firm's latent digital identity, maintaining an overall acceptable mean SMD ($SMD_{mean} < 0.1$).

Article in Press

This indicates that, despite the added complexity of semantic features, the Hybrid approach preserves the quality of the counterfactuals and provides a robust and contextually informed identification strategy. By capturing the multidimensional profile of each firm including structural, strategic, and digital dimensions it reduces the risk that estimated treatment effects are confounded by unobserved heterogeneity, ensuring credible comparisons between treated and control firms.

4. Conclusion and discussion

This paper proposes a methodological framework that integrates unstructured textual information into causal inference by incorporating website embeddings into the Variable Importance Matching (VIM) procedure. By capturing latent dimensions of firm characteristics present in website content but not reflected in traditional balance sheets, our approach constructs more credible counterfactuals. The empirical application to the *Kit Digital* program demonstrates that while financial covariates retain primary predictive power, textual embeddings contribute 27.66% of the total predictive weight in the hybrid model. This integration improves covariate balance, indicating that unstructured textual data provides additional informative dimensions that improve covariate balance and the quality of counterfactual matches. Beyond the specific program analyzed, the proposed pipeline is generalizable to any firm-level policy evaluation context in which unstructured textual data can serve as a proxy for unobserved characteristics. This positions website content analysis as a practical complement to traditional administrative data in program evaluation.

While our findings underscore the potential of textual data, several avenues for refinement remain. First, regarding the estimation procedure, implementing the L_1 regularization on a dedicated training set could further mitigate potential data contamination risks, ensuring the predictive weights remain robust to overfitting. Second, future research should explore alternative dimensionality reduction techniques (e.g., variational autoencoders) and more sophisticated text aggregation strategies to capture longitudinal shifts in firm narrative. Finally, extending this framework to other unstructured sources, such as social media sentiment or corporate sustainability reports, could provide a more holistic view of firm characteristics.

Overall, this research highlights that textual data is not merely supplementary but a step towards enriching causal evaluation pipelines. By transforming web-based information into actionable covariates, our methodology enhances the transparency and credibility of policy impact evaluations, providing a more granular understanding of firm-level outcomes in an increasingly digitalized economy.

Article in Press

References

Blazquez, D., & Domenech, J. (2018). Web data mining for monitoring business export orientation. *Technological and Economic Development of Economy*, *24*(2), 406–428. https://doi.org/10.3846/20294913.2016.1213193

Blazquez, D., Domenech, J., & Debón, A. (2018). Do corporate websites' changes reflect firms' survival? *Online Information Review*, *42*(6), 956–970. https://doi.org/10.1108/OIR-11-2016-0321

Criscuolo, C., Martin, R., Overman, H. G., & Van Reenen, J. (2019a). Some causal effects of an industrial policy. *American Economic Review*, *109*(1), 48–85. https://doi.org/10.1257/aer.20160034

Crosato, L., Domenech, J., & Liberati, C. (2021). Predicting SME's default: Are their websites informative? *Economics Letters*, *204*. https://doi.org/10.1016/j.econlet.2021.109888

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. In *Journal of Economic Literature* (Vol. 57, Issue 3, pp. 535–574). American Economic Association. https://doi.org/10.1257/jel.20181020

Imbens, G. W., & Rubin, D. B. (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge: Cambridge University Press.

King, G., & Nielsen, R. (2019). Why propensity scores should not be used for matching. *Political Analysis*, 27(4), 435–454. https://doi.org/10.1017/pan.2019.11

Kinne, J., & Lenz, D. (2021). Predicting innovative firms using web mining and deep learning. PLoS ONE, 16(4), e0249071. https://doi.org/10.1371/journal.pone.0249071

Lanners, Q., Parikh, H., Volfovsky, A., Rudin, C., & Page, D. (2023). Variable importance matching for causal inference. *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, in *Proceedings of Machine Learning Research*, 216, 1174–1184.

Pastor-Merino, A., Martínez-Barbero, X., & Domenech, J. (2025). Detecting AI adoption at scale: a web mining and LLM methodology. *Scientometrics*. https://doi.org/10.1007/s11192-025-05468-y

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3982–3992

Srhoj, S., Lapinski, M., & Walde, J. (2021). Impact evaluation of business development grants on SME performance. *Small Business Economics*, *57*(3), 1285–1301. https://doi.org/10.1007/s11187-020-00348-6

Article in Press