From Text to Economic Exposure: Measuring Intrinsic Risk in Urban Real Estate

Roberto Nares-Alcalá[1], Antonio Martínez-Raya[1], and Alejandro Segura-de-la-Cal[2(✉)]

[1] Department of Organizational Engineering, Business Administration and Statistics, Universidad Politécnica de Madrid, 28040 Madrid, Spain
[2] Department of Architectural Constructions and Control, Universidad Politécnica de Madrid, Avenida de Juan de Herrera 6, 28040 Madrid, Spain
alejandro.segura@upm.es

Abstract. Corporate disclosures usually describe the firm's strategies, expectations, and perceived risks. Beyond purely numerical indicators, these narratives provide a broader and often more nuanced view of the firm. In this sense, working with textual information allows researchers to capture and organize these insights in a more structured and meaningful way. The starting point of this paper is that traditional classifications, such as sector categories or fixed developer typologies, are often too rigid to reflect the diversity of the firms that are operating in the built environment. In practice, mixed-use models, infrastructure-linked assets, and the changing nature of urban functions tend to blur these boundaries. For this reason, the study has adopted an approach that considers firms as interpreters of economic and urban change, drawing on their own published language to capture underlying exposure. The empirical strategy builds a textual corpus for each firm, using sources such as regulatory filings, investor presentations, earnings calls, and business descriptions. Embeddings then convert these texts into semantic representations, enabling them to position firms within a continuous economic space. Rather than assigning each firm to a single category, similarity within this space is used to infer partial exposure to different economic narratives. This makes it possible to identify shared sensitivities to structural drivers, such as financing conditions, tenant composition, logistics demand, retrofit pressures, technological change, and urban redevelopment. Afterward, it incorporates these semantic exposures into a multi-factor risk model, which distinguishes common structural risk from firm-specific residual risk. By incorporating semantic exposures into a multi-factor risk model, the analysis distinguishes common structural risk from firm-specific residual risk. Real estate benefits particularly from this framework. Interest-rate shocks, liquidity stress, and capital allocation changes can alter firm risk in real estate before conventional classifications fully capture these alterations. More broadly, the study links firm-level narratives, asset pricing, and urban transformation, showing how the analysis of corporate text can contribute to a more adaptive understanding of risk.

Keywords: real estate · urban transformation · capital allocation · risk · semantic embeddings · natural language processing

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2027
O. Gervasi et al. (Eds.): ICCSA 2026 Workshops, LNCS 16765, pp. 282–297, 2027.
https://doi.org/10.1007/978-3-032-30536-7_18

1 Introduction

Urban real estate firms operate where financial markets, spatial change, infrastructure development, demographic shifts, and regulations meet. Their risk profile is not shaped just by accounting figures, asset values, or formal classifications. Interest rates also influenced it, as well as refinancing issues, tenant demand, logistics, energy availability, climate change, and redevelopment cycles. Many of these factors appear in company communications before they affect market prices. Annual reports, regulatory filings, investor presentations, business descriptions, and earnings-call transcripts give insights into how firms perceive and communicate changes in the built environment.

The focus of this paper is on dispersed economic knowledge. Hayek argues that relevant economic knowledge is spread among many actors; it is often local, contextual, and incomplete [1]. This is particularly true for urban real estate, where firms hold specific information about tenants, financing conditions, asset degradation, infrastructure limits, regulations, and local demand. Corporate narratives partially reflect how firms understand their economic surroundings. Although these documents are not impartial and may use strategic language, they contain organized information that can be systematically examined.

This paper builds on existing literature that uses text as data in economics and finance. Tetlock shows that media language reveals investor sentiment and financial outcomes [6]. Loughran and McDonald point out that financial language needs particular handling, especially in corporate filings [3]. Baker, Bloom, and Davis use newspaper text to create a measure of economic policy uncertainty [5]. Gentzkow, Kelly, and Taddy offer a general framework for using text as data in economics. Additionally, Hassan et al. review recent applications of textual analysis in economic research, firm disclosures, and risk narratives [2, 4].

Another relevant body of work uses corporate text to assess firm similarity and economic classification. Hoberg and Phillips build text-based industry networks from 10-K product descriptions, illustrating that language can measure firm similarities beyond fixed industry codes [10]. Their related research also uses text used to study product-market synergies and competition [11]. Bybee, Kelly, Manela, and Xiu show how large-scale business news can be turned into topic-based measures of economic conditions [12]. These studies indicate that textual data can yield organized economic variables.

The paper also draws from portfolio theory and factor-risk modeling. Markowitz draws attention to the importance of variance and covariance in portfolio selection [8]. Fama and French reveal that common risk factors explain shared variations in asset returns [13]. Fisher's theory of interest provides a basic account of discounting, time, and valuation [7]. Kelly, Pruitt, and Su connect observable firm characteristics to conditional risk exposures [14]. This paper follows the general structure of factor models but suggests that firm stories can create continuous semantic exposure matrices to complement or partially replace fixed industry classifications.

In spite of these contributions, several gaps exist. First, much of the text-as-data research focuses on sentiment, uncertainty, topic frequency, or product-market similarity. There has been less focus on turning firm narratives serving as economic exposure vectors for risk assessment. Second, traditional sector classifications impose rigid boundaries on firms whose business models are usually mixed. A real estate firm may have simultaneous

exposure to logistics, data infrastructure, healthcare demand, redevelopment, energy limits, or urban densification. Third, urban real estate studies often examine asset values, locations, regulations, and spatial changes, but less often link firm-level narratives to structure financial risk estimation. This creates a gap between qualitative analysis and quantitative risk assessment.

This paper discusses that gap by supplying a method for converting corporate accounts into semantic economic exposures. The contribution is in the technique. The goal is not to suggest that written material alone can gauge risk, or that corporate disclosures offer an unbiased view of reality. Instead, the paper proposes a standardized approach for cleaning textual information, converting it into embeddings, comparing it with meaningful economic categories, and normalizing the results into exposure weights. This approach is consistent with emerging research in urban decision-making, which emphasizes the integration of Artificial Intelligence with structured evaluation frameworks to bridge qualitative and quantitative dimensions in complex urban systems [15].

The proposed method unfolds in five steps. First, a firm-level corpus is created from corporate disclosures, including annual reports, regulatory filings, investor presentations, business descriptions, and earnings-call transcripts. Second, the texts are processed to eliminate duplicates, generic phrases, or uninformative content. Third, each firm corpus and each reference economic category are transformed into vector representations using a specified language-embedding model. This development relies on modern natural language processing, including attention-based architectures, contextual language models, and sentence embeddings designed for comparing semantic similarity [16–18]. Fourth, the semantic similarity between each firm and each category is calculated, for instance, using cosine similarity. Fifth, a parameter controlling the spread or focus of the resulting exposure profile converts into continuous exposure weights. A low value allows firms to maintain broad hybrid exposures, while a high value creates sharper classifications that correspond more closely with traditional categories.

This work is presented in three stages. First, Sects. 2 and 3 lay the conceptual groundwork of the proposal by connecting dispersed knowledge, economic coordination, and the role of firms as interpreters of transforming economic conditions. Second, Sects. 4 to 6 present the central methodology of the paper: using text as data, representing corporate narratives semantically, and transforming semantic similarity into economic exposure. Third, Sects. 7 and 8 apply this system to risk assessment and urban capital allocation, showing how semantic exposures can help measure intrinsic risk in the built environment. Section 9 concludes with a summary of the contributions, limitations, and future research directions.

2 Dispersed Knowledge and Economic Coordination

A central problem in economics concerns how societies coordinate production and allocate scarce resources. The knowledge relevant to economic decision-making is not concentrated in a single authority but dispersed across different actors, like individuals, firms, places, and institutions in general [1]. Prices, contracts, and organizational routines help to aggregate part of the knowledge, although a significant share remains embedded in local judgments, tacit capabilities, and expectations about the future.

This issue becomes particularly visible in sectors such as real estate (the sector that encompasses the largest share of global assets in the form of land, buildings, and infrastructure) and urban development. In these contexts, the determinants of value extend beyond current cash flows and include factors such as regulation, land-use constraints, financing conditions, technological change, demographic trends, and neighborhood transformation.

From this perspective, the economy can be seen as a system of distributed interpretation. Firms do not simply react to market signals; they also generate and communicate their own interpretations of changing economic conditions. Through disclosures, strategic plans, earnings presentations, investor materials, and regulatory filings, firms reveal how they understand the opportunities, risks, and competitive pressures. These narratives not only provide information to shareholders for buying or selling company shares but also offer a window into the underlying diversity of the economy. They reflect differences in business models, capital intensity, exposure to interest rates, sensitivity to urban demand, technological adoption, and organizational strategy. In this sense, corporate disclosures become an important medium through which dispersed knowledge becomes partially observable, provided that appropriate tools and methodologies are used for analysis and decision-making.

This insight is particularly important for capital allocation. In standard neoclassical frameworks, resource allocation is often described as the outcome of relatively clean price signals. In practice, however, investors and managers must interpret information that is incomplete and constantly evolving. This implies that a substantial amount of relevant information may not be fully reflected in prices, and can therefore be extracted to improve forecasts of future prices by analyzing the institutional processes used to produce, share, and evaluate firms' conditions. Firms compete not only in goods and asset markets but also in the credibility and clarity of the narratives they present about their future prospects. Market valuations capture more information than the value of their assets; they capture a stream of interpreted expectations about their growth, resilience, or exposure, which implies that the systematic collection of information is in itself a productive activity. Managers gather information to evaluate investment opportunities, identify rivals, benchmark performance, anticipate regulatory changes, and detect shifts in consumer demand. Analyzing the functioning of urban sectors, especially those highly dependent on the real estate market, becomes particularly complex; asset values present cross-dependencies with multiple overlapping spatial processes that include: i) sociological factors such as migration, local labor demand, or the household family model in the area, ii) extrinsic urban factors such as location, infrastructure, or zoning [19], or iii) economic factors such as financing costs. No single metric can fully capture these dynamics, so the language firms used to describe their activities become an additional source of evidence on where economic change is taking place and how capital may be reallocated.

The increasing digitization of corporate communication further expands the ability to capture this complexity. While earlier research often relied on balance sheet variables or broad industrial classifications, more recent empirical work incorporates richer textual signals. These signals do not resolve the problem of dispersed knowledge; rather, they allow agents to observe additional elements and make new decisions that influence

market prices, thereby altering the information conveyed by those prices. They allow researchers to examine the way firms classify themselves, how they frame risks, how they describe strategic adaptation, and how they position their activities within broader economic and urban systems. In this way, published information becomes both a decision tool for firms and an empirical proxy for otherwise unobservable heterogeneity in the economy [2].

3 Firms as Agents of Economic Interpretation and Valuation

If knowledge in the economy is dispersed, firms occupy a distinctive position as agents that collect, synthesize, and communicate fragments of that knowledge. A firm is not only a production unit but also an interpretive institution. The firm receives signals from suppliers, tenants, creditors, regulators, customers, municipalities, and competitors, and translates them into strategic narratives that not only combine quantitative and qualitative elements but also include important descriptions of business, economic, and social developments with high communicative value. This interpretive role matters because market valuation depends to a large extent on how the future is made understandable. Investors observe not only financial outcomes but also the informational structure through which management explains those outcomes.

The sources of this information are diverse. Regulatory documents, such as corporate annual reports (including, for example, 10-K forms regulated by the Securities and Exchange Commission), provide standardized disclosures on business segments, operating risks, financing conditions, litigation, market competition, and macroeconomic exposure. Investor relations materials add further detail through earnings presentations, asset reports, capital allocation updates, sustainability disclosures, and management commentary. Conference call transcripts provide insight into tone, emphasis, and managerial confidence, while corporate presentations often make strategic direction more explicit, particularly in transforming sectors. Taken together, these sources form a disclosure ecosystem through which firms contribute to a broader interpretation of their operations and of the economy [2].

For empirical research, these documents are valuable because they capture more than simple promotional information. They provide descriptions that are at once structured and heterogeneous of how firms perceive and communicate their environment. A logistics REIT, for example, may emphasize e-commerce penetration, tenant concentration, and the resilience of warehouse demand, while a data-center operator may focus on cloud infrastructure, AI workloads, and power constraints. In contrast, a residential platform may highlight household formation, affordability pressures, or migration patterns. Even within the same sector, firms often reveal quite different economic exposures through their narratives, which leads them to communicate those elements that reflect their particular interests and competitive positioning. This is where textual analysis becomes particularly useful, as it allows researchers to move beyond formal classifications and examine how firms describe the sources of their own value creation.

At the same time, these documents must be interpreted with caution. Corporate communication is inherently strategic and selective. Firms may emphasize favorable themes, downplay uncertainty, or adopt language that suggests innovation without there

necessarily being clear operational backing. This implies that textual information should not be treated as a direct reflection of fundamentals, as its value for firm analysis lies in revealing how management positions the company within a broader set of expectations [3]; this positioning is economically relevant because valuation depends not only on realized performance but also on the credibility and coherence of these narratives, insofar as whether a firm is perceived as defensive or cyclical, innovative or mature, or more or less sensitive to broader economic conditions can influence both its cost of capital and its investor base.

In real estate and the built environment, this interpretive dimension becomes particularly relevant. As these firms are grounded in large portfolios of built assets that are fixed in space, yet whose valuation evolves significantly over time due to factors such as demographic changes, income levels, or economic complexity, their relative value in turn depends on how firms interpret location-specific demand, redevelopment opportunities, financing conditions, regulatory constraints, and infrastructure linkages. For example, when a real estate company presents itself as an urban logistics platform, a digital infrastructure provider, or a life-science ecosystem operator, it is implicitly positioning itself within a broader economic narrative. These choices can influence analyst coverage, peer comparisons, and ultimately portfolio allocation. Firm documents, therefore, are not only descriptive but also form part of the valuation process itself.

Viewing firms as agents of economic interpretation also helps connect micro and macro perspectives. Corporate disclosures can show how broader economic forces are filtered through specific business models. Inflation, interest rate changes, demographic shifts, decarbonization, or urban restructuring do not affect all firms equally; their impact depends on firm-specific characteristics (such as the composition of assets, liabilities, tenants, technologies, or geographic exposure). It is at this point that textual data captures the process by revealing how firms interpret these socioeconomic forces and translate them into strategic priorities aimed at creating shareholder value and enhancing valuation. In this sense, firm-level text provides a useful lens for observing economic interpretation as it unfolds.

4 Text as Data in Economic and Financial Research

The transformation of text into data has become one of the most important methodological developments in contemporary economics and finance [4]. The basic idea is straightforward: written language contains information that can be systematically extracted, quantified, and related to firm-level outcomes. Rather than treating reports, transcripts, and disclosures as purely qualitative material, the use of machine learning and artificial intelligence tools transforms them into analyzable objects. This shift has led to a large body of literature on sentiment, uncertainty, disclosure quality, topic detection, managerial tone, and information asymmetry, significantly expanding the amount of information that can be extracted from such documents.

Early approaches relied primarily on word counts, dictionaries, and bag-of-words representations. These methods enabled important advances, particularly in measuring sentiment or uncertainty from financial news and corporate reports. However, they also present significant limitations. For instance, frequency-based techniques often ignore

context, syntax, and semantic substitution. Two firms may describe similar economic activities using different vocabularies, yet a simple keyword-based model may classify them as unrelated. Conversely, the repeated use of the same term may obscure substantial differences in meaning, depending on context. These issues become especially relevant in sectors characterized by technological convergence or hybrid business models [5], significantly limiting their effectiveness for the valuation purposes pursued in this paper.

The text-as-data approach is particularly promising for identifying economic exposure. Traditional classifications such as GICS or NAICS, which assign firms to fixed categories, may be useful for reporting and benchmarking, but they are taxonomic frameworks that impose discrete boundaries on economic realities that are often continuous and overlapping. Textual analysis allows researchers to observe firms more flexibly, as instead of asking only which sector a firm officially belongs to, it examines how the firm describes its revenues, assets, risks, customers, and strategic trajectory, thereby opening the possibility of measuring exposure as a gradient rather than as a binary membership.

In finance, this approach can enrich factor modeling, peer identification, event studies, and valuation analysis. Two firms may exhibit similar accounting ratios and yet have very different narrative exposures; alternatively, firms belonging to different official sectors may reveal common dependencies through their language. Text can therefore serve as a bridge between observable financial performance and the underlying economic structure [6], being particularly useful for capturing forward-looking or qualitative dimensions that remain only weakly reflected in traditionally analyzed accounting data.

The urban and real estate domain provides a particularly suitable context for this methodology. Urban systems are shaped by heterogeneous actors whose economic activities are intertwined with location, infrastructure, regulation, and social change. Developers, REITs, construction firms, utilities, mobility platforms, logistics operators, data-center firms, and property services companies all participate in the built environment, but in ways that go beyond conventional sector boundaries. Textual data allows researchers to trace these interdependencies. For example, a firm may formally operate in real estate and yet describe itself using the language of digital infrastructure, last-mile logistics, senior care, or energy transition; these narratives reveal how urban capital is being reorganized both within firms and across the broader economy.

For this reason, the text-as-data approach is not simply a technical complement to real estate research, but becomes a means of representing urban complexity. Through corporate language, we can capture elements such as neighborhood-level demand, tenant composition, urban reconfiguration, capital investment priorities, and exposure to local policies in a way that combines scalability and comparability. This represents a new capability that is particularly relevant when urban real estate is studied as an evolving system rather than as a static asset class. If cities are understood as spaces where capital, regulation, and infrastructure interact, then firm-level texts provide a distributed record of how these interactions are perceived and managed by market participants [2].

5 Semantic Representations of Economic Narratives

The methodological evolution of text analysis in economics and finance has moved from discrete lexical methods toward semantic representations capable of capturing meaning at a higher level of abstraction. Early approaches, such as bag-of-words, term-frequency, or dictionary methods, represent documents through observed word counts. These techniques remain useful for measuring tone, uncertainty, or the salience of predefined topics, but they are limited when the objective is to identify deeper economic similarity across firms. Two companies may operate in closely related activities while using different vocabularies, different reporting styles, or different narrative emphases. For this reason, more recent approaches rely on embeddings, which represent text as dense vectors in a continuous latent space.

Let D_i denote the textual corpus associated with firm i, constructed from regulatory filings, annual reports, investor presentations, earnings-call transcripts, or other corporate disclosures. The text is mapped into a vector representation through an embedding function.

$$e_i = \phi(D_i), e_i \in R^d,$$

where $\phi(\cdot)$ denotes the embedding model and d is the dimension of the semantic space. The resulting vector e_i is a compact numerical representation of the firm's economic narrative. Firms whose texts describe similar business models, risks, markets, or strategic orientations tend to lie closer to one another in this space.

The central advantage of this representation is that similarity is no longer based on exact word overlap, but on semantic proximity. To make this economically interpretable, firms can be compared to a set of reference categories.

Let

$$C_1, C_2, \ldots, C_k$$

denote K category descriptions, where each category represents a sectoral narrative, or a business model, or an economic theme. Embedding these category texts in the same latent space gives

$$C_k = \phi(C_k), k = 1, ..., K.$$

Once both firms and categories are represented in the same vector space, the semantic similarity can be obtained through cosine similarity:

$$s_{ik} = cos(e_i, c_k) = \frac{e_i^T c_k}{||e_i||||c_k||}$$

This measure captures the angle between the two vectors and therefore identifies similarity in direction rather than magnitude. And if the vectors are normalized, this can be reduced to the simple inner product

$$s_{ik} = e_i^T c_k.$$

The quantity s_{ik} measures the degree to which the narrative of firm i aligns with category k. A higher value indicates greater semantic proximity between the firm's published text and the economic content represented by that category. However, these raw similarity values are not yet directly interpretable as exposures; in this form, they indicate proximity but do not provide normalized weights across categories. To transform semantic similarity into economically meaningful exposures, it is necessary to normalize them using a softmax function:

$$p_{ik} = \frac{exp(\gamma s_{ik})}{\sum_{j=1}^{k} exp(\gamma s_{ij})}, k = 1, \ldots, K,$$

where $\gamma > 0$ controls the sharpness of the assignment. The resulting weights satisfy

$$p_{ik} \in (0, 1), \sum_{k=1}^{K} p_{ik} = 1$$

Therefore, they can be interpreted as continuous exposure weights or as soft memberships across categories. This formulation represents a substantial improvement over discrete classifications, as under a one-hot scheme, each firm is assigned to a single category, such that its exposure vector satisfies:

$$p_{ik} \in \{0, 1\}, \sum_{k=1}^{K} p_{ik} = 1$$

Such an approach is often too rigid for firms whose activities span multiple domains. By contrast, the embedding-based framework allows a firm to have partial exposures across several categories simultaneously, which is particularly important in contexts where sectoral boundaries are fluid, where firms combine multiple business functions, or where economic identity evolves.

The parameter γ plays an important role in this setting; higher values make the Soft-Max more concentrated, bringing it closer to a hard classification, while lower values produce more diffuse exposures, preserving ambiguity and overlap across narratives. In that sense, γ controls the degree to which semantic similarity is translated into concentrated or diversified economic exposure.

The full procedure can be summarized in three steps. First, embed the firm texts and category texts:

$$e_i = \phi(D_i), c_k = \phi(C_k).$$

Second, compute semantic similarity:

$$s_{ik} = \frac{e_i^T c_k}{||e_i||||c_k||}$$

Third, normalize the similarity scores into exposure weights:

$$p_{ik} = \frac{exp(\gamma s_{ik})}{\sum_{j=1}^{k} exp(\gamma s_{ij})}$$

Collecting all p_{ik} produces an exposure matrix

$$P \in R^{NxK},$$

whose rows summarize the semantic positioning of firms across economically meaningful categories.

Modern embedding models and large language models strengthen this framework because they produce contextual representations. The meaning of a word depends on the document in which it appears, and this matters greatly in economic and financial language. Terms such as "development", "platform", or "infrastructure" may describe very different economic activities depending on the type of firm. Contextual embeddings reduce this ambiguity by encoding the meaning of phrases and documents rather than isolated tokens. As a result, semantic similarity becomes a more precise measure of the underlying economic narrative, so that semantic representations act as a bridge between corporate text and economic interpretation. Text is transformed into vectors, vectors into similarity measures, and these into continuous exposure weights, resulting in a flexible representation of firms' economic identity that is better suited than rigid taxonomies to environments characterized by heterogeneity, overlap, and structural change.

6 From Semantic Similarity to Economic Exposure

The key analytical step consists of moving from semantic resemblance to an economically interpretable notion of exposure. Similarity in language is not, by itself, a financial variable, but becomes useful when it is linked to the idea that firms describing similar activities, constraints, and opportunities may tend to share common drivers of value. Under this interpretation, semantic similarity acts as a proxy for proximity in economic function, capturing the idea that firms are exposed not only to broad market movements but also to latent thematic forces reflected in their business narratives.

Within this framework, each firm's textual disclosures can be represented through embeddings and compared either with other firms or with a set of thematic prototypes. The resulting similarity scores can then be normalized into continuous exposure weights; rather than assigning a firm exclusively to categories such as "industrial," "retail," or "data center," the model allows for fractional exposure across several economic themes, which is particularly important for diversified, transitional, or hybrid firms. In this way, a company such as a platform operating at the intersection of storage, logistics, and digital infrastructure is not forced into a single category when its valuation depends on multiple overlapping narratives.

The economic meaning of these exposures lies in common sensitivity. If two firms use different words but converge semantically on issues such as energy demand, tenant retention, redevelopment optionality, life-science leasing, or e-commerce penetration, they may respond similarly to shocks affecting these themes. Semantic exposures approximate latent factor loadings rooted in firm narratives and can be interpreted as a map of the conceptual space within which valuation takes place [2].

This perspective also helps explain why textual analysis can matter for prices. From a fundamental standpoint, valuation can be understood as a problem of discounting uncertain future cash flows, where expectations about those flows depend on beliefs

regarding growth opportunities, competitive durability, financing conditions, and operating risks [7]. In this context, investors implicitly process firm narratives by extracting thematic exposures. An embedding-based approach formalizes this process, adding analytical value by transforming narrative proximity into a quantitative structure that can be related to returns, volatility, and cross-sectional differences in performance.

Once semantic exposures have been estimated, they can be used to extract the main factors influencing firm valuation, which do not necessarily coincide with traditional sectors. In the context of the built environment, these may correspond to dimensions such as interest-rate sensitivity, urban logistics intensity, digital infrastructure demand, demographic dependence, proximity to healthcare, or redevelopment optionality. The advantage of this approach is that these factors emerge from the language firms use to describe themselves, rather than being entirely imposed a priori, making the resulting structure potentially more adaptable to economic change.

At a theoretical level, the shift from similarity to exposure reflects a move from classification to measurement, where classification asks what a firm is, while exposure asks which forces it is close to. For asset pricing and urban analysis, the latter question is often more informative. Firms are not static types, but bundles of contingent relationships, and semantic methods provide a way to estimate these relationships between text and connect them to market valuation [8].

7 Economic Exposure and the Measurement of Intrinsic Risk

Once semantic exposures have been estimated, they can be incorporated into a standard multi-factor risk framework. Let $r_t \in R^N$ denote the vector of asset returns at date t. In a Barra-style specification, returns are decomposed into two components: systematic and idiosyncratic:

$$r_t = Xf_t + \varepsilon_t$$

where $X \in R^{N \times K}$ is the exposure matrix, $f_t \in R^K$ is the vector of factor returns, and ε_t is the vector of specific returns. In this context, semantic exposures extracted from text replace or complement traditional sector dummy variables, allowing firms to load continuously on multiple economic factors rather than being forced into a single category.

Factor returns are estimated cross-sectionally, while total risk is decomposed into two components: a systematic component driven by common factors and a firm-specific component. If Σ_F denotes the factor covariance matrix and β_i the vector of factor loadings for asset i, factor-driven variance is

$$\sigma_{F,i}^2 = \beta_i^\top \Sigma_F \beta_i$$

Specific variance is estimated from residual returns,

$$\sigma_{\beta,i}^2 = Var\left(\varepsilon_{i,t}\right),$$

so total variance becomes

$$\sigma_i^2 = \sigma_{F,i}^2 + \sigma_{\varepsilon,i}^2$$

This decomposition is central to valuation and risk control because it distinguishes between risk explained by shared economic structure and risk that remains intrinsic to the firm [8].

From a financial perspective, semantic exposure matters because structural changes in a firm's narrative may anticipate changes in its true economic risk. If a firm becomes more exposed to debt-sensitive activities, cyclical tenants, speculative development, or capital-intensive expansion, this shift may later be reflected in its realized betas, in the covariance structure of factor returns, and ultimately in its intrinsic risk [9]. Text thus provides a forward-looking signal of changing vulnerability. Rather than treating beta as a fixed parameter, the framework allows it to be understood as partly driven by evolving economic exposure [2], which may ultimately translate into changes in firm valuation.

This is particularly relevant in real estate. The sector is highly exposed to structural risks that are not always immediately reflected in accounting data. An increase in interest rates can compress asset values, leading to higher refinancing costs and translating into weaker transaction activity, thereby reducing the relative attractiveness of income-producing assets. A lack of liquidity can further amplify this process; for instance, if capital markets tighten or asset sales become more difficult, firms with short debt duration, aggressive external growth models, or fragile tenant bases may experience a significant increase in both systematic and idiosyncratic risk. In this sense, structural risk in real estate is not limited to market risk, but also emerges in the form of financing risk, duration risk, rollover risk, and valuation risk concentrated in specific business models.

A semantic risk model can help identify these vulnerabilities by grouping firms according to their underlying economic similarity rather than relying solely on formal sector classifications. Two firms may both be classified as real estate companies and yet exhibit very different sensitivities to interest rate shocks, liquidity stress, lease duration, technological obsolescence, or cyclical demand, while firms in different formal subsectors may share common structural exposures. A continuous exposure framework is therefore better suited to capturing the true propagation of risk within the built environment.

This is also relevant for robustness to tail risk. Volatility alone is an incomplete proxy for risk, as it does not capture the asymmetry and concentration of losses that emerge during stress episodes. In particular, higher-order risks associated with skewness, kurtosis, and non-normal return distributions may become critical in real estate markets exposed to interest rate repricing or liquidity shocks. By contrast, a multi-factor model does not eliminate tail risk, but it provides a more robust framework by isolating systematic channels of exposure before analyzing residual extreme risk separately.

In summary, the measurement of intrinsic risk improves when economic exposure is modeled as continuous, dynamic, and text-informed. In the real estate sector, whose capital structure, valuation volatility of underlying assets, financing conditions, and market liquidity play a decisive role in firm outcome, this approach offers a more realistic way to understand how structural risk enters valuation and how vulnerabilities can accumulate before being fully reflected in prices.

8 Urban Systems and Capital Allocation in the Built Environment

The broader significance of this framework becomes particularly clear when firms are understood as sensors of urban systems and cities are not viewed as mere physical agglomerations, but as dense configurations of capital, labor, infrastructure, regulation, and social demand. Firms operating within the built environment continuously register changes in these configurations; through leasing patterns, development pipelines, financing strategies, tenant feedback, and investment plans, they provide distributed signals about the evolving structure of urban economies.

From this perspective, the study of firms contributes directly to understanding urban development. Corporate narratives reveal where demand is intensifying, where infrastructure constraints are emerging, which types of space are becoming obsolete, and which asset classes are attracting strategic investment. The language used by firms can therefore be interpreted as a sensor network for the city, where an increase in narratives around themes such as cold storage, last-mile distribution, laboratory conversion, power availability, or demand for senior housing may indicate deeper transformations in the allocation of urban capital [2].

This matters because capital does not flow neutrally across space, but instead moves toward sectors, locations, and business models that appear capable of absorbing and reproducing value. When firms change, the capitalization of the sector changes with them; for example, a wave of firms repositioning around digital infrastructure can redirect investment toward data centers, telecom assets, energy interconnection, and related land markets, while a re-narration of logistics as strategic urban infrastructure can attract capital to warehouse corridors, intermodal nodes, and peripheral development areas. Similarly, if residential firms increasingly emphasize affordability, mobility access, or demographic resilience, these themes may shape both investor attention and interactions with urban policy.

The built environment thus emerges as a space where semantic interpretation and material allocation intersect, such that firm narratives influence valuation, valuation in turn affects the cost of capital, and the latter determines which urban projects become viable. This creates a feedback loop linking textual representation to physical transformation. Consequently, studying firms through their published texts is not only a financial exercise but also a way of analyzing how urban systems evolve.

For research in urban economics and real estate, this suggests a fruitful integration of market data, textual analysis, and spatial interpretation. Firms can be analyzed as intermediaries between macroeconomic forces and urban form, with their disclosures revealing how abstract drivers such as interest rates, digitalization, aging, or decarbonization materialize in buildings, land uses, and metropolitan investment patterns. A methodology capable of capturing these signals can help explain why some urban sectors attract persistent capital while others undergo repricing or decline.

This is the context in which textual analysis of firms offers more than an improvement in classification tools; it provides a way to observe the co-evolution of language, capital allocation, and urban transformation by treating firms as interpretive and sensing agents within the built environment. It thus becomes a tool that allows researchers to better understand how economic narratives are translated into a geography of investment.

9 Conclusions

This paper proposes a methodological model for transforming corporate narratives into measurable economic exposures and using those exposures to assess risk in urban real estate firms. The central argument is that firms operating in the built environment are not only holders of physical assets or members of fixed industrial categories; they are also interpreters of economic change. Their disclosures contain information about financing conditions, tenant demand, infrastructure constraints, redevelopment strategies, regulatory pressures, and spatial transformation. When systematically reviewed, these accounts can help identify shared sensitivities that are not fully captured by traditional sector classifications.

The paper's contribution is methodological rather than purely conceptual. It delivers a standardized sequence of steps for moving from text into risk assessment: corpus construction, text cleaning, embedding, semantic comparison, exposure normalization, acuity calibration, financial-noise filtering, factor-return estimation, and decomposition of total risk into systematic and intrinsic components. This workflow responds to the need for a replicable procedure that can be applied by other researchers and practitioners. It also clarifies the operational meaning of the proposed approach: corporate text is not treated as anecdotal evidence, but as an empirical input that can be converted into an exposure matrix and tested within a multi-factor risk model.

The framework contributes to three areas of research. First, it extends text-as-data methods by focusing not only on sentiment, uncertainty, or topic frequency, but on the construction of continuous economic exposures. Second, it contributes to real estate and urban studies by providing a way to connect firm narratives linked with structural transformations in the built environment. Third, it contributes to financial risk modeling by replacing or complementing discrete industry classifications with semantic exposure vectors that enable firms to load partially onto multiple economic narratives simultaneously.

The most important implication is that intrinsic risk can be defined more precisely once common semantic exposures have been controlled for. A firm may appear risky because it belongs to a volatile sector, is exposed to a common structural driver, or contains genuinely idiosyncratic vulnerabilities. The proposed model helps distinguish these cases. For example, two real estate firms may both be formally classified within the same broad sector but may differ sharply in exposure to refinancing conditions, logistics demand, healthcare tenants, energy infrastructure, data-center growth, or redevelopment optionality. Conversely, firms assigned to different official categories may share exposure to the same structural urban transformation. A semantic exposure model can reveal these overlaps, thereby improving risk attribution.

The paper additionally emphasizes that methodological transparency is essential. The embedding model must be specified, the reference categories must be documented, the acuity parameter must be calibrated rather than chosen arbitrarily, and the filtering procedure must be described in sufficient detail to permit replication. In this respect, the proposed framework should be understood as an empirical protocol. Its usefulness depends on whether other researchers can reproduce the exposure matrix, evaluate alternative category definitions, test different embedding models, and compare the resulting risk decomposition against conventional benchmarks.

Subsequent research should develop the empirical side of the framework in four directions. First, the model should be tested on larger international samples of listed real estate firms, REITs, infrastructure firms, developers, and urban service providers. Second, semantic exposures should be compared with traditional sector classifications to assess whether they improve explained variance, specific risk estimation, and portfolio risk attribution. Third, the acuity parameter should be calibrated through formal criteria, including exposure entropy, stability, and out-of-sample performance. Fourth, the model should be extended spatially by linking firm-level narratives to geographies, asset locations, and urban indicators such as density, transport accessibility, demographic change, climate exposure, and infrastructure investment.

In conclusion, the paper argues that corporate text can function as a bridge between qualitative economic interpretation and quantitative financial risk assessment. In urban real estate, where value is formed by long-lived assets, spatial restrictions, capital-market conditions, and structural transformation, this bridge is especially valuable. The proposed framework does not replace accounting data, market prices, or traditional risk models. Rather, it complements them by making firm narratives measurable and enabling them to enter a transparent risk-decomposition framework. This creates a more adaptable approach to measuring economic exposure and intrinsic risk in the built environment.

References

1. Hayek, F.A.: The use of knowledge in society. Am. Econ. Rev. **35**(4), 519–530 (1945)
2. Hassan, T.A., Hollander, S., Kalyani, A., van Lent, L., Schwedeler, M., Tahoun, A.: Text as data in economic analysis. J. Econ. Perspect. **39**(3), 193–220 (2025)
3. Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. J. Financ. **66**(1), 35–65 (2011)
4. Gentzkow, M., Kelly, B., Taddy, M.: Text as data. J. Econ. Literat. **57**(3), 535–574 (2019)
5. Baker, S.R., Bloom, N., Davis, S.J.: Measuring economic policy uncertainty. Q. J. Econ. **131**(4), 1593–1636 (2016)
6. Tetlock, P.C.: Giving content to investor sentiment. J. Financ. **62**(3), 1139–1168 (2007)
7. Fisher, I.: The Theory of Interest. Macmillan (1930)
8. Markowitz, H.: Portfolio selection. J. Financ. **7**(1), 77–91 (1952)
9. Hennessy, C.A., Strebulaev, I.A.: Policy risk and corporate investment. Rev. Financ. Stud. (2020)
10. Hoberg, G., Phillips, G.: Text-based network industries and endogenous product differentiation. J. Polit. Econ. **124**(5), 1423–1465 (2016)
11. Hoberg, G., Phillips, G.: Product market synergies and competition in mergers and acquisitions: a text-based analysis. Rev. Financ. Stud. **23**(10), 3773–3811 (2010)
12. Bybee, L., Kelly, B., Manela, A., Xiu, D.: Business news and business cycles. J. Financ. **79**(5), 3105–3147 (2024)
13. Fama, E.F., French, K.R.: Common risk factors in the returns on stocks and bonds. J. Financ. Econ. **33**(1), 3–56 (1993)
14. Kelly, B.T., Pruitt, S., Su, Y.: Characteristics are covariances: a unified model of risk and return. J. Financ. Econ. **134**(3), 501–524 (2019)
15. Guarini, M.R., Sica, F., Segura, A.: Artificial intelligence (AI) integration in urban decision-making processes: convergence and divergence with the multi-criteria analysis (MCA). Information **15**(11), 678 (2024)

16. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT 2019, pp. 4171–4186 (2019)
18. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-networks. In: Proceedings of EMNLP-IJCNLP 2019, pp. 3982–3992 (2019)
19. Guarini, M.R., Roma, A., Sabatelli, E., Segura, A.: Intrinsic and extrinsic attributes in real estate pricing: insights for sustainable urban planning strategies. Land Use Policy **153**, 107543 (2025)