The Journal of Finance and Data Science 12 (2026) 100199

Contents lists available at ScienceDirect

The Journal of Finance and Data Science

journal homepage: www.keaipublishing.com/en/journals/jfds



Research Article

Semantic drift in long-form financial disclosures in Portuguese

João Victor M. Correa [a], Jorge Miguel Bravo [b,c], Rodrigo Lanna Franco da Silveira [d], José César Cruz Júnior [e], Fernando Batista [f,g], Renato Moraes Silva [a,*]

[a] Interinstitutional Center for Computational Linguistics (NILC), Institute of Mathematics and Computer Science, University of São Paulo (USP), São Carlos, SP, Brazil
[b] NOVA IMS Universidade Nova de Lisboa, Lisbon, Portugal
[c] Université Paris-Dauphine PSL, Paris, France
[d] Institute of Economics, Center for Applied Economics, Agricultural and Environmental Research (CEA), University of Campinas, Campinas, Brazil
[e] CIH Trading LLC d/b/a Advance Trading Inc., Bloomington, IL, USA
[f] INESC-ID Lisboa, Rua Alves Redol, 9, 1000-029 Lisboa, Portugal
[g] Iscte - Instituto Universitário de Lisboa, Av. Forças Armadas, 1649-026 Lisboa, Portugal

ARTICLE INFO

Keywords:
Semantic drift
Corporate disclosures
Document embeddings
Financial NLP
Natural language processing
Portuguese language

ABSTRACT

Financial disclosures contain critical information that is not always immediately reflected in market prices. Tracking semantic change in these communications can surface narrative shifts, but it is difficult because financial documents are long, templated, and evolve gradually, which blurs meaningful drift with routine variation. Moreover, Portuguese-language financial disclosures remain underrepresented in the financial natural language processing literature. We propose a framework for measuring semantic drift in long-form Portuguese disclosures using document embeddings, chunking, and aggregation, with drift defined as consecutive cosine distance. Article-scope inference uses 3 models and 5 aggregation techniques (15 configurations per dataset) across five datasets spanning corporate and public-sector reporting. We validate drift against volatility signals using circular-shift tests, filing-date time windows, and Granger causality, complemented by descriptive event alignment and event-study diagnostics. We find that Vale exhibits a positive drift–volatility association under within-model correction, EDP shows robust drift-to-volatility Granger predictability across configurations, and SLC shows a robust filing-date window association. Conab exhibits a similar Granger pattern only when mapped to SLC stock as market proxy, making that result proxy-sensitive. Other model–dataset combinations show weaker or non-significant links, highlighting sensitivity to document type and template structure.

1. Introduction

On January 25, 2019, a tailings dam collapsed at the Brumadinho mine operated by Vale S.A. in Brazil. The disaster killed 270 people and became the deadliest mining accident in Brazilian history. In the company's quarterly financial reports that followed, Vale's language changed. The word “dam” appeared more frequently, surrounded by new context about safety protocols and remediation

* Corresponding author.
E-mail addresses: victormcorrea@usp.br (J.V.M. Correa), jbravo@novaims.unl.pt (J.M. Bravo), rlanna@unicamp.br (R.L.F. Silveira), ccruz@advance-trading.com (J.C. Cruz Júnior), fernando.batista@iscte-iul.pt (F. Batista), renatoms@icmc.usp.br (R.M. Silva).

Peer review under the responsibility of KeAi Communications Co., Ltd.

https://doi.org/10.1016/j.jfds.2026.100199
Received 5 March 2026; Received in revised form 7 June 2026; Accepted 26 June 2026
Available online 2 July 2026
2405-9188/© 2026 The Authors. Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

costs. Terms like “reparation” and “community” entered sections where they had not appeared before. This is semantic drift, a measurable shift in how a company talks about its operations over time.

These shifts matter because corporate language is not decorative. In capital markets, information disseminated through channels such as earnings reports and conference calls is the foundation of investment decisions, shaping perceptions of corporate value, performance, and market expectations (Maia and Bravo, 2026). Decades of research show that word choice in disclosures relates to firm outcomes, stock returns, and risk (Loughran and McDonald, 2011). In particular, Cohen et al. (2020) find that firms which barely revise the text of their filings from one period to the next earn higher subsequent returns than firms that rewrite heavily, a gap they attribute to investor inattention to qualitative changes. This makes the *rate* of disclosure-language change, not only its sentiment, an economically meaningful signal, providing a direct motivation for measuring semantic drift.

Most prior work focuses on sentiment scoring or counting words from predefined dictionaries. We track something different, namely how the overall meaning of documents changes from one period to the next. Our method uses document embeddings, a technique from natural language processing that represents text as numerical vectors in a high-dimensional space, where similar documents sit closer together (Le and Mikolov, 2014; Dai et al., 2015). Measuring the distance between consecutive document vectors lets us quantify how much a company's disclosure language has shifted.

The idea connects two research traditions. In computational linguistics, diachronic analysis tracks how word meanings evolve across decades (Hamilton et al., 2016; Kiyama et al., 2025). In machine learning, concept drift refers to changes in the statistical properties of data over time (Gama et al., 2014).

Detecting such drifts in text is particularly challenging due to the scarcity of annotated datasets covering extended periods, especially for underrepresented languages like Portuguese. A recent study proposed by Wanderley et al. (2025) addressed this by using a Portuguese-language fake news classification dataset centered around a high-impact event, the COVID-19 pandemic, to validate drift in short-form content. However, the study is limited by short timeframes and brief text lengths. The authors analyzed news within a two-year window (January 2020 to December 2021) with short average document lengths. To the best of our knowledge, no study has investigated semantic drift in long-form Portuguese documents over extended time spans, a setting that is essential for capturing gradual and complex narrative evolutions. The difficulty of analyzing such data is reflected in the broader financial NLP literature; for example, a recent comprehensive survey by Du et al. (2025) covers a wide array of applications but does not report studies within the financial domain specifically focused on text-based concept drift detection.

Financial disclosures provide an opportunity to address this gap, as they are among the few sources of high-quality textual data available with consistent periodicity over many years. Although these documents are unlabeled, market data provides a natural validation signal. Material events such as operational crises, regulatory changes, or commodity shocks tend to coincide with both revised disclosure language and elevated investor uncertainty. If our drift measure captures genuine content shifts, high-drift periods should co-move with higher return volatility, not because language drives price movements, but because both tend to respond to the same underlying changes. Known historical events offer a second, independent validation channel.

We describe a method for measuring semantic drift in long-form financial disclosures. Document-level inference is conducted using three models and five aggregation techniques, resulting in fifteen configurations per dataset. We validate against external signals including correlation with realized volatility under circular-shift testing, filing-date anchored time-window designs, Granger predictability tests, descriptive alignment with known events, and sensitivity to controlled out-of-domain injections.

This work is relevant to financial applications because it provides a systematic framework to track how disclosure language changes over time and to prioritize reports and periods for closer review, supporting financial market regulators in monitoring and enforcement. Our framework identifies narrative shifts that may reflect changes in corporate positioning, risk exposure, or other economically relevant dimensions. We interpret our validations as evidence about association and predictability structure, not as a claim of trading profitability.

From a data science perspective, this study contributes by investigating semantic analysis methods on real-world financial data, which is characterized by its high dimensionality and evolution over time. By exploring reports in Portuguese, a language still underrepresented in the field, the study broadens the empirical scope of the area and allows for the evaluation of the generalization, robustness, and sensitivity of semantic bias detection techniques under different data distributions.

In the field of machine learning, concept drift detection is fundamental for monitoring changes in data patterns over time, especially in dynamic scenarios such as finance. The proposed method can support analysis and forecasting systems by identifying structural changes that affect the stability of models trained on historical data.

In the context of natural language processing, this study addresses a still largely unexplored challenge: the detection of concept drift in Portuguese long documents. Many vector representation techniques assume limited context windows, which hinders the modeling of semantic changes distributed across extensive texts and time series of documents. By investigating aggregation strategies and vector representations suitable for long reports, this study contributes to the advancement of methods capable of capturing global and subtle semantic variations over time.

The main contributions of this study are summarized as follows.

- A computational pipeline for detecting concept drift in Portuguese corporate financial disclosures;
- An investigation of chunking and aggregation techniques for processing long-form documents aimed at preserving semantic signals across extensive financial reports;
- A comparative analysis of three embedding models (Qwen3-Embedding-8B, BERTimbau, and EuroBERT) and their impact on the sensitivity of semantic drift detection within a financial domain context;

- An extensive assessment of drift signals through complementary validation strategies, including volatility correlation with circular-shift testing and FDR control, filing-date anchored time-window designs, Granger causality tests, descriptive alignment with historical events, and robustness tests via shuffled ordering and out-of-domain injection.

The remainder of this paper is structured as follows. Section 2 presents the main related work in the field. Section 3 presents our experimental setup. Section 4 presents the results, while Section 5 discusses our results and findings. Finally, Section 6 concludes the paper.

2. Related work

Our study draws on financial text analysis, diachronic semantics, and concept drift detection.

2.1. Financial text analysis

Analysis of corporate disclosures has a long history in finance and accounting. Loughran McDonald (2011) showed that standard sentiment dictionaries, such as the Harvard General Inquirer, perform poorly on financial text because words like “liability” and “risk” have different connotations in finance than in everyday usage. They developed domain-specific word lists that better predict stock returns and abnormal trading volume following 10-K filings. Later surveys document that the field still leans heavily on cross-sectional sentiment, readability, and lexicon-based designs, with limited emphasis on temporal semantic change (Loughran and McDonald, 2020; Ravula, 2021; Du et al., 2025).

Studies more closely aligned with our approach use distributional representations of filings, topics, and embeddings. Adosoglou et al. (2021) applied Doc2Vec embeddings to measure year-over-year semantic changes in 10-K filings. Fengler and Phan (2024) tracks disclosure themes over time via embedding-informed topic construction. We extend this line of inquiry to Portuguese-language disclosures using modern transformer-based embeddings at finer temporal granularity, and we emphasize statistically disciplined inference for time series validation.

A second strand of research ties disclosure text directly to market outcomes. Tetlock (2007) shows that media pessimism predicts downward price pressure followed by reversal, providing evidence that qualitative text can be a market-moving signal. Brown and Tucker (2011) document that firms modify their Management's Discussion and Analysis (MD&A) sections from year to year and that the extent of these modifications is itself informative, foreshadowing later work on textual change. Building on this line of research, Cohen et al. (2020) show that the degree of textual revision constitutes a tradable signal. Comparing each filing to the same firm's prior-year filing of the same type, they find that “non-changers” outperform “changers” by up to 188 basis points per month, a gap they attribute to investor inattention to qualitative edits. They quantify change using several bag-of-words measures, including cosine similarity over raw term-frequency vectors and the Jaccard index over the set of terms, with the latter performing slightly better in forecasting return. Either way, lower similarity corresponds to greater semantic change, so their low-similarity “changers” align with what we later define as high-drift periods. We build on this tradition but differ on three axes. First, on language, we study Brazilian and European Portuguese rather than US English. Second, on representation, our primary signal comes from transformer embeddings rather than bag-of-words term counts, and we directly benchmark the two approaches (Table 7). Third, on scope, we track consecutive monthly and quarterly drift across five datasets in distinct sectors, whereas Cohen et al. (2020) compare each filing to its seasonal counterpart a year earlier to cancel filing seasonality. That design difference matters: their year-over-year comparison removes recurring seasonal patterns in standardized reporting templates, whereas our consecutive design retains these effects.

Most financial text research, including cross-sectional comparisons and textual analysis methods, such as text classification, sentiment scoring, and named entity recognition, treats documents as independent observations (e.g., Holowczak et al., 2019; Rybinski, 2020; Duarte et al., 2021; de Oliveira Carosia et al., 2021; Zerbinati et al., 2024; Huang and Shum Nolan, 2025). In contrast, we track *temporal* change within a single company's disclosures, measuring how language evolves rather than what it says at any given moment. This longitudinal view is natural for investors monitoring the same firm over multiple reporting periods.

2.2. Diachronic semantics

Computational linguistics has developed methods for tracking how word meanings change over time. Hamilton et al. (2016) found that word embeddings trained on historical corpora reveal statistical laws of semantic change. The word “gay,” for instance, shifted from meaning “happy” to its modern usage over the 20th century, and this shift is visible in the changing vector representations. Kulkarni et al. (2015) introduced methods for detecting when such changes are statistically significant rather than noise.

Benchmark and methodology work in lexical semantic change formalizes evaluation and highlights failure modes of contextual embeddings (Schlechtweg et al., 2020; Montariol et al., 2021; Kutuzov et al., 2022; Periti and Tahmasebi, 2024). Continuous formulations also appear in this literature, which aligns with our use of a drift time series rather than a single before/after comparison (Kiyama et al., 2025). Most importantly for our validation layer, Liu et al. (2021) combines contextual embeddings with permutation tests and Benjamini-Hochberg (BH) false discovery rate (FDR) control to obtain statistically reliable semantic shift detection under multiple comparisons. We adapt this same logic to document-level drift validation, where the families of simultaneous tests arise from aggregations, lags, and metrics.

2.3. Concept drift in machine learning

In machine learning, concept drift refers to changes in the statistical relationship between inputs and outputs over time (Gama et al., 2014). When such changes occur, models trained on historical data may become misaligned with current patterns. For example, a spam filter trained in 2010 may perform poorly in 2025 because spammer tactics and linguistic strategies have evolved. In textual applications, this phenomenon often manifests as evolving linguistic patterns that alter the relationship between textual features and prediction targets.

Early empirical evidence of temporal concept drift in NLP was provided by Batista et al. (2008), who showed that models trained on older corpora perform worse when applied to more recent data. Using maximum entropy models, they demonstrated that performance declines as the temporal gap between training and testing periods increases, highlighting the need for time-aware adaptation. Extending this evidence, Batista et al. (2012) confirmed that both capitalization and punctuation recovery are systematically affected by temporal variation, with performance decreasing as linguistic patterns evolve. Together, these studies provide early systematic evidence that language dynamics can materially affect model performance. This insight is particularly relevant in financial settings, where disclosure language evolves across regulatory cycles, macroeconomic regimes, and firm-specific events.

Beyond documenting performance degradation, the machine learning community has developed explicit mechanisms to detect distributional shifts. Hidalgo et al. (2019) proposed using cosine similarity between feature distributions to detect drift, while Feldhans et al. (2021) extended this idea to document embeddings. They showed that multivariate drift detectors applied to embedding spaces can capture distributional shifts that simpler univariate approaches miss.

Recent surveys and systems emphasize representation-level drift detection and adaptation in text streams (Garcia et al., 2025; Greco et al., 2025). While that literature focuses on maintaining predictive performance in streaming settings, our goal is different. We treat drift as the object of interest and validate it against external finance signals.

However, concept drift has been studied mainly in financial applications using price data or other numerical features. For instance, Panicachi Cocovilo Filho and Coelho (2024) evaluated machine learning-based predictors that incorporate mechanisms to handle concept drift for stock market forecasting. Their study utilized a historical database of stock prices from ten companies traded on the Brazilian stock exchange, collected over a 20-year period. The authors concluded that algorithms equipped with mechanisms that handle concept drift tend to achieve superior results. Specifically, the EOS-D approach, a variation of the ensemble of online learners algorithm that incorporates dynamic adaptations, obtained the best performance.

Our work applies these ideas to financial disclosures but with a different objective. Rather than maintaining model performance, we aim to extract economically meaningful signals from detected drift.

2.4. Text signals and market dynamics

Text to market linkage work often tests temporal structure with Granger-style designs or information-theoretic alternatives (Tabari et al., 2018; Scaramozzino et al., 2021; Liang et al., 2024; Drinkall et al., 2025). Related work also frames semantic shift as a stability issue for market prediction pipelines (Sun et al., 2025). We follow this direction only as validation. We use time-series tests to assess whether drift contains market-relevant structure, without treating the task as a return forecasting benchmark.

2.5. Document embeddings for long documents

Representing long documents as fixed-dimensional vectors is difficult. Transformer models have context limits (typically 512 tokens, roughly one page), so longer documents must be split into chunks and recombined. Tang and Yang (2024) compared pooling strategies for large language model embeddings and found that mean pooling often matches more complex methods; Lin et al. (2023) introduced learned aggregation for dense retrieval.

We build on this literature by evaluating aggregation strategies for measuring semantic change over time, where stability of document representations across periods matters most. Unlike retrieval tasks, our application requires that small changes in document content produce correspondingly small changes in the embedding, without artifacts from the aggregation process.

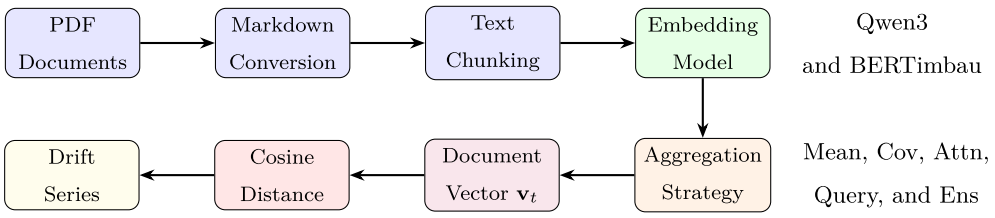


Fig. 1. Semantic drift measurement pipeline. PDF documents are converted to markdown, split into overlapping chunks, embedded using one of three models, aggregated into document vectors, and compared via cosine distance to produce a drift time series.

3. Methodology

To capture the evolution of narrative shifts in long-form financial disclosures, we developed a multi-stage semantic drift measurement framework. The pipeline is designed to overcome inherent limitations of embedding techniques when processing long texts while ensuring that the resulting document representations are robust enough to distinguish meaningful content changes from routine reporting variations. By systematically processing raw PDF documents through to the generation of drift time series, the framework allows for a comparative analysis of different embedding and aggregation techniques. Fig. 1 illustrates our semantic drift measurement pipeline. Unlike prior work that detects temporal drift indirectly through performance degradation (e.g., Batista et al., 2008), we directly measure semantic displacement in embedding space.

3.1. Problem formulation

Let $\mathscr{R} = \{r_1, r_2, \ldots, r_T\}$ be a time-ordered sequence of reports from the same source. Each r_t is one periodic filing (monthly or quarterly). After preprocessing and chunking, report r_t is represented by n_t chunks $\{x_{t,1}, \ldots, x_{t,n_t}\}$. The embedding model maps each chunk to $\mathbf{c}_{t,j} \in \mathbb{R}^p$, where p is the embedding dimension (typically ranging from 1024 to 4096 depending on the model).

An aggregation function $g(\cdot)$ maps chunk embeddings to one report-level vector:

$$\mathbf{v}_t = g(\mathbf{c}_{t,1}, \ldots, \mathbf{c}_{t,n_t}) \in \mathbb{R}^{p_g}. \tag{1}$$

where the output dimension depends on the aggregation strategy: $p_g = p$ for mean, attention-over-chunks, query attention, and ensemble, and $p_g = 2p$ for mean-covariance.

We measure drift at time t as the cosine distance between consecutive report vectors (Hidalgo et al., 2019):

$$\text{drift}_t = 1 - \cos(\mathbf{v}_t, \mathbf{v}_{t-1}) = 1 - \frac{\mathbf{v}_t \cdot \mathbf{v}_{t-1}}{\|\mathbf{v}_t\| \|\mathbf{v}_{t-1}\|}, \quad t = 2, \ldots, T \tag{2}$$

Cosine distance measures the angle between two vectors, ignoring their magnitudes. A value of 0 indicates identical semantic content (the vectors point in the same direction), while a value of 2 indicates completely opposite content. This metric is standard in information retrieval and text similarity applications (Salton and Buckley, 1988; Thongtan and Phienthrakul, 2019), and has been adapted for drift detection in data streams (Hidalgo et al., 2019).

This formulation builds on the framework for temporal semantic change from Hamilton et al. (2016) and the concept drift literature (Gama et al., 2014). Document embeddings let us represent semantic content as vectors (Le and Mikolov, 2014; Dai et al., 2015), and cosine distance has been validated for measuring document-level semantic differences (Qurashi et al., 2020; Zhang and Arefyev, 2022).

In our corpora (see Section 3.5), median consecutive drift ranges from 0.002 (best configuration for Itaú) to 0.034 (Vale), with 95th-percentile values between 0.01 and 0.25 depending on the dataset and model. Fig. 3 gives a concrete example of this signal for Vale (Qwen mean pooling), including the threshold used in the descriptive event-based view. The resulting time series $\{\text{drift}_2, \ldots, \text{drift}_T\}$ captures how much semantic content changed from one period to the next.

We also computed a baseline drift measure, defined as the cosine distance between the current report and the mean of the previous w reports (a rolling window approach). In our exploratory empirical analysis, consecutive drift produced sharper spikes at known event dates than the rolling baseline, which smooths abrupt changes by design. Because our primary interest is in detecting period-to-period shifts rather than long-run trends, we adopted consecutive drift as the default measure for all subsequent analyses. This choice is a design decision, not a formally benchmarked result, as there is no single standard or closed-form drift formulation universally adopted in the literature.

3.2. Document embeddings

The pipeline transforms raw PDF files into document-level embedding vectors through four stages (conversion, chunking, embedding, and aggregation).

An embedding is a learned numerical representation of text. Embedding models are trained on large corpora to map words, sentences, or documents into vectors such that semantically similar texts have similar vector representations (Mikolov et al., 2013; Reimers and Gurevych, 2019). Documents discussing “credit risk” and “loan defaults” would have similar embeddings; a document about “mining operations” would have a different one. Semantic similarity can then be measured with vector operations.

Since financial disclosures often span 50–200 pages, they far exceed the input limits of embedding models, which are typically 512 tokens. To address this, we split documents into overlapping chunks, embed each chunk separately, and aggregate the chunk embeddings into a single document vector. This is standard practice for long documents (Tang and Yang, 2024).

Raw PDFs are first converted to markdown using PyMuPDF4LLM,[1] which preserves document structure while stripping headers, footers, and page numbers. The cleaned text is then split into overlapping chunks using each embedding model's native tokenizer. We use 506 tokens per chunk with 96 tokens of overlap between consecutive chunks. Chunks shorter than 40 tokens are discarded, since such short tails are typically headers, footers, or partial sentences that carry little semantic content while exerting disproportionate

[1] https://github.com/pymupdf/RAG.

influence through padding and special tokens. The 40-token threshold was set empirically, and behaves consistently across all five datasets. Moreover, the 96-token overlap preserves surrounding context so no passage is irrecoverably lost. The 506-token budget was chosen to ensure parity with BERTimbau's 512-token limit, leaving a small buffer for tokenizer variation during decode–encode roundtrips. The same budget is applied to Qwen and EuroBERT, despite their much larger context windows (32,768 and 8192 tokens) to maintain a common chunking strategy across models. Using model-native tokenization ensures that chunk boundaries align with the model's vocabulary, avoiding mid-subword splits that could distort semantic content. All three models used in this study are HuggingFace-compatible, so tokenization uses AutoTokenizer in all cases. Chunk-size verification is logged during processing to confirm that chunks stay within the model token limit.

Each chunk is embedded independently using one of three models: Qwen3-Embedding-8B (Zhang et al., 2025), BERTimbau (Souza et al., 2020), and EuroBERT (Boizard et al., 2025). We selected these models to ensure robustness across different context windows and architectural scales. Qwen3-Embedding-8B (Zhang et al., 2025) is a large multilingual model producing 4096-dimensional vectors, accessed through the OpenRouter API. It ranks highly on the MTEB benchmark (Muennighoff et al., 2023) and supports Portuguese. BERTimbau (Souza et al., 2020) is a BERT model pretrained on Brazilian Portuguese text. We use the large-cased variant fine-tuned for semantic textual similarity, which produces 1024-dimensional vectors. This fine-tuning optimizes cosine similarity comparisons and captures Portuguese-specific patterns that multilingual models may miss. EuroBERT (Boizard et al., 2025) provides a compact multilingual alternative, consisting of a 210M-parameter encoder that produces 768-dimensional, L2-normalized embeddings under mean pooling, with a maximum sequence length of 8192 tokens. It adopts a LLaMA-style architecture (Touvron et al., 2023), incorporating rotary position embeddings, grouped-query attention, SwiGLU activations, RMSNorm, and FlashAttention. It is trained on European languages, including Portuguese, as well as code and mathematics. As the smallest of the three models, it serves as a useful test of whether a lightweight encoder can capture the same drift signal as substantially larger models.

3.3. Aggregation strategies

Converting chunk-level embeddings to a single document vector requires aggregation. A 100-page financial report may produce more than 200 chunks, each with its own embedding vector. These must be combined into one vector representing the document. The aggregation strategy affects which parts of the document influence the final representation most.

We evaluate five strategies, from simple averaging to attention-weighted combinations. Attention mechanisms assign different weights to chunks based on relevance, letting the aggregation focus on more informative parts of the document (Vaswani et al., 2017).

The simplest approach, **mean pooling**, averages all chunk embeddings:

$$\mathbf{v}_{\text{mean},t} = \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{c}_{t,i} \tag{3}$$

where $\mathbf{c}_{t,i}$ is the embedding of chunk i in report t, and n_t is the number of chunks in report t. Mean pooling often performs comparably to more complex methods (Tang and Yang, 2024).

Mean-covariance pooling concatenates the mean and per-dimension standard deviation of all chunk embeddings, capturing both first- and second-order statistics. The approach follows the statistics pooling strategy introduced for speaker verification (Snyder et al., 2018), which we adapt here for document-level aggregation:

$$\begin{aligned}\mathbf{v}_{\text{cov},t} &= [\boldsymbol{\mu}_t\,;\,\boldsymbol{\sigma}_t],\\ \boldsymbol{\mu}_t &= \frac{1}{n_t}\sum_{i=1}^{n_t}\mathbf{c}_{t,i},\\ \sigma_{t,k} &= \sqrt{\frac{1}{n_t}\sum_{i=1}^{n_t}(c_{t,i,k}-\mu_{t,k})^2+\epsilon}\end{aligned} \tag{4}$$

where $[\cdot;\cdot]$ denotes concatenation along the feature axis, $k = 1, \ldots, p$ indexes embedding dimensions, and ϵ is a small constant for numerical stability. The output has dimensionality $2p$, preserving distributional information about chunk-level variation that mean pooling discards.

For **attention over chunks**, we compute attention weights using the similarity between each chunk and the document mean:

$$a_{t,i} = \frac{\exp(\mathbf{c}_{t,i}^{\top}\boldsymbol{\mu}_t/\tau)}{\sum_{j=1}^{n_t}\exp(\mathbf{c}_{t,j}^{\top}\boldsymbol{\mu}_t/\tau)}, \qquad \mathbf{v}_{\text{attn},t} = \sum_{i=1}^{n_t}a_{t,i}\mathbf{c}_{t,i} \tag{5}$$

where τ is a temperature parameter. This approach upweights chunks that are representative of the document's central theme.

With **query attention**, instead of using the document mean as the query, we use a fixed query vector $\mathbf{q}$ that remains constant across documents (Lin et al., 2023):

$$b_{t,i} = \frac{\exp(\mathbf{c}_{t,i}^{\top}\mathbf{q}/\tau)}{\sum_{j=1}^{n_t}\exp(\mathbf{c}_{t,j}^{\top}\mathbf{q}/\tau)}, \quad \mathbf{v}_{\text{query},t} = \sum_{i=1}^{n_t} b_{t,i}\mathbf{c}_{t,i} \tag{6}$$

The query vector is the global mean across all documents in the corpus. In all experiments, we fix $\tau = 1.0$ (default softmax scaling) and keep this value constant across models and datasets for comparability. We do not perform a temperature sweep in this version.

Finally, the **ensemble** combines mean pooling and query attention through a weighted average:

$$\mathbf{v}_{\text{ensemble},t} = \lambda\,\mathbf{v}_{\text{mean},t} + (1-\lambda)\,\mathbf{v}_{\text{query},t} \tag{7}$$

with $\lambda = 0.5$ in our experiments, hedging between the stability of mean pooling and the selectivity of attention. Mean-covariance pooling adapts statistics pooling (Snyder et al., 2018) from speaker verification to document embeddings, while the ensemble is a straightforward blend of two existing strategies. To our knowledge, neither technique has previously been applied to measuring semantic drift in long-form documents.

Fig. 2 illustrates the conceptual difference between mean pooling and attention-based aggregation.

3.4. Bag-of-words lexical baselines

To benchmark transformer embeddings against traditional lexical representations, and to connect our drift measure to the textual-change literature in finance, we add document-level bag-of-words baselines in the spirit of Cohen et al. (2020). They compare each filing to its predecessor with several similarity measures, including cosine over raw term-frequency vectors and a Jaccard index over the binary set of terms, with the latter providing their strongest return predictive power.

We report two analogous lexical drift measures: a term frequency-inverse document frequency (TF-IDF) cosine drift, which weights terms by frequency, and a Jaccard drift, which reads only the set of terms present. We describe these as classical bag-of-words vector-space baselines rather than a direct replication, since we weight by TF-IDF rather than raw counts. Because cosine similarity and textual change are inversely related in the Cohen et al. (2020) framework, periods of low document-to-document similarity correspond to high-change (high-drift) periods in our terminology.

We fit a separete TF-IDF vectorizer per dataset over its full time series, with sub-linear term-frequency scaling, inverse-document-frequency weighting, L2 normalization, and a committed Portuguese stopword list to ensure that the vocabulary does not depend on the runtime environment. We apply document frequency pruning ($\min_{df} = 2$, $\max_{df} = 0.95$) but do not impose a max_features cap, allowing the vocabulary size to be determined by the corpus (ranging from approximately 7100 terms for SLC Agrícola to 25,000 for EDP). This avoids introducing an arbitrary hyperparameter, keeping the baseline simple and transparent.

The Jaccard baseline uses the same vocabulary, stopword list, and preprocessing pipeline, but represents each document using the binary term presence rather than weighted counts. All models, including the transformer-based representations, consume the same markdown text, without any additional numeric or table-specific filtering. Our PyMuPDF4LLM conversion removes headers, footers, and page artifacts. In templated commodity disclosures, tabular figures (e.g., planted area, production estimates, and prices) carry

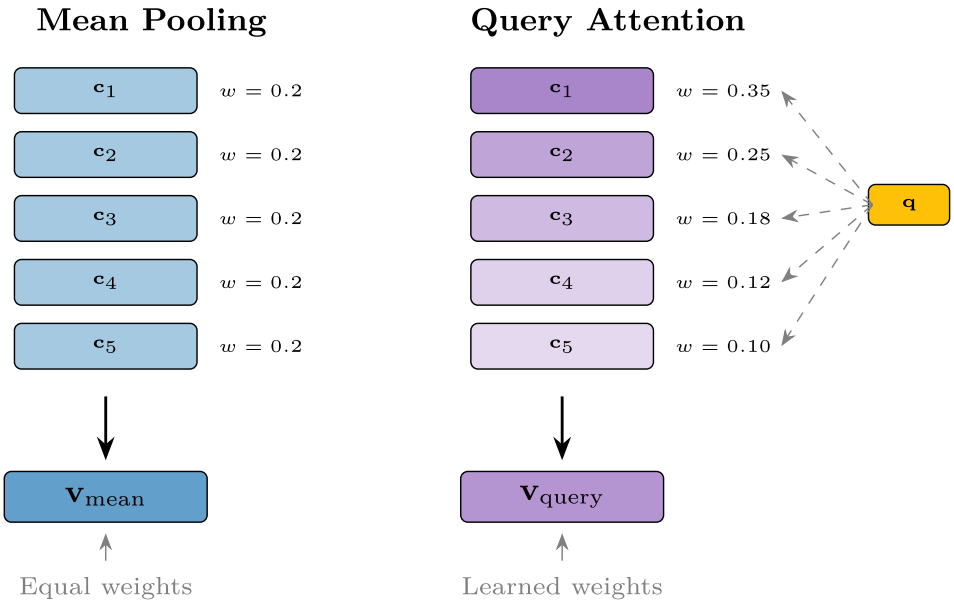


Fig. 2. Comparison of aggregation strategies. Mean pooling assigns equal weight to all chunks, while query attention computes similarity-based weights using a learned query vector $\mathbf{q}$, upweighting chunks most relevant to the query.

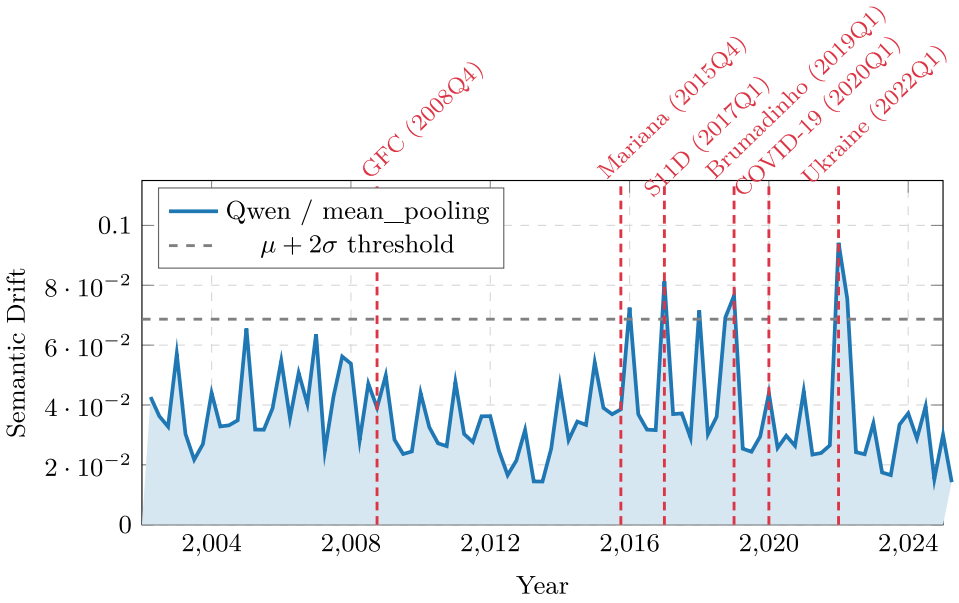


Fig. 3. Descriptive Vale drift timeline (Qwen mean pooling) with known events and a $\mu + 2\sigma$ threshold.

genuine semantic content rather than formatting noise. With identical input text across all representations, any difference in results reflects the representation rather than preprocessing.

Unlike contextual embeddings, sparse bag-of-words vectors have no token-level geometry for attention-based pooling to exploit, so the chunk-aggregation strategies above do not apply. We therefore treat each report as a single document, which is also the natural unit for the year-over-year similarity measure introduced by Cohen et al. (2020). For TF-IDF, drift is the same consecutive cosine distance used for the embedding models. For Jaccard, drift is one minus the Jaccard index between the term sets of consecutive documents, $1 - |A \cap B|/|A \cup B|$, corresponding to the binary set-overlap change studied by Cohen et al. (2020). Both baselines run the same four primary validation procedures as the embedding models (market correlation under circular-shift testing, lead-lag, Granger predictability, and the filing-date time-window design). One methodological difference relative to Cohen et al. (2020) is worth noting. They compare each filing to the same quarter of the prior year, thereby removing seasonal effects in reporting, whereas our consecutive formulation preserves them.

3.5. Datasets

We evaluate on periodic financial disclosures and institutional reports from five organizations in different sectors with different reporting frequencies.

- **Vale** is a multinational mining corporation and one of the world's largest iron ore producers. Its disclosures changed noticeably after the Mariana dam disaster in 2015 and the Brumadinho dam disaster in 2019.
- **Itaú Unibanco** is the largest private bank in Brazil. Its disclosures cover macroeconomic conditions and regulatory developments.
- **EDP Portugal** is a large European energy utility. The 2022 energy crisis offers a natural experiment for detecting event-driven semantic shifts.
- **Conab** (National Supply Company) publishes monthly grain market reports tracking Brazilian agricultural production. These reports follow a rigid template structure with strong seasonal patterns.
- **SLC Agrícola** is a Brazilian agribusiness with quarterly reporting. Compared to Conab, its disclosures are shorter and less templated, but still exhibit recurring seasonal and commodity-cycle language.

Table 1
Dataset characteristics and text statistics. All documents are in Portuguese. Pages from PDF metadata; chunks use each model's native tokenizer (max 506 tokens, 96-token overlap, min 40 tokens). Chunks/doc are medians; Q = Qwen3-Embedding-8B, B = BERTimbau, E = EuroBERT.

Dataset	Sector	Freq.	Docs	Period	Pages (p25/med/p75)	Chunks/doc (Q/B/E)
Vale	Mining	Quarterly	94	2002–2025	29/36/54	51/42/49
Itaú	Banking	Quarterly	70	2008–2025	130/148/160	336/289/299
EDP Portugal	Energy	Quarterly	81	2000–2025	94/138/186	244/189/220
Conab	Agriculture	Monthly	188	2009–2025	47/100/124	100/78/98
SLC Agrícola	Agribusiness	Quarterly	74	2007–2025	27/32/38	65/57/57

Table 1 reports summary statistics and characteristics of the datasets. All reports are publicly available through the respective investor relations websites and official institutional portals. Because consecutive drift requires two documents, the number of drift values is one fewer than the useable count. Market-overlap periods (drift vs. volatility): Vale 93, Itaú 69, EDP 80, Conab 187, SLC 73.

Document length changed over time for all five datasets. Because each chunk covers a fixed token budget, chunk counts per document track length directly; the trend is consistent across all three tokenizers (Qwen, BERTimbau, and EuroBERT). Comparing median chunks per document between the first and second halves of each series (values shown for the Qwen tokenizer), Vale (48–56), Conab (46–112), and SLC (55–79) grew, while Itaú (390–277) and EDP (274–227) shortened, reflecting changes in disclosure practices and regulatory scope rather than processing artifacts. Report format also varies both across and within companies. Conab follows a rigid template with fixed crop sections throughout its series, Itaú maintains a dense but stable structure, and Vale's reports shifted in both length and topical scope after the Brumadinho disaster. Using each dataset's article-scope best configuration, mean drift spans a 9.4-fold range, from 0.005 (Itaú) to 0.049 (EDP), exceeding the typical within-company coefficient of variation (median 0.75). This heterogeneity motivates per-dataset analysis rather than pooled estimation.

3.6. Validation methodology

A challenge in measuring semantic drift is the absence of ground truth. Unlike sentiment analysis, where annotators can label text as positive or negative, there is no objective standard for how much a document's meaning has "truly" changed. We validate against external signals that should correlate with meaningful semantic change if our measure works.

If semantic drift captures meaningful changes in corporate disclosure, high-drift periods should coincide with **market** uncertainty. We compute Spearman rank correlation between drift time series and realized volatility, computed as the standard deviation of daily returns within each reporting period. Spearman correlation measures monotonic relationships without assuming linearity and is robust to outliers.

For Conab, we use the stock price of SLC Agrícola as the market signal, as it is a major publicly traded company in the agricultural sector and is therefore more likely to reflect information relevant to these reports. While commodity prices such as corn or soybean could also be considered, a study by Silveira et al. (2025) finds that Conab reports have limited impact on corn futures prices. Statistical significance is assessed with a circular-shift permutation test (1000 shifts), which preserves autocorrelation in both series (Politis and Romano, 1994; Lancaster et al., 2018; Liu et al., 2021; Romano and Tirlea, 2022).

Because article-scope inference evaluates three models and five aggregations per dataset, we need a **multiple testing** policy. We report nominal circular-shift p-values and interpret results using BH FDR control within pre-declared families (Benjamini and Hochberg, 1995; Dror et al., 2017). BH-significant findings are treated as primary; nominal-only findings are reported as descriptive. For dependent tests at larger scale, alternative combination approaches exist (Bouamara et al., 2023).

To test predictive structure beyond contemporaneous association, we run **Granger causality** tests between drift and volatility using Augmented Dickey–Fuller (ADF) stationarity checks (Dickey and Fuller, 1979) to reduce spurious regression risk (Granger and Newbold, 1974; Hamilton, 1994) and Akaike Information Criterion (AIC)-selected lag order for Vector Autoregression (VAR) models (Akaike, 1974; Lütkepohl, 2005). These tests separate contemporaneous association from lagged predictability (Granger, 1969).

Period-level summaries can dilute effects tied to announcement dates. We therefore compute **filing-date time windows** around each filing date (e.g., $[-5, +63]$ trading days) and correlate drift with window volatility under the same circular-shift framework.

Major corporate events (disasters, regulatory actions, leadership changes) should trigger semantic shifts in subsequent disclosures. To explore this, we perform an **event alignment** by curating a table of known events for each dataset and compute the hit rate, defined as the proportion of events that coincide with drift spikes (values exceeding $\mu + 2\sigma$) within a one-period window. We also compute Precision@k by ranking periods by drift magnitude and counting a hit if a top-k period falls within the ± 1 event window. Best configurations are selected by highest hit rate, with Precision@5 as a tie-breaker; these rankings are descriptive given the small number of events.

This alignment serves as an empirical check rather than a formal validation. Since the event registry is hand-curated rather than exhaustive, it may inadvertently favor well-known crises, and the results should not be interpreted as absolute proof of correlation. Instead, this approach is intended to provide illustrative evidence of the relationship between semantic shifts and corporate events.

We also run filing-date **event study** tests following the market-model framework of MacKinlay (1997). For each filing, we estimate an OLS market model over a $[-120, -21]$ trading-day estimation window (roughly five months of data, ending 21 days before the filing to avoid pre-announcement contamination), yielding intercept $\widehat{\alpha}$ and slope $\widehat{\beta}$ against a benchmark index. Abnormal returns are $\mathrm{AR}_\tau = y_\tau - (\widehat{\alpha} + \widehat{\beta}\, y_{m,\tau})$, where y_τ is the asset return and $y_{m,\tau}$ is the benchmark return on trading day τ. The cumulative abnormal return (CAR) cumulates these residuals over a post-filing window; the annualized abnormal volatility (AVOL) is the annualized standard deviation of abnormal returns in a given window. We define $\Delta\mathrm{AVOL} = \mathrm{AVOL}_{\mathrm{post}} - \mathrm{AVOL}_{\mathrm{pre}}$ to measure the change in abnormal volatility from before to after the filing date.

We evaluate five window designs, all anchored at the filing date (day 0). **Symmetric windows** $w7$ and $w21$ use ± 7 and ± 21 trading days, comparing $[-7, -1]$ with $[+1, +7]$ and $[-21, -1]$ with $[+1, +21]$, respectively. **Buffered windows** $lag5_42$ and $lag5_63$ skip five trading days around the filing date, comparing $[-42, -5]$ with $[+5, +42]$ and $[-63, -5]$ with $[+5, +63]$, respectively, to avoid contamination from the announcement day itself. A fifth window, **quarterly between reports**, is dynamic: the pre-period runs from the previous filing date to the day before the current one, and the post-period from the day after to the next filing date. This design adapts to each company's reporting schedule.

Within each window, we split events into three **subgroups**: *all events* (the full sample), *high-drift* (filings whose drift exceeds the within-dataset median), and *low-drift* (the rest). This median split tests whether the magnitude of narrative change relates to the size of the volatility shift.

For variance comparison, we use **Levene's test** (Levene, 1960), which tests equality of variances between pooled pre-filing and post-filing returns within each subgroup. Unlike the *F*-test, Levene's test does not assume normality, which matters here because daily return distributions are heavy-tailed. To test whether individual event-level variance ratios are systematically positive (post > pre), we use a one-sided **Wilcoxon signed-rank test** (Wilcoxon, 1945) on $\log(\text{var}_{\text{post}}/\text{var}_{\text{pre}})$ across events. We also test ΔAVOL >0 with a one-sided Wilcoxon test.

Drift-scaling tests whether drift magnitude correlates with the size of the variance shift. We measure this as Spearman correlation between drift and the log variance ratio $\log(\text{var}_{\text{post}}/\text{var}_{\text{pre}})$, assessed under circular-shift permutation.

As a controlled perturbation, we run an **injection test** where we inject increasing percentages (5%, 10%, 25%, 50%) of text from three unrelated domains (technology, healthcare, retail) into documents and measure changes in consecutive drift. We interpret this as a sensitivity check, not a universal monotonicity test.

3.7. Experimental configurations

Document-level inference uses three models and five aggregation techniques, resulting in fifteen configurations per dataset. For market-correlation significance, we use circular-shift permutation tests with 1000 iterations.

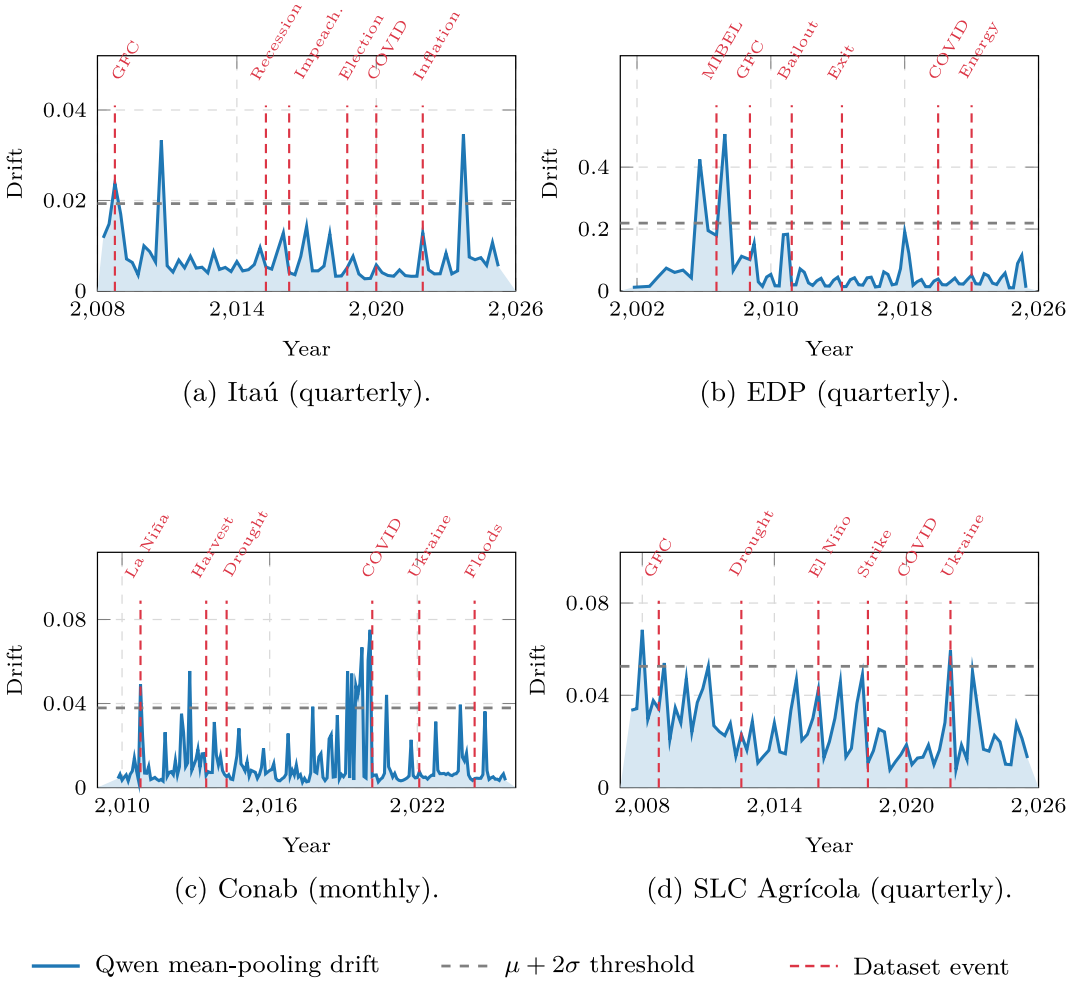


Fig. 4. Descriptive drift timelines (Qwen mean pooling) for Itaú, EDP, Conab, and SLC, each plotted against its own $\mu + 2\sigma$ threshold (gray dashed) and dataset-specific events (red dashed vertical lines, abbreviated above each panel). The panels share the construction of the Vale example in Fig. 3.

4. Results

This section presents descriptive, correlational, and causal evidence on the semantic drift signal. We begin with an event-based visual illustration, followed by contemporaneous and event-aligned market associations.

4.1. Event-based visual prelude

Before reporting inferential tests, we show a descriptive timeline that overlays semantic drift with known Vale events. Fig. 3 is intended as a visual hypothesis-building step, not as confirmatory evidence. The event registry is hand-curated and small, and multiple external events can plausibly map to the same drift spike. The figure suggests that some large drift values occur near major shocks (for example, Mariana, Brumadinho, and COVID-19), which is consistent with the hypothesis that disclosure language responds to material changes. We treat this as suggestive context only. Formal statistical validation follows below.

Fig. 4 extends this descriptive prelude to the remaining four datasets in a single four-panel view. Reading the panels together is essential, since the trajectories reveal how drift varies across datasets before any test is conducted. Vale's curve (Fig. 3) is dominated by a few sharp narrative excursions, with Brumadinho being the most prominent, whereas the templated quarterly and monthly series drift more gradually around their thresholds and respond to broad macroeconomic or agricultural conditions (the 2008 crisis, COVID-19, the 2022 commodity shock) rather than to individual disclosures. This visual contrast is exactly what the lexical-baseline comparison later quantifies (Section 4.13). When change is primarily narrative, as in Vale, semantic embeddings are required, while in the more templated series, simple term-overlap already captures much of the same movement. Because cosine similarity is $1 - \text{drift}$, these trajectories also correspond to the complementary similarity view used in prior textual-disclosure work. To avoid redundancy, we omit a separate similarity figure.

4.2. Market correlation

Table 2 reports Spearman correlations between semantic drift and stock volatility for the best-performing configuration per dataset. The "Best configuration" column identifies the model and aggregation pairing that yields the largest absolute drift–volatility correlation among the fifteen article-scope configurations (three models $\times$ five aggregations) for each dataset. The column ρ presents the Spearman correlation between the drift series and annualized realized volatility over n matched periods; the corresponding p-value is computed using a circular-shift permutation test (1000 shifts) that preserves serial dependence in the drift series. Because multiple configurations are evaluated per dataset, we treat BH-adjusted significance as the primary criterion, applying the BH correction within pre-declared families (across a model's five aggregations, the full set of fifteen configurations, and the three stock metrics), as detailed for Vale below. To keep the main text readable, the table shows one configuration per dataset; the complete results for all fifteen configurations across all datasets and metrics, together with the drift series and event registries, are released in the public repository referenced in Section 7.

Vale shows the strongest correlation. Of the 15 article-scope configurations tested on Vale, all five Qwen aggregations achieve statistical significance under circular-shift permutation testing (each with $p = 0.03197$) and remain significant under BH correction within the Qwen aggregation family ($q = 0.03197$). Under stricter families the result becomes marginal. BH across all 15 model-aggregation configurations gives $q = 0.096$, and BH across the three stock metrics (volatility, absolute return, post-period return) for the best configuration gives $q = 0.096$. Vale's drift-volatility association is therefore robust under the aggregation family but family-sensitive under broader corrections.

Fig. 5 visualizes the correlation strength across datasets and models.

Itaú, EDP, Conab, and SLC show weak or non-significant correlations with their respective market signals. This variation suggests that the relationship between semantic drift and market outcomes depends on the company and document type.

4.3. Event alignment

Event alignment remains descriptive, since the event registry is small and the mapping from a drift spike to a specific external event is not unique. For Vale, the best-configuration hit rate is high (0.83), but the permutation result is not below the 0.05 threshold ($p = 0.101$), so we treat this as suggestive rather than confirmatory. We therefore use event alignment as contextual evidence and prioritize the inferential tests tied to market series and filing windows.

Table 2
Market correlation results using circular-shift permutation p-values (1000 shifts). Bold indicates statistical significance at $p < 0.05$.

Dataset	Best Configuration	ρ	p-value	n
Vale	Qwen/attn_chunk	**0.359**	0.03197	93
Itaú	BERTimbau/mean_pooling	0.183	0.19680	69
EDP	BERTimbau/mean_cov	−0.097	0.57942	80
Conab	BERTimbau/mean_cov	0.091	0.44655	187
SLC	Qwen/query_attn	0.223	0.19381	73

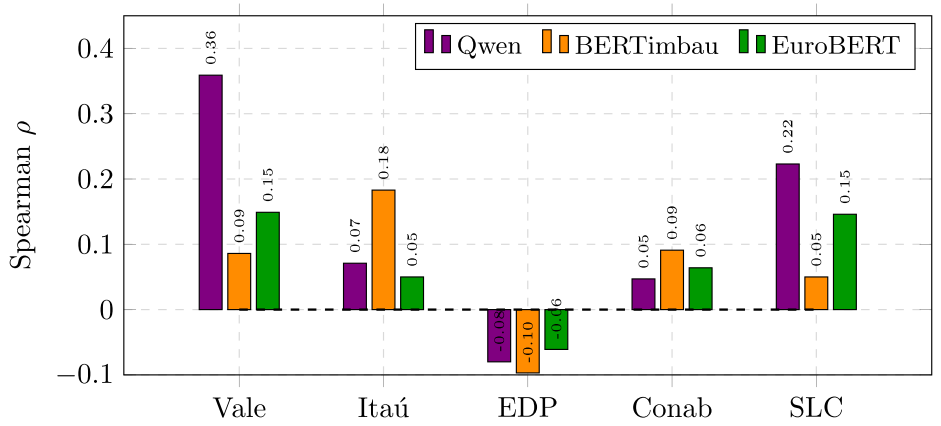


Fig. 5. Market correlation by dataset and embedding model (best aggregation per model). Only Vale achieves statistically significant correlation with stock volatility under circular-shift permutation testing. Correlations for Itaú, EDP, Conab, and SLC do not reach $p < 0.05$ in this setup. The dashed line marks zero correlation.

4.4. Lead-lag analysis

We test lead-lag structure by scanning cross-correlations between drift and volatility over a grid of lags and evaluating circular-shift permutation p-values. Table 3 summarizes, for each dataset, the strongest positive lead lag (+1 to +6) under the best configuration from Table 2; lag 0 results are reported separately in market-correlation tables.

4.5. Granger causality

Correlation is contemporaneous and does not imply temporal predictability. We therefore run Granger tests with ADF stationarity checks and AIC-selected VAR lags Granger (1969); Akaike (1974); Dickey and Fuller (1979); Lütkepohl (2005);. EDP is the dataset with the strongest Granger evidence. Drift Granger-causes volatility in all 15 model and aggregation configurations (p range 0.00048–0.00737), while the reverse direction is non-significant in all configurations. Under the SLC stock proxy, Conab drift also Granger-causes volatility in all 15 configurations (p range 0.00290–0.04716, reverse non-significant), though this result is conditional on the proxy choice and should be interpreted as secondary evidence.

4.6. Filing-date time-window volatility

To avoid averaging away announcement-centered effects, we compute a filing-date anchored window volatility signal and correlate it with drift using circular-shift permutation tests. Table 4 reports the [−5, +63] trading-day window results using the same best configuration per dataset as Table 2. SLC is the only robust BH-significant association in this window run.

4.7. Drift-tail volatility

Beyond correlation, we test whether the highest-drift tail concentrates higher volatility than the remaining periods using one-sided circular-shift tests with global BH correction. The strongest robust result is for SLC in the $p5d_p42d$ window, where the top-20% drift tail has higher volatility than the rest (mean difference +0.1065, ratio 1.319, global BH $p = 0.01066$). Itaú also shows robust tail effects in short symmetric windows, including $3d_3d$ (top-10%, +0.2138, ratio 1.726, global BH $p = 0.01066$) and $7d_7d$ (top-10%, +0.1673, ratio 1.573, global BH $p = 0.01066$).

Table 3
Lead-lag cross-correlation between drift and volatility for each dataset (best configuration per dataset), reporting the strongest positive lead lag (+1 to +6). Positive lags indicate drift leading volatility. None of these peak positive–lag correlations reach $p < 0.05$ under circular-shift testing.

Dataset	Configuration	Lag	ρ	p-value
Vale	Qwen/attn_chunk	+1	0.290	0.05994
Itaú	BERTimbau/mean_pooling	+5	−0.098	0.58242
EDP	BERTimbau/mean_cov	+3	−0.251	0.07293
Conab	BERTimbau/mean_cov	+1	0.164	0.15784
SLC	Qwen/query_attn	+3	0.280	0.08492

Table 4
Time-window correlation between drift and window volatility for [−5, +63] trading days anchored at filing dates, using the best configuration per dataset from Table 2 *p*-values use circular-shift permutations; BH controls FDR across all dataset and metric tests in this window run.

Dataset	ρ	*p*-value	BH *p*	*n*
Vale	0.223	0.08691	0.43457	93
Itaú	−0.021	0.90110	0.98402	69
EDP	0.067	0.72228	0.98402	74
Conab	0.128	0.33566	0.97652	187
SLC	**0.357**	**0.000999**	**0.01499**	73

4.8. CAR and AVOL around filing dates

Using a market-model event study setup MacKinlay (1997), we evaluate drift links to CAR and to abnormal volatility shifts (AVOL). In current outputs, we do not observe robust drift-scaling for CAR or for ΔAVOL under global BH correction. However, EDP exhibits a robust short-window abnormal volatility increase after filings. For *w7*, mean ΔAVOL is 0.0687 and a one-sided Wilcoxon test rejects $\Delta\text{AVOL} \leq 0$ with global BH $p = 0.00014$.

4.9. Variance–shift tests around filing dates

Table 5 reports BH-corrected Levene results comparing post-filing versus pre-filing return variance. For EDP, post-filing variance exceeds pre-filing variance in both the full event pool and the high-drift subgroup. For Vale, the strongest evidence appears in high-drift filings (longer buffered and quarterly windows), with additional BH-significant effects in some low-drift windows. Filing dates coincide with volatility regime changes even when drift magnitude does not scale with the size of that change.

4.10. Macroeconomic correlation

We considered correlating drift with broad market uncertainty proxies as an exploratory check. Because these signals are not part of the validated inference policy used for the main results, we omit them from the primary evidence bundle.

4.11. Cross-corpus drift: Conab and SLC Agrícola

Conab's monthly agricultural reports and SLC Agrícola's quarterly filings cover overlapping commodity markets (soybeans, corn), so a natural question is whether drift in one corpus leads or follows drift in the other. We align the two drift series to common quarterly periods and compute lagged Spearman cross-correlations with circular-shift *p*-values. Fig. 6 shows the result for all three embedding models. With Qwen embeddings, a negative correlation appears at lag +2 ($\rho = -0.43$, raw $p = 0.012$), but this does not survive BH correction across the nine tested lags ($q_{\text{BH}} = 0.108$). BERTimbau shows a weaker version of the same pattern ($\rho = -0.22$ at lag +2, $q_{\text{BH}} = 0.74$). For EuroBERT, the largest absolute correlation appears at lag −4 ($\rho = 0.35$, raw $p = 0.099$, $q_{\text{BH}} = 0.59$), while lag +2 is weaker ($\rho = -0.22$, $q_{\text{BH}} = 0.59$). The descriptive pattern is consistent with a delayed inverse relationship (Conab drift precedes an opposite shift in SLC drift two quarters later), but the effect is not statistically confirmed under our correction policy.

4.12. Model and aggregation comparison

For Vale, we can compare all article-scope model and aggregation combinations directly. Table 6 reports Spearman correlations with circular-shift *p*-values for drift versus annualized volatility.

Fig. 7 shows Spearman correlations across all model-aggregation combinations for Vale.

The model comparison follows the same qualitative pattern. In Vale, Qwen shows the strongest correlations across aggregations, while EuroBERT and BERTimbau stay in a similar lower-correlation range. Across datasets, EuroBERT does not show systematic gains in inferential significance.

We see three complementary factors for Qwen's advantage on Vale, presented as a hypothesis rather than a benchmark claim. The first is capacity and scale: Qwen3-Embedding-8B has approximately 8B parameters and 4096-dimensional outputs, significantly

Table 5
Levene variance-shift tests around filing dates (post-filing vs. pre-filing return variance). BH correction is applied within each dataset across window and subgroup combinations.

Dataset	Window	Subgroup	BH *p*
EDP	*w7*	All events	**0.00016**
EDP	*w7*	High-drift	**0.0022**
Vale	*lag*5_42	High-drift	**0.010**
Vale	*lag*5_63	High-drift	**0.005**
Vale	Quarterly	High-drift	**0.016**

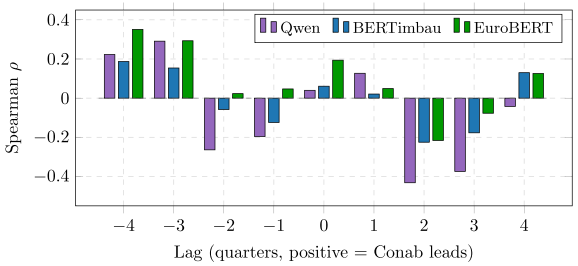


Fig. 6. Cross-correlation between Conab drift and SLC Agrícola drift at quarterly lags for Qwen, BERTimbau, and EuroBERT. Positive lags mean Conab leads. None of the lag effects survives BH correction across the tested lag family.

Table 6
Vale drift versus volatility across all article-scope configurations (3 models and 5 aggregations). BH p applies BH correction within each model across its five aggregations.

Model	Aggregation	ρ	p-value	BH p
Qwen	mean_pooling	**0.358**	0.03197	0.03197
Qwen	mean_cov	**0.303**	0.03197	0.03197
Qwen	attn_chunk	**0.359**	0.03197	0.03197
Qwen	query_attn	**0.358**	0.03197	0.03197
Qwen	Ensemble	**0.357**	0.03197	0.03197
BERTimbau	mean_pooling	0.069	0.57842	0.72303
BERTimbau	mean_cov	0.031	0.85215	0.85215
BERTimbau	attn_chunk	0.083	0.51648	0.72303
BERTimbau	query_attn	0.086	0.51648	0.72303
BERTimbau	Ensemble	0.077	0.53746	0.72303
EuroBERT	mean_pooling	0.147	0.13387	0.16733
EuroBERT	mean_cov	0.093	0.49051	0.49051
EuroBERT	attn_chunk	0.149	0.12787	0.16733
EuroBERT	query_attn	0.146	0.12388	0.16733
EuroBERT	Ensemble	0.149	0.12388	0.16733

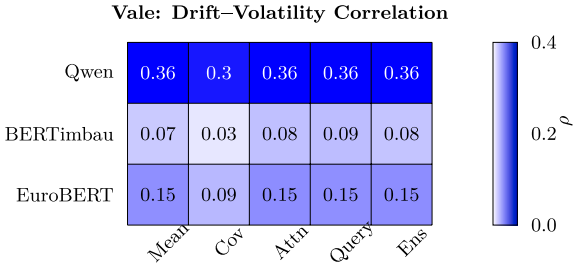


Fig. 7. Spearman correlation (ρ) between semantic drift and stock volatility for all model-aggregation combinations on Vale. Darker cells indicate stronger correlation. Qwen yields the largest correlations, while EuroBERT and BERTimbau remain lower and broadly similar.

outscaling BERTimbau-large (335M parameters and 1024 dimensions) and EuroBERT (210M parameters and 768 dimensions), leaving more room to encode subtle topical reorganization. The second is the training objective: Qwen is optimized using a contrastive retrieval and semantic-similarity loss and ranks at the top of the Massive Text Embedding Benchmark (MTEB), so its embedding geometry is explicitly aligned with cosine distance, the quantity captured by our drift measure. The third is the nature of Vale's post-Brumadinho shift, a broad topical reorganization, in which safety, remediation, and governance language enters multiple sections simultaneously. A higher-capacity model tracks this distributed change more faithfully than encoders working under smaller context windows. The gap is largest exactly where the change is most semantic rather than lexical.

4.13. Lexical baseline comparison

Table 7 sets the two document-level lexical baselines, TF-IDF and Jaccard, against each dataset's transformer best configuration across the four primary tests. The pattern is dataset-dependent rather than a clean sweep. Vale is the clearest discriminator: the transformer correlates with volatility (market $\rho = 0.359$, $p = 0.032$) while both lexical baselines stay flat (0.049 for TF-IDF, -0.017 for Jaccard), and the same gap holds across the lead-lag and time-window tests. This is where semantic representation earns its place, as

Table 7
Document-level lexical baselines (TF-IDF and Jaccard) versus the transformer best configuration per dataset across the four primary tests. The market block reports the correlation between drift and annualized volatility (ρ), along with its circular-shift p-value and BH-adjusted q_{BH}; the lead-lag column reports the strongest positive drift → volatility lead (lag in periods and Spearman ρ); the Granger column reports the drift → volatility p-value at the AIC-selected lag; the time-window block repeats the [−5, +63] filing-date design of Table 4. Bold indicates $p < 0.05$. The n column is the market-overlap count; the time-window n is identical except for EDP ($n = 74$).

Dataset	Representation	n	Market vs. volatility			Lead-lag (lag:ρ)	Granger p	Time window [−5, +63]		
			ρ	p	q_{BH}			ρ	p	q_{BH}
Vale	Qwen/attn_chunk	93	**0.359**	**0.032**	0.096	+1: 0.29	0.073	0.223	0.087	0.435
	TF-IDF	93	0.049	0.790	0.790	+3: −0.11	0.313	0.036	0.893	0.893
	Jaccard	93	−0.017	0.953	0.953	+3: −0.13	0.163	−0.021	0.912	0.912
Itaú	BERTimbau/mean_pooling	69	0.183	0.197	0.469	+5: −0.10	0.649	−0.021	0.901	0.984
	TF-IDF	69	0.216	0.088	0.264	+6: −0.19	0.163	0.084	0.484	0.806
	Jaccard	69	0.235	0.088	0.264	+6: −0.20	0.274	0.069	0.521	0.840
EDP	BERTimbau/mean_cov	80	−0.097	0.579	0.855	+3: −0.25	**0.007**	0.067	0.722	0.984
	TF-IDF	80	−0.046	0.803	0.958	+3: −0.21	0.564	0.058	0.802	0.864
	Jaccard	80	−0.022	0.902	0.902	+3: −0.23	0.552	0.085	0.684	0.840
Conab	BERTimbau/mean_cov	187	0.091	0.447	0.670	+1: 0.16	**0.047**	0.128	0.336	0.977
	TF-IDF	187	0.116	0.247	0.740	+1: 0.17	0.064	0.120	0.337	0.806
	Jaccard	187	0.095	0.287	0.860	+1: 0.15	0.101	0.077	0.450	0.840
SLC Agrícola	Qwen/query_attn	73	0.223	0.194	0.291	+3: 0.28	0.429	**0.357**	**0.001**	**0.015**
	TF-IDF	73	0.076	0.619	0.929	+3: 0.15	0.603	0.179	0.156	0.779
	Jaccard	73	0.179	0.251	0.376	+3: 0.32	0.357	**0.353**	**0.001**	**0.015**

Vale's drift is a broad topical reorganization rather than keyword turnover. The two significant Granger results, EDP and Conab, likewise belong to the transformer models.

On the more templated series, the lexical baselines close the gap. TF-IDF slightly exceeds the transformer on the Itaú and Conab market correlations (0.216 versus 0.183, and 0.116 versus 0.091), where period-to-period change is largely lexical churn. The sharpest case is SLC Agrícola in the filing-date time-window test, where Jaccard ($\rho = 0.353$, $p = 0.001$, $q_{BH} = 0.015$) matches the transformer ($\rho = 0.357$, $p = 0.001$, $q_{BH} = 0.015$) to within noise, both surviving BH correction, while TF-IDF weighting does not ($\rho = 0.179$, $p = 0.156$).

On these templated quarterly filings the binary set-overlap measure that Cohen et al. (2020) found most predictive tracks period-over-period change as well as the embedding-based representation; here, the transformer does not outperform it but matches it. We therefore interpret the table as showing semantic embeddings dominating where drift is narrative and topical (Vale), and lexical overlap being sufficient where change is templated and keyword-driven (SLC Agrícola, Conab, Itaú). This is consistent with the investor-inattention mechanism of Cohen et al. (2020), under which market co-movement is driven by how meaning is reorganized, except when disclosures exhibit little meaningful reorganization at all.

4.14. Validation of drift signal

We treat injection-based perturbations as a sensitivity check for consecutive drift, testing three out-of-domain sources (technology, healthcare, retail) at four dosages per dataset. At 50% injection the technology domain produces drift changes of Vale −16.8%, Itaú +7.4%, EDP −30.0%, Conab +10.5%, SLC −12.6%. Three of the five datasets show negative shifts, consistent with the same external text homogenizing adjacent documents and compressing consecutive drift. Itaú and Conab are the exceptions. For Itaú, technology injection increases drift, possibly because its financial vocabulary is distant enough from technology text that the added material introduces between-period variation rather than convergence. For Conab, the positive shift may reflect a similar vocabulary gap between agricultural reports and technology content. The healthcare domain reinforces this pattern, yielding an even larger positive shift for Itaú (+19.4% at 50%). Across all domains, drift responds to content perturbation but does not monotonically increase with dosage for every domain pairing, consistent with our interpretation of injection as a sensitivity check rather than a calibration tool.

5. Discussion

This section interprets the empirical findings, discusses their practical implications, and outlines key limitations.

5.1. Variation across companies and document types

Results vary in how semantic drift relates to external signals across companies and document types. For Vale, a mining company whose disclosures are exposed to large operational and commodity shocks, semantic drift aligns with market volatility at lag 0 under circular-shift testing. Shifts in language and market uncertainty co-occur rather than strictly leading or lagging. Vale also shows significant pre-to-post variance shifts around filings (Table 5), strongest in high-drift windows, linking large narrative changes to volatility regime transitions around those dates.

For Itaú, period-level drift versus volatility correlations are small and do not reach $p < 0.05$ under circular-shift testing in the validated configurations, and lead-lag scans show no statistically reliable peaks under the same criterion. At the same time, filing-centered tail tests indicate that the highest-drift periods can concentrate higher short-window volatility, which is consistent with announcement-centered effects that are not well captured by period aggregation.

For EDP, the most consistent signal is temporal rather than contemporaneous. Granger tests show that drift Granger-causes volatility across all article-scope configurations, while the reverse direction is not significant. This pattern is consistent with lagged predictability from drift to volatility in this sample, without establishing a causal mechanism. Post-filing return variance at *w7* is also significantly higher than pre-filing variance (Table 5), so EDP filings coincide with a shift in the volatility regime.

For SLC, the strongest evidence is tied to filing-date designs. A [−5, +63] window shows a robust association between drift and window volatility, and drift-tail tests again concentrate higher volatility in high-drift windows.

These contrasting observations show that the method detects meaningful signals, but interpreting those signals depends on company characteristics and document type. The method is general; the nature of information captured varies across disclosure contexts.

As a robustness check, we ran exploratory tests that did not reach significance. We split the sample at 2016-01-01, the approximate midpoint of the overlapping data range across datasets, and tested whether the drift-volatility slope differs between the two halves. The split is not tied to any known structural event; it simply divides the sample into roughly equal groups. For SLC the interaction term flips sign across regimes but does not survive the permutation test ($p_{\text{perm}} = 0.13$, $q_{\text{BH}} = 0.53$). Targeted window checks (Conab drift against SLC stock; spike-window variance for Itaú and SLC) produce only nominal hits ($p < 0.05$) that fail BH correction. We report these null findings to document where the current evidence ends.

5.2. Factors influencing low correlation in agricultural data

Conab's agricultural reports show weak contemporaneous correlation with SLC stock volatility (best quarterly $\rho = 0.091$, circular-shift $p = 0.447$), and event alignment is not statistically confirmed (hit-rate permutation $p = 0.261$, MRR = 0.016). At the same time, Conab exhibits lagged Granger predictability under the same proxy, so the evidence is mixed rather than null. We considered several explanations.

Conab reports follow a standardized format with fixed sections for each crop. Most content variation occurs within constrained fields (production estimates, area planted), which limits the scope for **semantic shift** in the first place.

Agricultural commodity markets also follow strong seasonal cycles. If drift inherits that cyclicality, any residual market correlation could reflect shared **seasonality** rather than informational content. Disentangling seasonal from informational drift remains an open direction for future work.

This links back to the textual-change literature in finance. Cohen et al. (2020) compare each filing to the same quarter of the prior year precisely to cancel filing seasonality, whereas our consecutive drift inherits it, which partly explains why Conab's contemporaneous coupling is weak. The lexical baselines sharpen this interpretation. On the templated Conab and Itaú series, TF-IDF matches or slightly exceeds the transformer market correlation (Table 7), because most period-to-period change there reflects lexical churn within fixed fields; and on SLC Agrícola's filing-date window test, the Jaccard overlap measure ties the transformer ($\rho \approx 0.35$, $q_{\text{BH}} = 0.015$ for both), corresponding to the binary set-overlap signal Cohen et al. (2020) found most predictive. Where the change is genuinely semantic, as in Vale, the transformer performs substantially better while both lexical baselines remain flat. We interpret this as semantic embeddings dominating where drift is narrative and topical, and lexical overlap being sufficient where change is templated and keyword-driven, a qualitative distinction that is linked in Cohen et al. (2020) to subsequent returns.

A third factor is timing. Agricultural production cycles may not align with the monthly reporting frequency, creating a disconnect between when information changes and when reports are published.

SLC Agrícola operates in the same sector yet shows robust filing-window drift associations in its own quarterly earnings reports (Table 4, BH $p = 0.015$) and significant drift-tail volatility concentration. The weak Conab signal therefore reflects report format and publication frequency rather than any inherent insensitivity of agricultural markets to textual information.

5.3. Practical implications

Each finding maps to a different stakeholder, and each comes with limits set by the evidence.

Securities regulators have increasingly shifted from compliance-based supervision to risk-based supervision (RBS) in order to identify, assess, and mitigate the most significant threats to market integrity, investor protection, and financial stability. This approach requires proactive risk identification and the calibration of supervisory intensity to regulated entities' risk profiles, enabling oversight that is more efficient, proportionate, and effective.

RBS relies heavily on the collection and analysis of extensive, high-quality data (e.g., key risk indicators) to inform supervisory judgments. Securities regulators who review large volumes of periodic filings could use drift as a screening tool. In the EDP dataset, a Portuguese energy utility whose filings fall under the *Comissão do Mercado de Valores Mobiliários* (CMVM), drift is associated with subsequent volatility across all fifteen article-scope configurations in this sample ($p < 0.008$ in every case). Large narrative changes in these disclosures coincided with periods of higher market volatility, and a system that flags filings whose drift exceeds a historical threshold would let analysts focus on the most changed documents first. The same logic applies to other jurisdictions. In October 2024, Brazil's C*omissão de Valores Mobiliários* (CVM) adopted resolutions requiring ISSB-aligned sustainability reporting for fiscal years

beginning on or after January 2026 (Comissão de Valores Mobiliários, 2024). As these disclosures accumulate, drift monitoring could track whether firms converge toward shared language or diverge into company-specific framing.

Investor-relations and compliance teams inside firms could treat drift as a consistency check on successive report drafts. For Vale, the positive drift-volatility association ($\rho = 0.36$, significant under the within-model aggregation family) and the robust variance shifts around filing dates suggest that large narrative revisions track periods of market stress. The association loses significance under stricter correction families ($q_{\text{BH-metrics}} = 0.096$, $q_{\text{BH-configs}} = 0.096$), so any internal dashboard would need firm-specific calibration rather than a fixed cutoff.

Buy-side analysts face a different use case. SLC and Itaú's highest-drift filings concentrate significantly elevated short-window volatility under global BH correction ($q < 0.011$). Extreme narrative changes carry information about event-window risk even when period-level correlations are weak. Ranking filings by drift magnitude within a sector would let a screening tool surface the reports most worth reading in full.

Conab now illustrates a mixed case rather than a strict null. With SLC as proxy, lagged predictability appears in Granger tests, but contemporaneous correlation and event alignment remain weak. Templated reports with rigid structure and seasonal content likely dampen immediate drift-to-market coupling, so this setting remains sensitive to proxy choice. Regulators designing disclosure standards face a real tension here, since comparability benefits from templates, but templates limit what automated tools can extract.

Embedding drift matters for model risk management too. If an embedding model receives a silent update, or a new version tokenizes Portuguese text differently, the drift series shifts for reasons unrelated to disclosure content. Qwen3-Embedding-8B, BERTimbau, and EuroBERT produce different aggregation rankings in several of our datasets (Section 4), which makes model-level drift tracking a practical requirement for production deployments.

5.4. Limitations

Several methodological constraints should be considered when interpreting the findings.

With 70 to 188 documents per dataset (69–187 periods overlapping with market data) and small event registries for alignment testing, **statistical power** is limited. The significant results for Vale, EDP, Conab (proxy-based Granger), and SLC are encouraging, but replication on larger corpora would strengthen confidence.

The **construct validity** of drift also deserves scrutiny. Cosine distance between consecutive document embeddings does not distinguish semantic content changes from structural ones. Shifts in table layouts, boilerplate sections, header formatting, or report templates can all move the embedding vector without reflecting a genuine change in disclosed information. Our markdown conversion strips some formatting, but residual structure effects remain possible. A systematic analysis of confounds (e.g., correlating drift with chunk count, document length, or within-document embedding variance) would help isolate genuine semantic change from formatting artifacts, and remains an open direction for future work.

All our documents are in Portuguese. Whether results generalize to English 10-K filings or other regulatory contexts remains untested, so **language and domain specificity** is an open question.

We also lack labeled examples of "true" semantic drift. Our validation relies on external correlates (market data, events), which may capture only some meaningful linguistic changes. Without **ground truth**, the boundary between genuine informational shift and noise is difficult to draw.

On the infrastructure side, Qwen3-Embedding-8B embeddings were generated through the OpenRouter API (qwen/qwen3-embedding-8b, created October 2025) during 2025–2026. API-served models can receive silent weight updates or quantization changes, so exact **numeric reproduction** requires either pinning a model snapshot or re-embedding from a locally hosted checkpoint. To mitigate this, our pipeline computes and stores a SHA-256 hash manifest for every embedding file immediately after generation, recording the digest, file size, and model metadata. Any future re-embedding that produces a different hash would flag a model-side change. BERTimbau embeddings, produced from a fixed HuggingFace checkpoint, do not share this concern but are hashed identically for completeness.

6. Conclusion

Financial disclosures change over time, and tracking these semantic shifts may reveal signals useful for several applications. For example, monitoring these changes can help anticipate shifts in corporate strategy or financial health, supporting alpha generation and investment monitoring. Semantic trajectories can also enhance regulatory screening and forensic accounting by detecting subtle language shifts used to obscure risk or non-compliance. They can also enable quantitative analysis of policy communication shifts, such as changes in monetary stance. While textual drift analysis can be used to detect significant shifts, existing literature predominantly focuses on short-form documents, datasets with limited temporal range, or on English-language texts.

To address this gap, we proposed a framework that measures semantic drift in Portuguese disclosures with document embeddings. The pipeline chunks documents, embeds each chunk with three complementary models, and aggregates chunk representations through multiple pooling strategies to capture how disclosure language changes from one reporting period to the next.

Testing on quarterly reports from mining, banking, energy, and agribusiness firms, plus monthly agricultural market reports, we find.

- The main contribution is a replicable method for measuring semantic drift in long-form disclosures, including a comparison of long-document aggregation strategies and their stability.

- External validation is context-dependent. Vale shows a positive drift–volatility association that is significant under the within-model aggregation family but sensitive to stricter multiple-testing families. EDP exhibits robust Granger predictability from drift to volatility across all article-scope configurations. SLC shows a robust filing-date window association at $[-5, +63]$. Conab exhibits a similar Granger pattern, but only when mapped to SLC stock as market proxy, making that finding proxy-sensitive and secondary to the EDP result.
- Lead-lag scans do not yield statistically reliable peak–lag correlations under circular-shift testing in the validated configurations.
- Tail tests strengthen the announcement-centered picture. For SLC and Itaú, the highest-drift tail concentrates higher window volatility under global BH correction.
- CAR and AVOL event studies do not show robust drift-scaling, but EDP shows a robust short-window abnormal volatility increase after filings (AVOL, $w7$).
- Aggregation choices matter, but simple pooling can be competitive. Mean pooling performs comparably to attention-based methods in several settings.

Across datasets, the qualitative pattern remains stable across the three models. Model-level rankings vary in some cases, but the main inferential conclusions remain similar overall.

The main contribution is the method itself, a way to quantify semantic change in corporate disclosures using document embeddings. Validation results show that the method produces economically meaningful signals, though interpreting those signals depends on company characteristics and document type. The method offers a quantitative tool for researchers and practitioners interested in tracking how corporate communication evolves.

Future work will prioritize validation design and interpretability over more complex aggregation. In our experiments, simple pooling is often competitive and, for some model–dataset pairs, different aggregation choices yield very similar drift series and external-validation outcomes. The next step is therefore to strengthen how drift is interpreted and stress-tested by expanding the set of firms and disclosure types, formalize and document event registries, and apply time-series significance testing consistently across validations. We also plan to develop attribution methods that link drift spikes to the underlying text and to study how preprocessing choices such as boilerplate and template structure affect what the drift signal captures.

CRediT authorship contribution statement

João Victor M. Correa: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing, Visualization. **Jorge Miguel Bravo:** Conceptualization, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Rodrigo Lanna Franco da Silveira:** Conceptualization, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **José César Cruz Júnior:** Conceptualization, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Fernando Batista:** Conceptualization, Investigation, Methodology, Validation, Writing – original draft, Writing – review & editing. **Renato Moraes Silva:** Conceptualization, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Writing – original draft, Writing – review & editing.

Data and code availability

The code and data supporting this study are publicly available. The corporate disclosures analyzed here are public filings, and the source documents are released together with the analysis code, which regenerates the reported results from those documents. The repository, including the result tables and drift series, is available at https://github.com/nilc-nlp/semantic-drift-disclosures.

Declaration of competing interest

The authors declare the following financial interests (e.g., any funding for the research project)/personal relationships (e.g., the author is an employee of a profitable company) which may be considered as potential competing interests:

Financial interests: None.

Personal relationships: One of the authors is employed by Advance Trading Inc., Bloomington, Illinois, USA. The company had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

The remaining authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We gratefully acknowledge the support provided by São Paulo Research Foundation (FAPESP; grant #2024/17834-6).

References

Adosoglou, G., Lombardo, G., Pardalos, P.M., 2021. Neural network embeddings on corporate annual filings for portfolio selection. Expert Syst. Appl. 164, 114053. https://doi.org/10.1016/j.eswa.2020.114053.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19, 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Batista, F., Mamede, N., Trancoso, I., 2008. Language dynamics and capitalization using maximum entropy. In: Proc. of ACL-08: HLT, Short Papers. ACL, pp. 1–4.

Batista, F., Moniz, H., Trancoso, I., Mamede, N.J., 2012. Bilingual experiments on automatic recovery of capitalization and punctuation of automatic speech transcripts. IEEE Transactions on Audio, Speech and Language Processing, Special Issue on New Frontiers in Rich Transcription 20, 474–485.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society Series B: Statistical Methodology 57, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.

Boizard, N., Gisserot-Boukhlef, H., Alves, D.M., Martins, A., Hammal, A., Corro, C., Hudelot, C., Malherbe, E., Malaboeuf, E., Jourdan, F., Hautreux, G., Alves, J., El-Haddad, K., Faysse, M., Peyrard, M., Guerreiro, N.M., Fernandes, P., Rei, R., Colombo, P., 2025. Eurobert: Scaling Multilingual Encoders for European Languages. https://arxiv.org/abs/2503.05500.arXiv:2503.05500.

Bouamara, N., Laurent, S., Shi, S., 2023. Sequential Cauchy Combination Test for Multiple Testing Problems with Financial Applications. https://arxiv.org/abs/2303.13406.arXiv:2303.13406.

Brown, S.V., Tucker, J.W., 2011. Large-sample evidence on firms' year-over-year MD&A modifications. J. Account. Res. 49, 309–346. https://doi.org/10.1111/j.1475-679X.2010.00396.x.

Cohen, L., Malloy, C., Nguyen, Q., 2020. Lazy prices. J. Finance 75, 1371–1415. https://doi.org/10.1111/jofi.12885.

Comissão de Valores Mobiliários, 2024. Resoluções CVM 217, 218 E 219. Regulatory Resolutions. https://www.gov.br/cvm/pt-br/assuntos/noticias/2024/cvm-edita-as-resolucoes-217-218-e-219. published 29 October 2024. Make CBPS Technical Pronouncements No. 01 (IFRS S1) and No. 02 (IFRS S2) mandatory for publicly traded companies, applicable to fiscal years beginning on or after 1 January 2026.

Dai, A.M., Olah, C., Le, Q.V., 2015. Document embedding with paragraph vectors. In: Arxiv. https://arxiv.org/abs/1507.07998. arXiv:1507.07998.

Dickey, D.A., Fuller, W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root. J. Am. Stat. Assoc. 74, 427–431. https://doi.org/10.1080/01621459.1979.10482531.

Drinkall, F., Zohren, S., McMahon, M., Pierrehumbert, J.B., 2025. Stories that (are) move(d by) markets: a causal exploration of market shocks and semantic shifts across different partisan groups. In: Findings of the Association for Computational Linguistics: ACL 2025. Association for Computational Linguistics, Vienna, Austria, pp. 2889–2904. https://aclanthology.org/2025.findings-acl.149/.

Dror, R., Baumer, G., Bogomolov, M., Reichart, R., 2017. Replicability analysis for natural language processing: testing significance with multiple datasets. Transactions of the Association for Computational Linguistics 5, 471–486. https://doi.org/10.1162/tacl_a_00074.

Du, K., Zhao, Y., Mao, R., Xing, F., Cambria, E., 2025. Natural language processing in finance: a survey. Inf. Fusion 115, 102755. https://doi.org/10.1016/j.inffus.2024.102755.

de Oliveira Carosia, A.E., Coelho, G.P., da Silva, A.E.A., 2021. Investment strategies applied to the brazilian stock market: a methodology based on sentiment analysis with deep learning. Expert Syst. Appl. 184, 115470. https://doi.org/10.1016/j.eswa.2021.115470.

Duarte, J.J., Montenegro González, S., Cruz Jr., J.C., 2021. Predicting stock price falls using news data: evidence from the Brazilian market. Comput. Econ. 57, 311–340. https://doi.org/10.1007/s10614-020-10060-y.

Feldhans, R., Wilke, A., Heindorf, S., Shaker, M.H., Hammer, B., Ngonga Ngomo, A.-C., Hüllermeier, E., 2021. Drift detection in text data with document embeddings. In: Intelligent Data Engineering and Automated Learning – IDEAL 2021. Springer International Publishing, Cham, pp. 107–118. https://doi.org/10.1007/978-3-030-91608-4_11.

Fengler, M.R., Phan, M.T., 2024. Unveiling themes in 10-k disclosures: a new topic modeling perspective. SSRN Electron. J. https://doi.org/10.2139/ssrn.4994935.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., Bouchachia, A., 2014. A survey on concept drift adaptation. ACM Comput. Surv. 46, 1–37. https://doi.org/10.1145/2523813.

Garcia, C.M., Abilio, R., Koerich, A.L., Britto, A.d.S., Barddal, J.P., 2025. Concept drift adaptation in text stream mining settings: a systematic review. ACM Trans. Intell. Syst. Technol. 16, 1–67. https://doi.org/10.1145/3704922.

Granger, C.W.J., 1969. Investigating causal relations by econometric models and cross-spectral methods. Econometrica 37, 424–438. https://doi.org/10.2307/1912791.

Granger, C.W.J., Newbold, P., 1974. Spurious regressions in econometrics. J. Econom. 2, 111–120. https://doi.org/10.1016/0304-4076(74)90034-7.

Greco, S., Vacchetti, B., Apiletti, D., Cerquitelli, T., 2025. Unsupervised concept drift detection from deep learning representations in real-time. IEEE Trans. Knowl. Data Eng. 37, 6232–6245. https://doi.org/10.1109/TKDE.2025.3593123.

Hamilton, J.D., 1994. Time Series Analysis. Princeton University Press, Princeton, NJ. https://press.princeton.edu/books/hardcover/9780691042893/time-series-analysis.

Hamilton, W.L., Leskovec, J., Jurafsky, D., 2016. Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1489–1501. https://aclanthology.org/P16-1141/.

Hidalgo, J.I.G., Mariño, L.M.P., de Barros, R.S.M., 2019. Cosine similarity drift detector. In: Artificial Neural Networks and Machine Learning – ICANN 2019: Text and Time Series. Springer, pp. 669–685. https://doi.org/10.1007/978-3-030-30490-4_53.

Holowczak, R., Louton, D., Saraoglu, H., 2019. Testing market response to auditor change filings: a comparison of machine learning classifiers. J. Financ. Data Sci. 5, 48–59. https://doi.org/10.1016/j.jfds.2018.08.001.

Huang, C.C., Shum Nolan, P., 2025. Dumb money? Social network attention herding, sentiment, and markets. J. Financ. Data Sci. 11, 100169. https://doi.org/10.1016/j.jfds.2025.100169.

Kiyama, H., Aida, T., Komachi, M., Ogiso, T., Takamura, H., Mochihashi, D., 2025. Analyzing continuous semantic shifts with diachronic word similarity matrices. In: Proceedings of the 31st International Conference on Computational Linguistics, pp. 1613–1631. https://aclanthology.org/2025.coling-main.109/.

Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S., 2015. Statistically significant detection of linguistic change. In: Proceedings of the 24th International Conference on World Wide Web, pp. 625–635. https://doi.org/10.1145/2736277.2741627.

Kutuzov, A., Velldal, E., Øvrelid, L., 2022. Contextualized embeddings for semantic change detection: lessons learned. Northern European Journal of Language Technology 8. https://aclanthology.org/2022.nejlt-1.9/.

Lancaster, G., Iatsenko, D., Pidde, A., Ticcinelli, V., Stefanovska, A., 2018. Surrogate data for hypothesis testing of physical systems. Phys. Rep. 748, 1–60. https://doi.org/10.1016/j.physrep.2018.06.001.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In: Proceedings of the 31st International Conference on Machine Learning. In: https://proceedings.mlr.press/v32/le14.html.

Levene, H., 1960. Robust tests for equality of variances. In: Olkin, I. (Ed.), Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling. Stanford University Press, Palo Alto, pp. 278–292.

Liang, W., Li, Z., Chen, W., 2024. Enhancing Financial Market Predictions: Causality-Driven Feature Selection. https://arxiv.org/abs/2408.01005.arXiv:2408.01005.

Lin, S.-C., Li, M., Lin, J., 2023. Aggretriever: a simple approach to aggregate textual representations for robust dense passage retrieval. In: Transactions of the Association for Computational Linguistics, vol. 11, pp. 436–452. https://aclanthology.org/2023.tacl-1.26/.

Liu, Y., Medlar, A., Glowacka, D., 2021. Statistically significant detection of semantic shifts using contextual word embeddings. In: Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 104–113. https://aclanthology.org/2021.eval4nlp-1.11/.

Loughran, T., McDonald, B., 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66, 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x.

Loughran, T., McDonald, B., 2020. Textual analysis in finance. Ann. Rev. Financ. Econ. 12, 357–375. https://doi.org/10.1146/annurev-financial-012820-032249.

Lütkepohl, H., 2005. New Introduction to Multiple Time Series Analysis. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-27752-1.
MacKinlay, A.C., 1997. Event studies in economics and finance. J. Econ. Lit. 35, 13–39. https://www.jstor.org/stable/2729691.
Maia, R.d.R., Bravo, J.M., 2026. A new scale for evaluating disclosure in earnings calls on emerging markets. Emerging Science Journal 10, 247–275. https://doi.org/10.28991/ESJ-2026-010-01-012.
Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient estimation of word representations in vector space. arXiv. https://arxiv.org/abs/1301.3781. arXiv:1301.3781.
Montariol, S., Martinc, M., Pivovarova, L., 2021. Scalable and interpretable semantic change detection. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4642–4652. https://aclanthology.org/2021.naacl-main.369/.
Muennighoff, N., Tazi, N., Magne, L., Reimers, N., 2023. MTEB: massive text embedding benchmark. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pp. 2014–2037. https://aclanthology.org/2023.eacl-main.148/.
Panicachi Cocovilo Filho, L.F., Coelho, G.P., 2024. Evaluating the impact of drift detection mechanisms on stock market forecasting. Knowl. Inf. Syst. 66, 723–763. https://doi.org/10.1007/s10115-023-02025-y.
Periti, F., Tahmasebi, N., 2024. A systematic comparison of contextualized word embeddings for lexical semantic change. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). Association for Computational Linguistics, Mexico City, Mexico, pp. 4262–4282. https://aclanthology.org/2024.naacl-long.240/.
Politis, D.N., Romano, J.P., 1994. The stationary bootstrap. J. Am. Stat. Assoc. 89, 1303–1313. https://doi.org/10.1080/01621459.1994.10476870.
Qurashi, A.W., Holmes, V., Johnson, A.P., 2020. Document processing: methods for semantic text similarity analysis. In: Proceedings of the 2020 International Conference on Innovations in Intelligent Systems and Applications (INISTA), pp. 1–6. https://doi.org/10.1109/INISTA49547.2020.9194665.
Ravula, S., 2021. Text Analysis in Financial Disclosures. https://arxiv.org/abs/2101.04480.arXiv:2101.04480.
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. https://aclanthology.org/D19-1410/.
Romano, J.P., Tirlea, M.A., 2022. Permutation testing for dependence in time series. J. Time Anal. 43, 781–807. https://doi.org/10.1111/jtsa.12638.
Rybinski, K., 2020. The forecasting power of the multi-language narrative of sell-side research: a machine learning evaluation. Finance Res. Lett. 34, 101261. https://doi.org/10.1016/j.frl.2019.08.009.
Salton, G., Buckley, C., 1988. Term-weighting approaches in automatic text retrieval. Inf. Process. Manag. 24, 513–523. https://doi.org/10.1016/0306-4573(88)90021-0.
Scaramozzino, R., Cerchiello, P., Aste, T., 2021. Information theoretic causality detection between financial and sentiment data. Entropy 23, 621. https://doi.org/10.3390/e23050621.
Schlechtweg, D., McGillivray, B., Hengchen, S., Dubossarsky, H., Tahmasebi, N., 2020. Semeval-2020 task 1: unsupervised lexical semantic change detection. In: Proceedings of the Fourteenth Workshop on Semantic Evaluation, pp. 1–23. https://aclanthology.org/2020.semeval-1.1/.
Silveira, R.L.F.d., Silva, R.M., Mattos, F.L., Cruz Júnior, Capitani, D.H.D., 2025. The reaction of corn futures markets to US and Brazilian crop reports. J. Futures Mark. 45, 1298–1323. https://doi.org/10.1002/fut.22601.
Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S., 2018. X-vectors: robust DNN embeddings for speaker recognition. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 5329–5333. https://doi.org/10.1109/ICASSP.2018.8461375.
Souza, F., Nogueira, R., Lotufo, R., 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In: Intelligent Systems: 9Th Brazilian Conference, BRACIS 2020. Springer, pp. 403–417. https://doi.org/10.1007/978-3-030-61377-8_28.
Sun, Z., Xiao, C., Harit, A., Yu, J., 2025. Quantifying Semantic Shift in Financial Nlp: Robust Metrics for Market Prediction Stability. https://arxiv.org/abs/2510.00205. arXiv:2510.00205.
Tabari, N., Biswas, P., Praneeth, B., Seyeditabari, A., Hadzikadic, M., Zadrozny, W., 2018. Causality analysis of Twitter sentiments and stock market returns. In: Proceedings of the First Workshop on Economics and Natural Language Processing. Association for Computational Linguistics, Melbourne, Australia, pp. 11–19. https://aclanthology.org/W18-3102/.
Tang, Y., Yang, Y., 2024. Pooling and attention: what are effective designs for LLM-based embedding models? arXiv. https://arxiv.org/abs/2409.02727. arXiv:2409.02727.
Tetlock, P.C., 2007. Giving content to investor sentiment: the role of media in the stock market. J. Finance 62, 1139–1168. https://doi.org/10.1111/j.1540-6261.2007.01232.x.
Thongtan, T., Phienthrakul, T., 2019. Sentiment classification using document embeddings trained with cosine similarity. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pp. 407–414. https://aclanthology.org/P19-2057/.
Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G., 2023. Llama: Open and Efficient Foundation Language Models. https://arxiv.org/abs/2302.13971. arXiv:2302.13971.
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30. In: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.
Wanderley, M.G., Ferraz, L.B.S., Almeida, T.A., Silva, R.M., 2025. A moving target: detecting concept drift in brazilian portuguese fake news. In: Proceedings of the 16th Symposium in Information and Human Language Technology (STIL 2025), pp. 490–501. https://sol.sbc.org.br/index.php/stil/article/view/37849.
Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1, 80–83.
Zerbinati, M.M., Roman, N.T., Di Felippo, A., 2024. A corpus of stock market tweets annotated with named entities. In: Proceedings of the 16th International Conference on Computational Processing of Portuguese, vol. 1. Association for Computational Linguistics, Santiago de Compostela, Galicia, Spain, pp. 276–284. https://aclanthology.org/2024.propor-1.28/.
Zhang, B., Arefyev, N., 2022. The document vectors using cosine similarity revisited. In: Proceedings of the Third Workshop on Insights from Negative Results in NLP. Association for Computational Linguistics, Dublin, Ireland, pp. 129–133. https://doi.org/10.18653/v1/2022.insights-1.17.
Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J., 2025. Qwen3 embedding: advancing text embedding and reranking through foundation models. arXiv. https://arxiv.org/abs/2506.05176. arXiv:2506.05176.