Journal of
Risk and Financial Management

MDPI

Article

Media Sentiment and Reputational Risk in a Public Mortgage Institution: A Daily Index of Media-Derived Reputational Pressure for INFONAVIT

Ana Lorena Jiménez-Preciado, José Carlos Trejo-García *, Miguel Ángel Martínez-García and Francisco Venegas-Martínez

Escuela Superior de Economía, Instituto Politécnico Nacional, Av. Plan de Agua Prieta 66, Miguel Hidalgo, Mexico City 11350, Mexico; ajimenezp@ipn.mx (A.L.J.-P.); mmartinezga@ipn.mx (M.Á.M.-G.); fvenegas1111@yahoo.com.mx (F.V.-M.)
* Correspondence: jtrejog@ipn.mx

Abstract

Reputational risk is hard to quantify, and more so for public institutions in emerging markets. We build a daily index of media-derived reputational pressure for INFONAVIT, Mexico's largest public housing lender, from 138,789 news articles published by 9495 outlets between November 2019 and July 2024. Transformer-based sentiment classification, neural topic modeling, emotion detection and time-series anomaly analysis are combined into a Reputational Risk Score (RRS), bounded between 0 and 100, that measures media pressure, not reputational losses. Agreement between the BERT classifier and the monitoring service's human coders is negligible (Cohen's kappa = −0.0019), so absolute sentiment levels cannot be read; replacing the classifier with two Spanish-language models leaves the index almost unchanged (r = 0.941 and 0.952). Rolling z-scores flag 63 crisis days in 53 episodes over 1223 days with at least 30 articles, and the highest peaks correspond to fraud investigations, operational changes affecting borrowers and incidents at institutional offices. The fraud-related detections survive all 48 detection settings, and the main detections are robust to alternative weights and coverage thresholds, but they require the tone components, and media volume carries 78% of the variance of the composite. Because episodes are interpreted ex post, we present an auditable record of media pressure rather than external validation.

Keywords: reputational risk; media-derived reputational pressure; natural language processing; BERT; sentiment analysis; topic modeling; housing finance; INFONAVIT; emerging markets; media analytics; risk management

Check for updates

Academic Editor: Thanasis Stengos

Received: 29 September 2026
Revised: 21 October 2026
Accepted: 22 October 2026
Published: 27 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Reputational risk, defined as the potential for losses arising from a negative perception by stakeholders that undermines an institution's ability to maintain or establish business relationships (Basel Committee on Banking Supervision [BCBS], 2009), has become one of the most critical risk categories for financial institutions. Unlike market, credit, and operational risks, which rely on established quantitative frameworks and regulatory capital requirements under the Basel Accords, reputational risk is still primarily assessed using indirect qualitative indicators, expert judgment, and ex post event studies (Fiordelisi et al., 2013; Perry & de Fontnouvelle, 2005). This measurement gap is particularly pronounced among public-sector financial institutions in developing countries, where institutional credibil-

ity hinges on the image projected in the media and on public trust, yet where a robust architecture for advanced risk analysis is lacking.

In its 2009 enhancements to the Basel II framework, the BCBS identified reputational risk as a risk behind other risks, stating that "reputational risk is multifaceted and reflects the views of other market participants" (Basel Committee on Banking Supervision [BCBS], 2009). However, the framework did not provide guidance on how to quantify reputational risk, and no major initiative since then has done so. These shortcomings are likely to be even more pronounced in the case of government-sponsored housing finance agencies because they combine policy objectives, financial intermediation, and social services (Eckert & Gatzert, 2017; Gatzert, 2015).

The National Workers' Housing Fund Institute (INFONAVIT, Instituto del Fondo Nacional de la Vivienda para los Trabajadores) provides examples of both the challenges and the stakes involved. INFONAVIT, the largest mortgage lender in Mexico, has over six million current homeowners and total assets of about eight percent of GDP (Monkkonen, 2012, 2019). The institution's hybrid nature, to provide social housing in a fiscally responsible manner, exposes it to reputational risk that has been linked to political support, loss of confidence among workers and reduced access to financial markets. Constant news coverage of governance issues, high rates of housing vacancy and the COVID-19 pandemic have had real effects on the business climate (Reyes, 2021; Monkkonen, 2019).

Recent advances in Natural Language Processing (NLP), and specifically with transformer-based architectures (Vaswani et al., 2017; Devlin et al., 2019), have created new opportunities to measure media sentiment at scale. Sentiment indices from text are useful to predict stock returns (Tetlock, 2007), quantify economic policy uncertainty (Tobback et al., 2018), capture narrative dynamics that drive macroeconomic fluctuations (Shiller, 2017), and quantify corporate social irresponsibility exposure (Kölbel et al., 2017). Most existing applications, however, focus on English-language texts concerning publicly traded firms in developed markets. There are still few applications in Spanish-language media, public-sector institutions, and the housing finance sector.

To fill these research gaps, this study proposes a media-based framework for monitoring media-derived reputational pressure at INFONAVIT. The framework integrates transformer-based sentiment classification, neural topic modeling, emotion detection, media-volume dynamics, crisis-keyword intensity and time-series anomaly detection into a daily Reputational Risk Score. The study contributes in four ways. We first assemble a large-scale corpus of 138,789 Spanish-language news articles from 9495 media outlets, spanning November 2019 through July 2024. We then apply a multi-method NLP pipeline to the corpus that combines multilingual BERT sentiment classification (Devlin et al., 2019), BERTopic neural topic modeling (Grootendorst, 2022), and pysentimiento emotion analysis (Pérez et al., 2021). The pipeline is integrated into a Reputational Risk Score (RRS), a four-component composite index that synthesizes sentiment polarity, media volume, negativity concentration, and crisis-keyword intensity into a single daily metric scaled from 0 to 100. Finally, we apply time-series decomposition and anomaly-detection techniques to identify crisis episodes and describe how media-derived reputational pressure evolves over time.

Although previous studies have measured reputational risk using social media sentiment, market indicators or survey-based approaches, little evidence exists regarding public mortgage institutions operating under governmental mandates. Moreover, existing studies rarely integrate topic modeling, transformer-based sentiment classification and emotion analysis into a unified reputational risk framework. Therefore, this study addresses these limitations by proposing an integrated NLP framework capable of identifying periods of heightened reputational pressure from digital media.

In particular, the study addresses five gaps in the literature:

1. Measurement gap. Quantitative measures of reputational risk rely mainly on market reactions to operational loss announcements (Perry & de Fontnouvelle, 2005; Gillet et al., 2010; Fiordelisi et al., 2013), which cannot be computed for institutions without listed equity.
2. Institutional gap. There is little quantitative evidence on reputational risk in public or government-sponsored housing finance institutions, whose standing depends on political support and public trust as much as on financial performance (Carpenter & Krause, 2012).
3. Language gap. Text-as-data research in finance is dominated by English-language sources (Loughran & McDonald, 2016; Gentzkow et al., 2019), and Spanish-language media remain largely unexplored.
4. Methodological gap. Sentiment, topics, emotions, media volume and crisis vocabulary are rarely combined into a single, transparent and reproducible daily indicator whose sensitivity to weights and detection parameters is documented.
5. Validation gap. Off-the-shelf transformer classifiers are frequently applied without auditing their agreement with human coding or the stability of their bias over time.

As far as we are aware, this is the first time transformer-based NLP methods have been used to build a reputational risk index for a public housing finance institution. The framework contributes both to the academic literature on reputational risk measurement (Fiordelisi et al., 2013; Eckert & Gatzert, 2017) and to the practical toolkit that risk managers in Latin American public financial institutions can draw on.

The paper proceeds as follows. Section 2 reviews the literature on reputational risk, NLP in finance, and INFONAVIT's institutional context. Section 3 describes the data, NLP pipeline, RRS formulation, and validation strategy. Section 4 presents empirical results. Section 5 discusses the findings, their implications and the study's limitations. Section 6 concludes.

2. Literature Review

2.1. Reputational Risk in Financial Institutions

In the financial sector, reputational risk is not a new topic of study. Its conceptualization has changed significantly in the past 20 years; earlier, reputation was viewed as an intangible resource and reputational risk as the resulting loss of market value when things go wrong (Perry & de Fontnouvelle, 2005; Cummins et al., 2006). Perry and de Fontnouvelle (2005) measured reputational risk by splitting the market-value impact of announcements of operational losses into an expected operational-loss part and a residual reputational-loss part. They found that the reputational losses were often bigger than the direct operational losses that caused them. This event-study approach was extended to the insurance sector by Cummins et al. (2006), who found that operational loss events generate negative abnormal returns that are larger than announced loss amounts.

Fiordelisi et al. (2013) identified determinants of reputational risk in banking and found that size, leverage and the character of the operational loss event together predict the size of resulting reputational losses as well as profitability. In a companion paper, Fiordelisi et al. (2014) formalized the relation between operational and reputational risk and defined reputational risk as an amplifier that increases the impact of other risk events.

Eckert and Gatzert (2017) propose an integrated modeling system for operational risk assessment which includes reputation risk. This analysis is a direct response to the Basel Committee on Banking Supervision's (BCBS) observation that reputational risk is multi-dimensional and interlinked with all other risk types. Based on empirical evidence, Gatzert (2015) investigates the link between corporate reputation and financial performance, showing that media sentiment has a significant impact on business performance. Meanwhile,

Aula and Mantere (2020) adopt a strategic management perspective, stressing the influence of the media landscape on organizational communication and the discrepancy between the slow process of reputation formation and the speed with which it can collapse.

The nature of reputational risk has also changed. Digital and social media accelerate the diffusion of adverse information and allow stakeholders to shape organizational reputation in real time rather than through periodic evaluations by intermediaries (Etter et al., 2019), so reputational damage can build up within hours. Studies of banks and insurers show that reputation risk management has become a strategic, enterprise-wide concern, although its measurement remains largely qualitative (Gatzert & Schmit, 2016; Heidinger & Gatzert, 2018). Investors also respond to media-reported environmental, social and governance incidents (Gantchev et al., 2022; Serafeim & Yoon, 2023), which makes timely monitoring of coverage a risk management task rather than only a communications exercise.

Reputational risk is also closely linked to operational uncertainty. Market value losses after operational loss announcements tend to exceed the announced amounts, and the excess is attributed to reputational damage, particularly in cases of internal fraud (Gillet et al., 2010; Sturm, 2013; Fiordelisi et al., 2014). For a housing finance institution, the relevant operational events include internal and external fraud, failures in payment and collection processes, service disruptions and abrupt changes in credit conditions, each of which can generate operational losses and adverse coverage at the same time. For this reason, we treat a media-based index as a complement to, not a substitute for, operational risk indicators.

Still, there has been little written on public financial institutions (such as those owned by the government) or government-sponsored enterprises (GSEs) that provide housing finance. While the literature on reputational risk of banks is rich, most of it focuses on publicly traded banks in developed economies, whereas for public institutions, reputational risk manifests through political support, regulatory treatment, and public confidence in housing policy, in addition to financial metrics (Witzany, 2017). Public administration research describes this as a reputational logic of accountability, in which agencies manage the judgments of multiple audiences to preserve their autonomy and political support (Carpenter & Krause, 2012; E. M. Busuioc & Lodge, 2016; M. Busuioc & Lodge, 2017).

2.2. Natural Language Processing for Financial Risk Assessment

Deep learning and transformer-based models (Vaswani et al., 2017) have allowed computational text analysis to predict financial risks with far greater accuracy. The application of general-purpose sentiment dictionaries may be poorly suited to financial risk assessment because they do not consider the specialized nature of the financial market and the vocabulary associated with that domain (Loughran & McDonald, 2011). Tetlock (2007) finds that pessimism in news coverage, measured by the number of negative words found in a corpus of articles published in The Wall Street Journal, forecasts decreased stock market returns and increased trading volume.

Devlin et al. (2019) introduced BERT (Bidirectional Encoder Representations from Transformers), which was subsequently adopted in the finance literature. BERT is superior to dictionary-based measures, as it relies on bidirectional attention to model context and semantics, allowing for a finer-grained classification of sentiment. Araci (2019) further trained BERT on financial-domain text to obtain FinBERT and showed that it outperforms both dictionary-based methods and the best machine-learning sentiment classifier reported in the literature. Huang et al. (2023) further developed FinBERT for structured information extraction from financial texts.

For Spanish-language texts, Cañete et al. (2020) developed BETO, a BERT model pre-trained on a large Spanish corpus, while Pérez et al. (2022) introduced RoBERTuito, a RoBERTa model pre-trained on 500 million Spanish tweets. The above models are extended

by the pysentimiento toolkit (Pérez et al., 2021) that provides sentiment analysis, emotion detection, and hate-speech detection for Spanish. These advances have made possible large-scale sentiment analysis of Spanish-language media, which is critical for the study of reputational dynamics of Latin American financial institutions.

Topic modeling has evolved from the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) to the recent neural network-based methods. BERTopic (Grootendorst, 2022) combines sentence transformers with UMAP (McInnes et al., 2018) for dimensionality reduction and HDBSCAN (McInnes et al., 2017) for clustering to produce interpretable topic representations. The method is effective with multilingual datasets and consistently outperforms LDA on different benchmarks.

2.3. Media Sentiment Indices and Financial Markets

A growing body of research has built sentiment indices from media content to support economic and financial analysis. For example, Shapiro et al. (2022) constructed a daily news sentiment index for the US and demonstrated that it provides useful information about economic activity that is not captured by traditional consumer sentiment surveys. More broadly, Gentzkow et al. (2019) offer a thorough review of text-as-data methods and observe that these methods are being used more and more in economic research. Bholat et al. (2015) have investigated text mining for central banks and demonstrated how such an approach can be used to support monetary policy and financial stability decisions.

Tobback et al. (2018) constructed an economic policy uncertainty index for Belgium based on a text classification. They find that machine learning approaches perform better than keyword-based approaches. Similarly, Kölbel et al. (2017) found that media coverage of corporate social irresponsibility increases financial risk as measured by credit default swap spreads. Carretta et al. (2011) described evidence of a significant market reaction to press coverage regarding governance in European banking. Deephouse (2000) provided a theoretical framework to understand media reputation as a strategic asset; Engle and Ruan (2019) proposed news-based indicators as early warning signals of systemic risk. The term narrative economics was introduced by Shiller (2017), who argues that major economic events are driven by narratives that go viral.

Existing approaches to measuring reputational risk can be grouped into four families, which Section 5.1 compares with the index proposed here. Market-based event studies infer reputational losses from abnormal returns around operational loss announcements (Perry & de Fontnouvelle, 2005; Gillet et al., 2010; Sturm, 2013); they yield monetary estimates but require listed equity and identified events. Survey- and ranking-based measures capture stakeholder perceptions directly (Fombrun & Shanley, 1990) but are costly and available only at low frequency. Commercial media-screening databases of environmental, social and governance incidents, such as those used by Kölbel et al. (2017) and Gantchev et al. (2022), offer broad firm coverage, but their scoring rules are proprietary and incident-based. Finally, news-based text indices measure policy uncertainty, economic sentiment or climate risk at high frequency (Baker et al., 2016; Engle et al., 2020; Shapiro et al., 2022; Kalamara et al., 2022; Ashwin et al., 2024; Bybee et al., 2024), but they are designed for macroeconomic or market-wide aggregates rather than for a single institution. None of these families provides a transparent, daily, institution-specific indicator for a non-listed public lender, which is the gap the RRS addresses.

Despite its breadth, this literature has rarely addressed non-English-language media, public-sector financial institutions, or the housing finance sector. This study aims to fill this gap by developing a reputational risk index based on the Spanish-language media coverage of a public housing finance institution.

2.4. INFONAVIT: Institutional Context

INFONAVIT, established in 1972, is a tripartite public institution that receives its funding from mandatory contributions by employers equal to 5% of the salaries of their employees. Over the past five decades, it has become Mexico's main source of finance for social housing, originating more than 12 million mortgage loans. By 2024, its active portfolio exceeded six million loans, making it one of the largest mortgage lenders in Latin America.

The literature has widely covered the evolution of Mexico's housing finance system. Monkkonen (2011) argues that the larger mortgage financing access that INFONAVIT and FOVISSSTE provided transformed the country's housing market but also had some unintended consequences such as urban sprawl and longer commutes. Building on this work, Monkkonen (2012) shows that housing finance reforms reinforced socioeconomic segregation by directing lower-income households toward peripheral developments with limited access to employment opportunities and essential public services.

The number of vacant houses is also a concern for the institution in terms of its image in the public eye. Housing vacancy is a potentially important source of reputational exposure for INFONAVIT. Estimates by Monkkonen (2019) indicate that there are almost five million unoccupied housing units in Mexico. Reyes (2021) also connects these vacancy patterns to institutional decision making and their larger effects on urban development. Most recently, Jiménez-Preciado et al. (2025) studied ESG risks in the Mexican mortgage industry and found that reputational performance is an important element for long-term organizational sustainability.

INFONAVIT's impact on the housing sector in Mexico and its contribution to the national economy in general has been significant. However, existing studies have mainly concentrated on housing policy, urban development and mortgage finance. To our knowledge, no previous study has built a systematic indicator of the institution's media-derived reputational pressure using natural language processing (NLP) techniques.

3. Data and Methodology

The methodological approach presented in this section is grounded in a quantitative, longitudinal, and observational research design. This section describes the analytical workflow developed to combine several Natural Language Processing (NLP) techniques with time-series analytics, resulting in a unified Reputational Risk Score (RRS). The unit of textual analysis is the individual news headline referring to INFONAVIT, while the temporal unit used to construct the indicator is the calendar day. The methodological framework operationalizes media-derived reputational pressure as a multidimensional observable construct reflected in the tone, intensity, thematic content, and abnormal concentration of news coverage. Therefore, the proposed Reputational Risk Score should be interpreted as a media-monitoring indicator of changes in the institution's media environment rather than as a direct estimate of financial losses attributable to reputation. We retain the name Reputational Risk Score for continuity with the risk-management literature from which the construct is drawn, but the index is to be read throughout the paper as a measure of media-derived reputational pressure and not as a measure of reputational risk itself.

3.1. Data Description

3.1.1. Data Source and Collection

We retrieved a news content corpus on INFONAVIT from GA Comunicación, a commercial media-monitoring service accessed under an institutional subscription, because news data constitute one of the earliest observable manifestations of reputational pressure, reflecting institutional controversies before visible effects in financial or operational indicators. Accordingly, news headlines were selected as the primary unit of analysis to

construct a forward-looking reputational monitoring framework. The full data were organized into six annual Microsoft Excel files from November 2019 to July 2024, including publication date, outlet name, author, genre, space allocated to the item, geographic origin of the outlet, headline, actor, subject, topic, editorial orientation, information source, and publication type.

The service monitors national, state and local outlets across print, digital, radio and television, delivers a daily file of items mentioning the institution together with a scanned or linked clipping, and assigns each item a topic category and a tone label (Positivo, Neutral, Negativo) through its own human coders. The provider does not publish a codebook and we did not have access to its coding manual, to coder identifiers or to any double-coded subsample, so neither the decision rules behind the tone field nor its inter-coder reliability can be reported. We therefore treat these labels as a second, independent classification rather than as a gold standard, and Section 3.1.2 documents the quality problems we found in the delivered files.

These files were merged into one data set with 143,613 total records for 9495 unique media outlets, 635 topic categories, and 64 states or areas where INFONAVIT operates throughout Mexico.

3.1.2. Data Preprocessing and Quality Assurance

The purpose of text preprocessing is to reduce linguistic variability and preserve semantic information required for transformer-based representation learning. Instead of traditional bag-of-words approaches, transformer models require only limited preprocessing because contextual embeddings retain syntactic and semantic relationships.

Before applying the NLP models, the six annual files were subjected to a structured data audit. To guarantee data accuracy, the audit examined the consistency of the files, and the following multi-step preprocessing pipeline was implemented:

1. Deduplication. We detected and removed duplicate records by means of an exact match on the date of publication, headline and source. This eliminated 4824 duplicate records, resulting in 138,789 articles for our final dataset.
2. Tone label normalization and audit. The tone field of the delivered files does not always contain a tone label. Across the 143,618 delivered records, 21.2% carry a value that is not a tone, and the rate varies strongly by year (0% in 2019, 2020 and 2023; 50.2% in 2021; 19.4% in 2022; 36.5% in 2024; Table A4). Two different problems are mixed in that figure. Most of the 2021 and 2022 cases are feminine spellings of valid labels (Positiva, Negativa) that a naive normalization step would map to Neutral; we recover them, which relabels 15,729 articles (1934 from Neutral to Negative and 13,793 from Neutral to Positive). The remaining 13,846 records (10.0% of the corpus, almost all in 2024) contain topic strings such as Reforma pensiones in the tone field, a column shift we cannot repair; these are treated as missing, and C_3 is computed only over articles with a valid label. The corrected distribution of the 124,943 valid labels is Positive 50.7%, Neutral 44.6% and Negative 4.7%.
3. Date standardization. Publication dates, delivered in two formats (with and without a time stamp), were parsed to ISO 8601 format. Records with missing or unparseable dates were excluded from time-series analyses but retained for cross-sectional analyses.
4. Text cleaning. Headlines were cleaned lightly, to preserve their semantic content (just whitespace, UTF-8, special characters). No stemming or lemmatization was applied, since BERT models do not require it.
5. Coverage audit. Monthly article counts revealed an inconsistent delivery between June and November 2020: the June–August records were delivered in a different format and September–November contain only 524 articles (521, 2 and 1). We exclude this

block (4300 articles) from all time-series analyses. A day enters the index only if it has at least 30 articles; this removes 301 low-coverage days, including October 2022, for which the provider delivered only 130 articles. The daily index is therefore computed on 1223 days covering 129,644 articles, and results with minimum thresholds of 10 and 1 articles per day are reported in Section 4.8.

3.1.3. Ethical Considerations and Data Availability

The data consist of published news items compiled by a commercial media-monitoring service; no personally identifiable data were used. Media monitoring data were gathered via an institutional subscription and used consistently with the provider's contract terms. Raw data will not be provided due to contractual obligations, although all code and aggregated results are publicly available at https://github.com/AnaJZP/infonavit-sentiment (accessed on 21 October 2026) (see the Data Availability Statement). No ethics approval was required for this study as human subjects were not involved.

3.2. NLP Pipeline

All preprocessing and NLP analyses were implemented in Python (version 3.10). The main libraries were transformers, torch, sentence-transformers, BERTopic, UMAP-learn, HDBSCAN, pysentimiento, pandas, and statsmodels. All code is openly available at https://github.com/AnaJZP/infonavit-sentiment. The analysis relies on three complementary NLP tools, each addressing a different aspect of media coverage.

3.2.1. Sentiment Analysis: Multilingual BERT

The sentiment-classification stage estimates the overall polarity associated with each news headline. The transformer architectures learn contextual semantic relationships that help to alleviate the classification errors caused by sarcasm, negation and domain-specific terminology. The sentiment polarity was computed with the nlptown/bert-base-multilingual-uncased-sentiment model, a BERT variant trained for sentiment classification in six languages including Spanish (Devlin et al., 2019). We chose this model over Spanish-only alternatives (e.g., BETO) for three reasons: (i) it outputs a five-point ordinal scale (1 to 5 stars) that captures sentiment intensity; (ii) it was trained on product and service reviews, a register in which evaluative language is explicit; and (iii) it handles the occasional English-language articles in the corpus. Section 4.2 shows that this choice carries a domain shift that the validation exercise makes plain, and Section 5.1 reports how far the results depend on it.

The ordinal output was mapped to a ternary classification: Negative (1 or 2 stars), Neutral (3 stars), Positive (4 or 5 stars). The underlying score was also rescaled to the interval $[-1, +1]$ as $s_i = \frac{(stars_i - 3)}{2}$ for use in the continuous RRS computation. Inference was performed on headlines, as these represent the most concentrated expression of editorial stance. Processing of the 138,789 headlines required approximately 4 h on a single NVIDIA GPU.

Because this model was not trained on Spanish news, we also classified the full corpus with two Spanish-language models, BETO (finiteautomata/beto-sentiment-analysis) and RoBERTuito (pysentimiento/robertuito-sentiment-analysis), and rebuilt the index with each of them in place of C_1. Section 4.2 reports the comparison. Fine-tuning a model on institutional Spanish news would require an annotated corpus that does not exist for this domain and that was beyond the scope of this study; we state it as the first priority for future work.

3.2.2. Topic Modeling: BERTopic

To identify the thematic structure of INFONAVIT media coverage, we used BERTopic (Grootendorst, 2022), which combines: (1) sentence embeddings via the paraphrase-multilingual-MiniLM-L12-v2 sentence transformer; (2) dimensionality reduction using UMAP (McInnes et al., 2018), configured with n_neighbors = 15, n_components = 5, min_dist = 0.0, and cosine metric; (3) clustering via HDBSCAN (McInnes et al., 2017) with min_cluster_size = 100 and Excess of Mass selection; and (4) topic representation via class-based TF-IDF with custom Spanish stop words. The min_topic_size was set to 100 to balance granularity against interpretability. BERTopic's dynamic topic modeling was used to track topic prevalence over time.

3.2.3. Emotion Analysis: Pysentimiento

To capture affective dimensions that polarity alone misses, we used the pysentimiento toolkit (Pérez et al., 2021), which uses RoBERTuito (Pérez et al., 2022) to classify text into joy, sadness, anger, surprise, disgust, and fear. The daily distributions of emotions were aggregated to highlight the periods of increased negative affect. Even when overall sentiment remains close to neutral, such emotional shifts can act as early indicators of reputational pressure.

3.3. Reputational Risk Score (RRS) Formulation

To capture the several dimensions along which media coverage of an institution can vary, we construct a composite Reputational Risk Score (RRS) that aggregates four complementary dimensions of media coverage into a single daily indicator. Let t be the day of observation and A_t the set of news articles published on day t, such that $n_t = |A_t|$ is the total number of articles available on day t.

Component 1: Sentiment Risk (C_1). Daily sentiment risk inverts the average continuous sentiment score:

$$C_{1,t} = 1 - \frac{1}{n_t} \sum_{i \in A_t} \frac{(s_i + 1)}{2} \quad (1)$$

where s_i lies in $[-1, +1]$. More negative sentiment produces a higher risk score.

Component 2: Volume Anomaly (C_2). This component captures abnormal media attention through z-score normalization:

$$C_{2,t} = \Phi\left(\frac{n_t - \mu_n}{\sigma_n}\right) \quad (2)$$

where μ_n and σ_n are the mean and standard deviation of the daily article count over the estimation sample and Φ is the standard normal cumulative distribution function, so days with above-average coverage raise the score.

Component 3: Negativity Ratio (C_3). The measure is the share of negative articles:

$$C_{3,t} = \frac{n_t^{(-)}}{n_t} \quad (3)$$

where $n_t{}^{(-)}$ is the number of articles published on day t that the monitoring service's human coders labeled Negative, and the denominator is the number of articles of that day carrying a valid tone label (Section 3.1.2). C_3 therefore draws on a classification that is independent of the BERT model used in C_1.

Component 4: Crisis Keyword Intensity (C_4). This measures the proportion of news articles that contain pre-defined terms related to crises (the 30 Spanish terms listed in Appendix A, e.g., fraud, corruption, scandal, complaint, lawsuit, irregularity, eviction, seizure, impunity, investigation, fine, loss, debt and delinquency). Matching is case-insensitive

substring matching on headlines. The generic term demanda was excluded because in this corpus it mostly refers to demand for housing or credit, and was replaced by demanda judicial, demandado and demandan. Let K denote the resulting set of terms (Table A1) and I[·] the indicator function:

$$C_{4,t} = \frac{1}{n_t} \sum_{i \in A_t} I[\exists\, w \in K : w \in headline_i] \quad (4)$$

The composite index combines the four components with weights:

$$RRS_t = 100 \times (0.35C_1 + 0.25C_2 + 0.25C_3 + 0.15C_4) \quad (5)$$

The weighting scheme captures the relative importance of each dimension in the measurement of media-derived reputational pressure. Sentiment polarity is the most direct indicator of media tone (Tetlock, 2007; Shapiro et al., 2022) and thus is given the largest weight (0.35). We assign a weight of 0.25 to the volume anomaly, acknowledging the potential that high media coverage may be an indirect proxy for higher media-derived pressure (Kölbel et al., 2017; Deephouse, 2000). The share of explicitly negative articles, i.e., the negativity ratio, is also weighted 0.25, as it complements the continuous measure of sentiment (Loughran & McDonald, 2011). Finally, crisis-related keywords receive a weight of 0.15, on the assumption that crisis-specific language helps identify emerging reputational events (Engle & Ruan, 2019). Because each component lies in [0, 1] and the weights sum to one, the RRS is bounded between 0 and 100 (0 = no media-derived reputational pressure; 100 = the theoretical maximum). No min–max normalization is applied to the components or to the composite, so the index stays on a fixed 0–100 scale and its values are comparable across days within a given specification; in our sample the index ranges from 17.92 to 58.68. Values obtained under different weighting schemes are not directly comparable in level, since each scheme defines a different composite. Section 4.8 therefore compares specifications through their correlation with the baseline, the overlap of their crisis days and the episodes they detect, rather than through the numerical levels of the index. Because the weights are theory-driven, Section 4.8 reports equal, sentiment-heavy, volume-heavy, sentiment-excluded and PCA-based alternatives. Two caveats follow from the construction, and we state them here rather than in the robustness section. First, C_1 and C_3 both summarize tone, from two different classifications, and the crisis lexicon in C_4 shares vocabulary with fraud-related coverage, so the four components are not independent measurements; 84.8% of the headlines matched by the lexicon are also classified as negative by BERT. Second, because C_2 is a normal CDF of a highly variable count while the other components are bounded shares with small daily variation, the volume component accounts for 78% of the variance of the composite, against 13% for C_1 and 4% for C_3. The RRS is therefore dominated by abnormal attention, and we recommend reading it together with the four component series rather than as a single calibrated risk measure.

3.4. Time-Series Analysis

The RRS was decomposed using Seasonal-Trend decomposition based on Loess (STL) to separate the time series into trend, seasonal, and residual components. The monthly cycle was represented by a 30-day seasonal period, and a 7-day period is reported as a robustness check. Because STL requires an evenly spaced series, days removed by the coverage filters (Section 3.1.2) are linearly interpolated for the decomposition only; they are not used for crisis detection:

$$RRS_t = T_t + S_t + R_t \quad (6)$$

Crisis episodes are identified through a rolling z-score approach:

$$z_t = \frac{(RRS_t - \mu_{t,w})}{\sigma_{t,w}} \quad (7)$$

computed over a trailing window of w = 14 days. The mean and standard deviation are computed over the w preceding valid observations and exclude the current day, so the score is not shrunk by the observation it evaluates. A day is flagged when $z_t > 2$. This threshold is a convention and not a distributional claim: the realized z-scores are right-skewed (skewness 0.73, excess kurtosis 1.84) and autocorrelated (first-order autocorrelation 0.293), so 5.21% of days exceed 2 instead of the 2.28% implied by normality. Section 4.8 reports the empirical quantile of the realized distribution, a robust version based on the median and the median absolute deviation, a window defined in calendar days and a change-point alternative. Adjacent flags within 3 days are merged into contiguous episodes. The window is measured in valid observation days. These parameters are conventional rather than optimized, so Section 4.8 reports all 48 combinations of windows $w \in \{7, 14, 21, 30\}$, thresholds $\tau \in \{1.5, 2.0, 2.5, 3.0\}$ and merge gaps $g \in \{1, 3, 5\}$.

3.5. Validation Strategy

The sentiment classifier was evaluated against manually annotated headlines: we compare the BERT sentiment predictions with the cleaned manual labels, reporting Cohen's kappa, overall accuracy, a 3 × 3 confusion matrix, and per-class precision, recall, and F1. Because those manual labels came from human coders of unknown reliability, we interpret disagreement between BERT and the labels as potentially reflecting noise in the manual process rather than error on BERT's part alone. The exercise thus measures how far the two approaches correspond and stops short of elevating the manual labels to a gold standard.

In addition to sentiment analysis, BERTopic was used to identify recurring themes in the coverage. These themes describe what the media wrote about when coverage intensified; topic modeling groups texts by lexical and semantic similarity and does not establish the institutional causes of that coverage or the sources of variation in the index. The performance of the model was evaluated using topic coherence (NPMI), representative documents and expert judgment.

Assessing the index itself raises a circularity problem: a media-based indicator is expected to rise when an event receives heavy coverage, so the coincidence of its peaks with publicized events cannot be taken as external validation. We therefore assess the RRS in three ways that do not rely on that coincidence. First, we examine how the classifier's bias behaves over time, since a bias that moves slowly relative to the detection window is the condition under which a biased sentiment score can still track relative changes. Second, we audit the content of detected episodes with information that does not enter the index: a reading of the most repeated headlines on each peak day and the BERTopic topics of the articles. Third, we examine the stability of detections under alternative weights, detection parameters, minimum-coverage thresholds, outlet subsamples and the removal of syndicated duplicates (Section 4.8).

4. Results

4.1. Descriptive Statistics

The neutral level of coverage (44.6% of the articles with a valid tone label) is representative of the type of media coverage on public institutions, where a large portion of news articles are informational content on INFONAVIT's housing programs, mortgage announcements and policy updates. The relatively small amount of negative coverage (4.7%), by contrast, should be read with caution. This pattern can reflect the overall tone of the media, but it can also be a function of the classification criteria of the monitoring service. One possible explanation is that a relatively strict threshold was chosen for classifying negative sentiment. The distribution of coverage across 9495 media outlets further demonstrates the

extensive public visibility of INFONAVIT at local, regional and national levels (Table 1 and Figure 1).

Table 1. Descriptive statistics of the INFONAVIT media corpus (November 2019 to July 2024).

Variable	Value	Notes
Total articles (after deduplication)	138,789	
Duplicates removed	4824	Exact match on date + headline + outlet
Unique media outlets	9495	
Unique topics (monitoring service)	635	
States/regions	64	Including national-level outlets
Coverage period	Nov 2019 to Jul 2024	57 months
Mean daily articles	84.2	
Median daily articles	66	Right-skewed distribution
Manual tone: Neutral	55,742	44.6% of valid labels
Manual tone: Positive	63,324	50.7% of valid labels
Manual tone: Negative	5877	4.7% of valid labels
Records without a valid tone label	13,846	10.0% (topic strings in the tone field)

Note: Tone labels are those assigned by the media monitoring service's human coders, not by the BERT classifier. Tone percentages are computed over the records with a valid tone label after the cleaning described in Section 3.1.2; records whose tone field contained topic strings (13,846, 10.0%) are treated as missing, and 15,729 feminine spellings of valid labels were recovered (Section 3.1.2 and Table A4). The June to November 2020 delivery is excluded for the reasons given in Section 3.1.2.

The wide and heterogeneous media corpus contributes to the empirical basis of the proposed Reputational Risk Score. Because the sentiment indicator aggregates thousands of media outlets from different geographical regions and editorial positions, no single outlet drives the daily series. We do not claim that the breadth of the corpus removes reporting bias more generally. Where this can be tested, Section 4.8 points the other way: restricting the corpus to the twenty largest outlets lowers the correlation with the baseline index to 0.640 and recovers two of the five highest-peak episodes, and removing syndicated duplicates recovers four of five. Aggregation limits the influence of any one outlet, but the composition of the corpus does affect the results.

4.2. Sentiment Analysis Results and Validation

When we applied the multilingual BERT model to the 138,789 headlines, the resulting ternary distribution was Negative = 77,466 (55.8%), Neutral = 8602 (6.2%), and Positive = 52,721 (38.0%). The resulting continuous score had a mean of −0.13, a median of −0.292, and a standard deviation of 0.436. Table 2 compares both classifications over the 124,943 articles with a valid manual label.

Table 2. Comparison between BERT predicted sentiments and manual tone labels.

Category	Manual *n* (%)	BERT *n* (%)	Difference (pp)
Negative	5877 (4.7%)	70,884 (56.7%)	+52.0
Neutral	55,742 (44.6%)	7754 (6.2%)	−38.4
Positive	63,324 (50.7%)	46,305 (37.1%)	−13.6

Note: Both distributions are computed over the 124,943 articles with a valid manual tone label after the recovery described in Section 3.1.2. The difference column reports percentage points (BERT minus manual) and measures disagreement between two classifications rather than the error of either one, since neither the manual labels nor the BERT predictions are treated as a gold standard (Section 4.2).

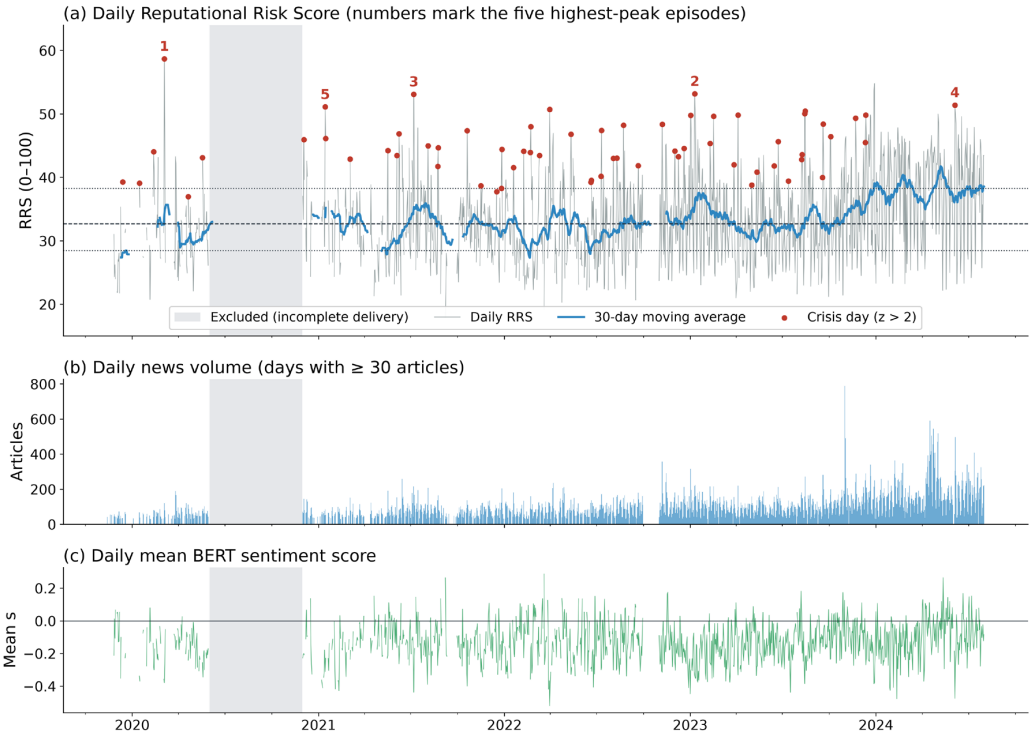


Figure 1. Daily Reputational Risk Score (RRS) for INFONAVIT, November 2019 to July 2024, computed on days with at least 30 articles. (**a**) Daily RRS (gray), 30-day moving average (blue), 25th, 50th and 75th percentiles (dotted and dashed lines) and crisis days with $z > 2$ (red); numbers identify the episodes in Section 4.5. (**b**) Daily number of articles. (**c**) Daily mean BERT sentiment score. The shaded area marks the excluded June–November 2020 delivery, and gaps correspond to days below the coverage threshold.

The BERT predictions differ substantially from the manual labels. Table 3 and Figure 2 present the corresponding confusion matrices, and Figure 3 shows the distribution of the continuous BERT scores within each manual tone category.

Table 3. Confusion matrix of manual labels (rows) and BERT predictions (columns).

	BERT Negative	BERT Neutral	BERT Positive	Total (Manual)
Manual Negative	4317	239	1321	5877
Manual Neutral	28,848	3836	23,058	55,742
Manual Positive	37,719	3679	21,926	63,324
Total (BERT)	70,884	7754	46,305	124,943

Note: Rows are the manual labels assigned by the monitoring service's human coders and columns are the BERT predictions. The percentages quoted in the text are row-normalized, so they give the share of articles carrying a given manual label that BERT assigned to each class. The matrix records disagreement between two classifications and is not an accuracy assessment.

Over the articles with a valid manual label, Cohen's kappa between the two classifications is −0.0019, overall agreement is 24.1% and the macro F1 is 0.211; per-class F1 scores are 0.112 for Negative, 0.121 for Neutral and 0.400 for Positive. Agreement is thus negligible. The interpretation of these numbers follows Cohen (1960) and Landis and Koch (1977), with the caveat that neither classification is a reference standard.

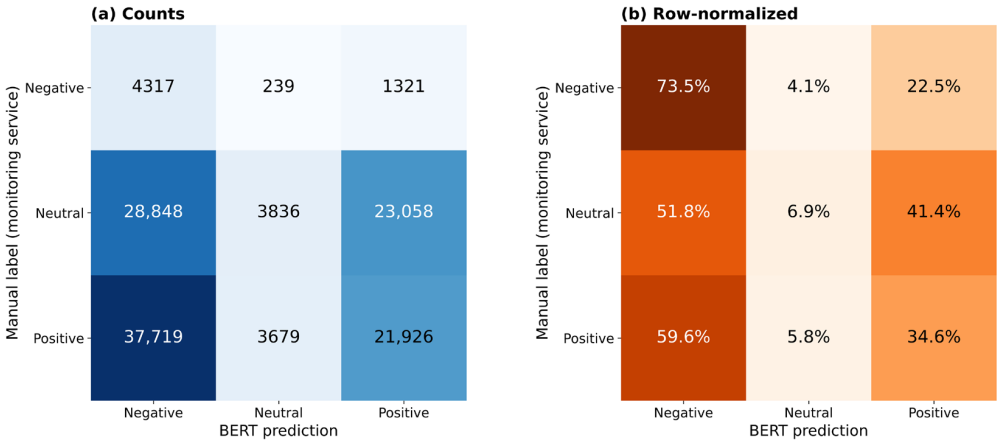


Figure 2. Confusion matrix of BERT predictions on sentiments vs. manually annotated labels: (**a**) counts and (**b**) row-normalized shares. Rows are manual labels and columns are BERT predictions.

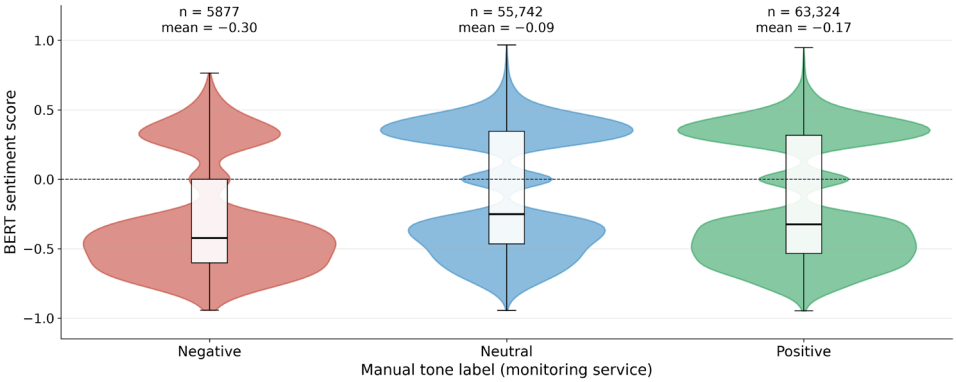


Figure 3. Distribution of the continuous BERT sentiment score within each manual tone category (articles with a valid label, $n = 124{,}943$). The box inside each violin shows the median and interquartile range. The three distributions overlap almost completely, which is the distributional counterpart of the disagreement reported in Table 3.

This level of disagreement is far larger than what most NLP validation studies report (Shapiro et al., 2022; Araci, 2019), and it calls for a careful reading rather than a simple accuracy verdict. Two characteristics of the validation setting may explain this discrepancy. The model we use was fine-tuned on product and service reviews, a register in which explicit evaluation is the norm; carried over to institutional news headlines, it tends to read routine or administrative phrasing as mildly critical, which pushes a large share of items into the negative class. The manual labels, in turn, came from a monitoring service whose coders appear to have reserved the negative tag for overtly adverse coverage. The dominant pattern in the confusion matrix exemplifies both mechanisms: BERT assigns to Negative 51.8% of the articles labeled Neutral by the coders and 59.6% of those labeled Positive. We therefore treat neither classification as ground truth and, more consequentially for what follows, do not read the absolute level of the sentiment component. What the RRS relies on is how sentiment moves over time relative to its own recent history, and that relative signal is far less sensitive to a constant classification offset than the raw negative share is. Whether the offset is in fact constant is an empirical question that we test directly. Across the 51 months with data, the share of headlines that BERT labels negative (Figure A1) averages 56.7% (SD 5.4 percentage points) with no significant trend (slope −0.05 pp per month, $p = 0.31$, $R^2 = 0.02$). Among headlines that the human coders did not label negative, the BERT negative share is also trendless (slope +0.02 pp per month, $p = 0.70$). The bias is therefore large and shows no linear drift. That is not the same as stability. The monthly

share still varies by several percentage points around its mean, and a nonsignificant slope cannot rule out changes that are not linear. Nor do we claim that the classifier gives a valid measure of sentiment for Mexican institutional news; showing that would require fine-tuning and evaluating it on an annotated sample of INFONAVIT headlines. What limits the effect of this variation on the results is the alert rule. Each day is compared with its own trailing 14-day window, so a change in the bias that builds up over weeks or months is absorbed by the window mean and does not by itself produce alerts. A change that happens within a few days, for example when coverage turns to a topic the model reads poorly, is not absorbed, and we treat it as a limitation (Section 5.3).

Two Spanish-language models provide a further check. Classifying the whole corpus with BETO and with RoBERTuito gives negative shares of 15.2% and 25.5%, against 55.8% for the multilingual model and 4.7% for the human coders. The Spanish models are therefore much closer to the coders in level, which is consistent with the domain shift in a model trained on product reviews, but they do not agree with the coders either (kappa 0.025 for BETO and −0.0034 for RoBERTuito), and they agree only moderately with each other (kappa 0.411). The disagreement is thus a property of the task and of the labels, not only of one model.

What matters for this paper is whether the index depends on that choice. Rebuilding C_1 with each Spanish model and leaving everything else unchanged gives series that correlate 0.941 (BETO) and 0.952 (RoBERTuito) with the baseline, detect the same five highest-peak episodes and flag 64 and 66 crisis days against 63 in the baseline (Table A5). The level of the sentiment component changes with the model; the episodes it helps detect do not.

While the overall sentiment categorizations diverge substantially from the human assessments, the aim of this research is not to replicate human coding but to observe temporal shifts in the media landscape. In risk rating, comparative changes in ratings are far more useful than absolute sentiment scores. Hence, the practical value of the proposed framework lies in detecting abnormal deviations in coverage whose content can then be audited (Section 4.5).

4.3. Topic Modeling Results

Besides the descriptive clustering, the identified topics also contribute to the understanding of the overall Reputational Risk Score by showing which themes dominate the coverage. Rather than just evaluating changes in the tone of the media, BERTopic identifies recurring themes in the coverage associated with such changes, including governance, fraud, labor reforms and housing policy. This thematic breakdown makes the framework more relevant in practice, as risk management practitioners can tell apart operational, political, regulatory and social themes that recur when media pressure rises. BERTopic identified 79 distinct topics from the 138,789 headlines, with 69,497 articles (50.1%) assigned to the outlier topic (−1), a pattern commonly observed when BERTopic is applied to short texts. The remaining 69,292 articles (49.9%) were assigned to coherent topical clusters. Table 4 presents the top 15 topics.

Three features of this solution qualify how much weight the topic layer can carry. Half of the corpus (50.1%) is assigned to the outlier cluster, one topic (Housing Credits) absorbs 62.3% of the classified articles, and the average NPMI coherence of the fifteen largest topics is −0.047 (from −0.314 to 0.279), which is low and reflects the short, formulaic nature of headlines. We also did not re-estimate the model under alternative seeds and hyper-parameters, so topic stability is untested. We therefore use the topics descriptively, to characterize what the coverage of a detected episode was about, and not as measurements entering the index.

Table 4. Top 15 BERTopic topics by number of articles.

Topic	Label	Top Keywords	*n*	% Neg.
0	Housing Credits	pesos, credito (credit), vivienda (housing), pensiones (pensions)	43,195	60%
1	Hurricane Otis	otis, acapulco, reconstruccion (reconstruction), guerrero	1482	48%
2	Labor Reform	reforma (reform), empresas (companies), ilegal (illegal), ley (law)	1402	54%
3	Fraud and Alerts	fraude (fraud), alerta (alert), advierte (warns), evitar (avoid)	1179	94%
4	Social Security	imss, hogar (home), real (real), diciembre (December)	1039	59%
5	Women's Access	mujeres (women), acceso (access), creditos (credits), programa (program)	963	39%
6	Elections 2024	sheinbaum, claudia, xochitl, galvez	933	32%
7	Urban Development	urbano (urban), desarrollo (development), ciudades (cities)	889	11%
8	Labor Market	laboral (labor), empleo (employment), trabajo (work), trabajadores (workers)	848	41%
9	Institutional Crisis	presidencial (presidential), quiebra (bankruptcy), amlo	655	72%
10	Union Activity	sindicato (union), lider (leader), sindical (union-related), sede (headquarters)	632	78%
11	Electoral Politics	politico (political), candidatos (candidates), morena	605	42%
12	Business Recognition	empresas (companies), diez (ten), reconoce (recognizes)	592	58%
13	Regional Governance	estatales (state-level), municipales (municipal), yucatan	592	43%
14	Consumer Queries	quien (who), pasa (happens), caso (case), que (what), seguro (insurance)	583	82%

Note: Topics are BERTopic clusters and the negative shares in the last column are based on BERT predictions, so they are not comparable with the shares in Figure 4, which come from the monitoring service's own topic categories and its human tone labels. The 69,497 articles assigned to the outlier topic (−1) are excluded. Topic labels were assigned by the authors from the top keywords and representative documents; they describe what the coverage was about and do not identify the institutional causes of that coverage.

The recovered themes map onto recognizable facets of the institution (Figure 4). The Housing Credits topic is by far the largest cluster, and although 60% of its articles are negative, this in part reflects that any mention of the credit system tends to be framed critically in the coverage analyzed. A second, governance-oriented group of clusters covers labor reform, fraud, and institutional crisis (Topics 2, 3, and 9). The Fraud and Alerts topic exhibits the highest proportion of negative coverage (Topic 3, 94%), whereas urban development and women's access to housing credit (Topics 7 and 5) draw the least (11% and 39%, respectively). One point of presentation deserves note: Figure 4 does not show BERTopic clusters but the topic categories assigned by the monitoring service, several of which correspond to specific INFONAVIT programs and instruments (such as Mejoravit and Shared Responsibility), together with their manual tone labels and ranked by negative share. Table 4, in contrast, reports BERTopic clusters with BERT-based negative shares. The two views are complementary rather than identical, which is why the topic names and negative shares differ between them.

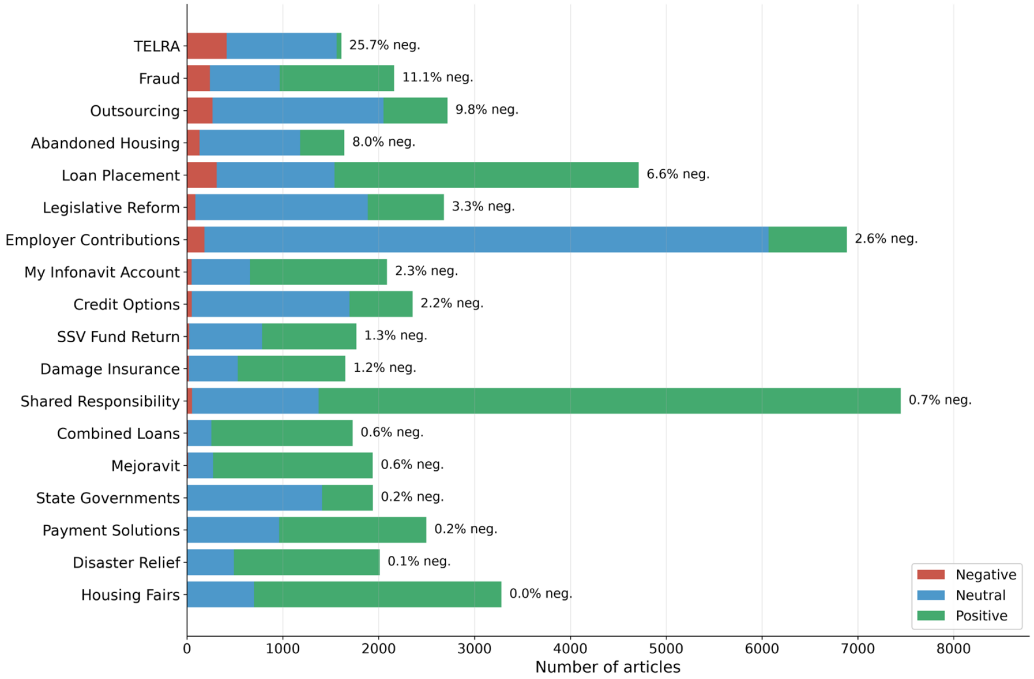


Figure 4. Manual tone breakdown (negative, neutral, positive) for the twenty most frequent topic categories assigned by the monitoring service, ordered by negative share (shown next to each bar). These provider categories differ from the BERTopic clusters in Table 4.

4.4. Reputational Risk Score Time Series

This section characterizes the temporal behavior of the composite index, beginning with its overall distribution before turning to the seasonal-trend decomposition and the detection of crisis episodes. Summary statistics for the daily RRS over the whole study period are shown in Table 5 and its monthly evolution in Figure 5.

Table 5. Reputational Risk Score (RRS), summary statistics (days with at least 30 articles).

Statistic	Value
Mean	33.56
Median	32.66
Std. Dev.	6.77
Min	17.92
Max	58.68
Skewness	0.453
Excess kurtosis	−0.334
25th percentile	28.43
75th percentile	38.26
95th percentile	45.48
Crisis days ($z > 2$)	63
Crisis episodes (merged)	53
Days in the index (≥30 articles)	1223

Note: Statistics are computed over the 1223 days with at least 30 articles; days below that threshold and the excluded June to November 2020 period do not enter. Because the components are not min–max normalized, the values lie on a fixed 0 to 100 scale and are comparable across days within this specification. They are not comparable in level with the values obtained under the alternative weighting schemes in Table A2, since each scheme defines a different composite.

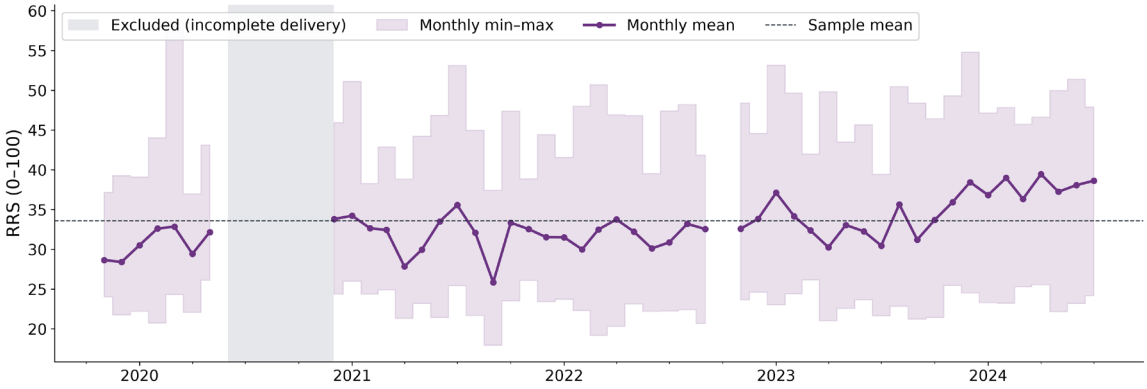


Figure 5. Monthly Reputational Risk Score (RRS) for INFONAVIT, November 2019–July 2024: monthly mean of daily values (line), monthly minimum–maximum range (band) and sample mean (dashed line). The shaded area marks the excluded June–November 2020 delivery; months without valid days are left blank.

Over the 1223 valid days the RRS averaged 33.56 (SD 6.77) and ranged from 17.92 to 58.68. The distribution is moderately right-skewed (0.453) with slightly negative excess kurtosis (−0.334): high values are more frequent than low ones, but the tails are not heavier than those of a normal distribution. Spikes are therefore identified relative to recent history (Section 3.4) rather than from the unconditional distribution. The yearly mean rises from 28.5 in 2019 to 37.9 in 2024, mainly through the volume component. Because C_2 standardizes daily counts against the full-sample mean, the growth of coverage lifts its average from 0.22 in 2019 to 0.71 in 2024, while the average of C_3 falls from 0.11 in 2020 to 0.02 in 2024. Part of this late increase reflects the monitored universe rather than the institution: the number of distinct outlets delivering at least one article per month grows from 260 in November 2019 to 1347 in July 2024 (+21.45 outlets per month, $p < 0.001$). Section 4.8 rebuilds C_2 with past-only baselines and with a per-outlet normalization to show how much of the level shift survives. The trailing z-score used for detection is not affected by these level shifts (Figure 5).

The monthly profile shows long stretches close to the sample mean, punctuated by months that contain isolated high-risk days. The rise after mid-2023 should be read together with the growth in coverage volume described above rather than as a deterioration of INFONAVIT's reputation per se. Reputational pressure in this series therefore appears as short-lived spikes superimposed on a slowly moving level, which motivates combining trend decomposition with local anomaly detection.

4.4.1. STL Decomposition

With a 30-day period (Figure 6), the trend component ranges from 27.09 to 41.16. It stays close to 29–31 through 2022 (yearly means of 28.9 in 2019, 31.0 in 2020, 31.0 in 2021 and 30.6 in 2022), rises to 33.5 in 2023 and peaks in February 2024 (41.16), with a 2024 mean of 39.0; as noted above, this late increase mirrors the growth in coverage. The seasonal component ranges from −15.12 to 16.87 and captures within-month regularities in publication. The largest residual (32.47) occurs on 4 March 2020, the day of the fraud-investigation coverage discussed in Section 4.5. The variances of the trend, seasonal and residual components equal 30.4%, 24.9% and 75.4% of the observed variance; the shares do not add up to one because STL components are not orthogonal. With a 7-day period, the trend range (24.25 to 42.91), the date of the largest residual and the late-sample rise are unchanged, while the residual share falls to 47.7%. The interpolated segments (500 of 1723 calendar days, including the excluded 2020 block) produce a smooth trend by construction and should not be interpreted. With close to 29% of calendar days interpolated, we read the trend only

as a descriptive summary of the observed segments and draw no substantive conclusion from its shape.

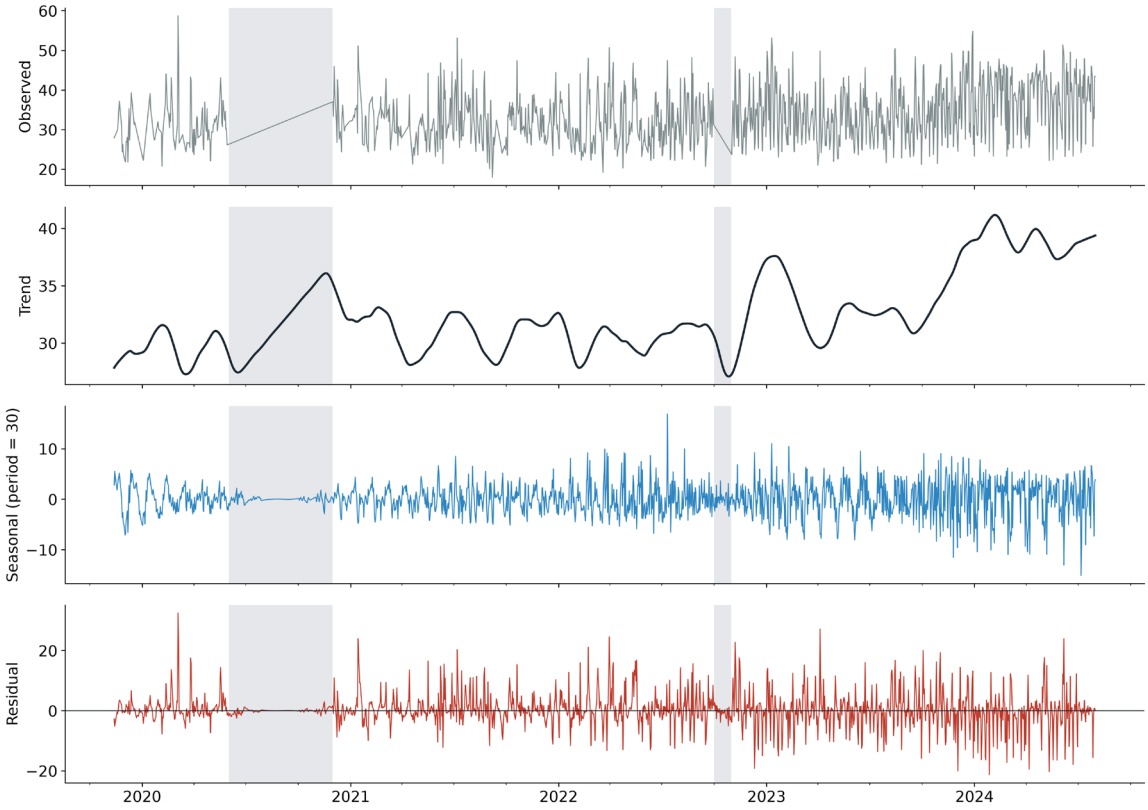


Figure 6. STL decomposition of the daily RRS (period = 30 days). From top: observed RRS, trend, seasonal component and residual. Shaded areas mark interpolated blocks (June–November 2020 and October 2022).

The STL decomposition separates three features of the media-derived index. The trend component describes its slow evolution, which in this sample largely follows the growth in coverage volume rather than any change in the institution's reputation. The seasonal component reflects regularities in publication. The residual component isolates abrupt departures of the index from its recent level; the alerts in Section 4.5 fall on such departures, but the residual does not by itself identify crises or separate them from other sources of intense coverage.

4.4.2. Component Correlation Analysis

We also examined the Pearson correlations among the four components as a check on collinearity, that is, to see whether any component is close to a linear function of another and would therefore add little to the composite. The corresponding correlation matrix is shown in Table 6 and graphically represented as a heatmap in Figure 7.

The four components are weakly correlated. The strongest positive association is between sentiment risk and crisis keywords ($r = 0.274$), followed by negativity and crisis keywords ($r = 0.216$) and sentiment risk and negativity ($r = 0.137$). The last correlation is informative because C_1 (BERT) and C_3 (human coders) come from independent classifications, and their positive association indicates that the two sources move in the same direction despite their disagreement in levels. Volume is nearly uncorrelated with sentiment risk and crisis keywords ($|r| \leq 0.029$) and negatively correlated with negativity ($r = -0.235$), so it contributes largely separate information.

The correlations among the four components are weak, so none of them is close to a linear function of another. This result is limited in scope. Low Pearson correlations indicate weak linear association between the component series; they do not show that the components carry non-overlapping information, nor that they capture distinct dimensions of reputational risk. What they do indicate is that the composite is not driven by one signal counted several times, which is a precondition for combining several indicators rather than relying on sentiment polarity alone.

Table 6. Pearson correlation matrix of RRS components.

	C_1 (Sentiment)	C_2 (Volume)	C_3 (Negativity)	C_4 (Keywords)
C_1 (Sentiment)	1.000	0.029	0.137	0.274
C_2 (Volume)		1.000	−0.235	−0.026
C_3 (Negativity)			1.000	0.216
C_4 (Keywords)				1.000

Note: Pearson correlations between the four daily component series over the days with at least 30 articles. Low correlations indicate weak linear association between the series. They do not show that the components carry non-overlapping information or that they capture distinct dimensions of reputational pressure. What they do indicate is that the composite is not driven by a single signal counted several times.

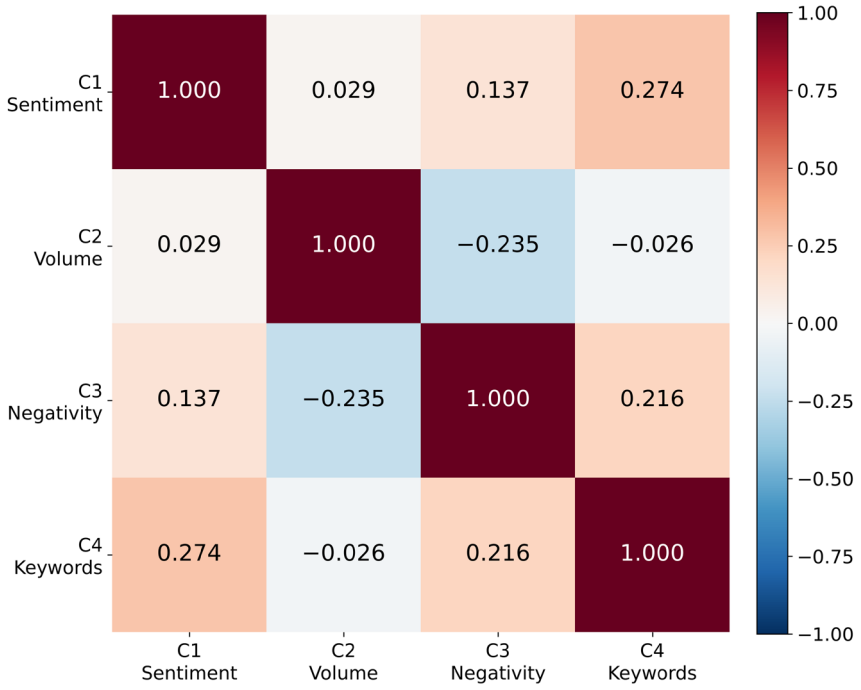


Figure 7. Pearson correlation matrix of the four RRS components (days with at least 30 articles).

4.5. Crisis Detection

Using the rolling z-score method, we identify 63 crisis days, which merge into 53 episodes. Table A9 lists all 53 episodes with their dominant topic and most repeated headline. Table 7 lists the five episodes with the highest peak RRS, together with the dominant coverage on the peak day. The description comes from a reading of the most repeated headlines on that day and from the BERTopic topics of the episode's articles, neither of which enters the index.

Table 7. The five detected episodes with the highest peak RRS (rolling z-score > 2, 14-day window, 3-day merge gap).

Episode	Period	Duration (Days)	Peak RRS	Dominant Coverage on the Peak Day
1	4 March 2020	1	58.68	Financial Intelligence Unit (UIF) investigation of an alleged fraud against INFONAVIT (over MXN 5 billion)
2	10 January 2023	1	53.16	Deadline to convert loans from minimum-wage units (VSM) to pesos; crowded offices; update of the UMA reference unit
3	7 July 2021	1	53.10	50% debt discount for early loan repayment; temporary office closures
4	4 June 2024	1	51.39	Vandalism of LGBTIQ+ flags at INFONAVIT offices; announcement of cumulative lending
5	14–15 January 2021	2	51.11	Arrest of lawyers by the Attorney General's Office (FGR) over a fraud against INFONAVIT; debt-forgiveness controversy

Note: The five episodes with the highest peak RRS under the baseline specification. The dominant coverage column reports an ex post reading of the most repeated headlines on the peak day; it describes what the coverage was about and is not independent validation of the detection, since the episodes were identified first and read afterwards. Episode 3 is best read as a potential false alarm, and Episode 2 is recovered in only half of the detection settings in Table A3 (Sections 4.5 and 4.8).

The episodes differ in nature. Episodes 1 and 5 are adverse governance events: coverage of a UIF fraud investigation (59 of 120 articles labeled negative by the human coders) and of arrests related to a fraud against the institution (91 of 138). Episode 2 is an operational event, the crowding of offices at the deadline to convert loans to pesos, and Episode 4 is an incident at institutional offices that received wide coverage (304 articles). Episode 3 is best read as a potential false alarm: coverage of a 50% discount for early repayment contains 5 human-coded negative articles out of 215 (2.3%, against 4.9% on an average day) and is flagged because of high volume, the keyword deuda (debt) and BERT's tendency to read administrative credit language as negative. What the evidence establishes is that the episode is not associated with adverse human-coded coverage. It does not establish that the detection was mistaken, since the index is designed to flag unusual concentrations of coverage whatever their tone, and we therefore avoid calling it a false alarm outright. Conversely, the largest coverage peak in the sample, Hurricane Otis (788 articles on 1 November 2023), is not flagged: its coverage was relief-oriented, with almost no human-coded negative articles that day, and followed several days of already high volume, so the trailing z-score stayed below the threshold (at most 1.71).

Crisis days also differ in content from other days. Headlines that BERTopic assigns to the Institutional Crisis, Fraud and Alerts, and Labor Reform topics are 3.6, 2.5 and 2.1 times more frequent on crisis days than on other valid days, whereas the dominant Housing Credits topic is equally frequent (ratio 1.07). The Fraud and Alerts and Labor Reform topics share some vocabulary with the crisis lexicon (e.g., fraude, ilegal), so those ratios are partly mechanical; the Institutional Crisis topic does not. These audits are descriptive and ex post. They show what the detected days contain, but they do not establish that the index anticipates reputational damage, which would require an independently compiled event record or institutional outcome data (Section 5.3).

4.6. Geographic and Media Source Analysis

The media coverage of INFONAVIT is quite different from one region of the country to another. Figure 8 shows the volume of coverage and the rate of negative coverage by state, and Table 8 lists the ten outlets that produced the most articles.

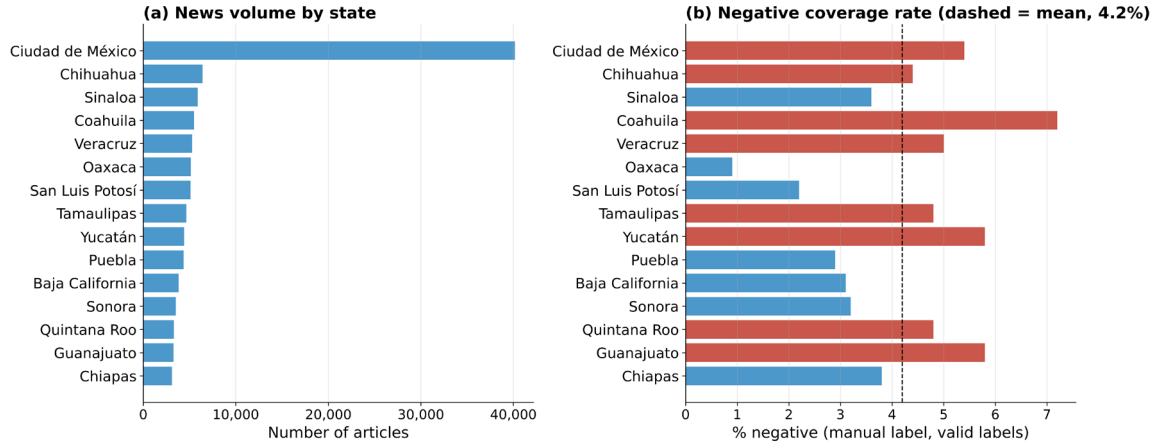


Figure 8. INFONAVIT media coverage by state for the fifteen states with the most articles: (**a**) number of articles and (**b**) share of articles labeled negative by the monitoring service among articles with a valid tone label (dashed line: mean of the fifteen states; red bars, states above the mean; blue bars, states below the mean). Ciudad de México (including records coded as CDMX) accounts for 29.2% of all coverage, followed by Chihuahua (4.7%), Sinaloa (4.3%) and Coahuila (4.0%).

Table 8. Top 10 media outlets by article volume.

Rank	Outlet	Articles	% of Total
1	Milenio	1640	1.2%
2	El Economista	1280	0.9%
3	La Jornada	1250	0.9%
4	El Universal	1172	0.8%
5	El Heraldo de México	1110	0.8%
6	El Financiero	1101	0.8%
7	Mi Bolsillo	1016	0.7%
8	Reforma	954	0.7%
9	La Verdad Noticias	792	0.6%
10	Excélsior	779	0.6%

Note: Shares are computed over the 138,789 articles in the analysis corpus. All outlets enter the index with equal weight, with no adjustment for audience reach or editorial influence. Section 4.8 reports how the results change when the corpus is restricted to the twenty largest outlets.

The top 20 outlets account for 12.7% of total coverage, and the long tail of 9495 outlets reflects a strong presence of local media. Mexico City accounts for 29.2% of all coverage, which includes national outlets headquartered there. Among the fifteen states with the most coverage, the share of articles labeled negative by the human coders, computed over articles with a valid label, is highest in Coahuila (7.2%), Guanajuato (5.8%), Yucatán (5.8%) and Mexico City (5.4%); the mean across these fifteen states is 4.2%, and Sinaloa (3.6%) and Oaxaca (0.9%) are the least negative. Digital outlets (Portal, 76% of coverage) are more often labeled negative than print media (17% of coverage) by BERT (57.7% vs. 50.8%) but less often by the human coders (4.2% vs. 6.9%); Figure A2 reports the human-coded negative share by media type and year, another instance of the divergence documented in Section 4.2.

4.7. Emotion Analysis Results

Although sentiment measures positive and negative polarity, reputational crises are frequently characterized by specific emotional patterns such as fear, anger or distrust. Consequently, emotion detection provides complementary information that cannot be captured through sentiment alone. The pysentimiento emotion classifier identified surprise, anger, and fear as the dominant emotions across the corpus. Surprise tends to cluster around policy statements and changes to political institutions. Anger is most prevalent on themes of fraud and governance. Fear concentrates on events such as the beginning of the COVID-19 pandemic and debates on institutional change.

Multiple episodes of increased anger and fear do not align with changes in the aggregate positive-negative-neutral split, which suggests that the emotion measures pick up variation that the sentiment split does not reflect. This supports keeping both in the pipeline, although it does not show that the two measure distinct constructs.

4.8. Robustness Analysis

Table 9 summarizes how the detections change under alternative specifications. For each variant, we report the correlation with the baseline RRS on common days, the number of crisis days and episodes, the Jaccard overlap between its crisis days and the baseline crisis days, and how many of the five episodes in Table 7 it detects (at least one flagged day within one day of the episode).

Table 9. Robustness of crisis detection to alternative specifications (baseline: weights 0.35, 0.25, 0.25, 0.15; w = 14; $\tau = 2$; g = 3; at least 30 articles per day).

Specification	r with Baseline	Crisis Days	Episodes	Jaccard (Crisis Days)	Table 7 Episodes Detected
Baseline weights	1.000	63	53	1.000	5/5
Equal weights	0.993	64	53	0.841	5/5
Sentiment-heavy weights (0.50, 0.20, 0.20, 0.10)	0.981	69	55	0.784	5/5
Volume-heavy weights (0.25, 0.35, 0.25, 0.15)	0.981	69	55	0.784	5/5
Sentiment excluded (0.00, 0.40, 0.35, 0.25)	0.949	67	54	0.625	4/5
PCA-based weights (0.30, 0.09, 0.28, 0.33), diagnostic only	0.922	70	56	0.529	5/5
Minimum 10 articles per day	0.998	71	59	0.718	5/5
Minimum 1 article per day	0.997	75	61	0.605	4/5
Unique headlines per day	0.961	53	45	0.706	4/5
Top-20 outlets only (≥10 articles per day)	0.640	36	31	0.193	2/5
Without the BERT component (C_1)	0.949	67	54	0.625	4/5
Without the human-coded component (C_3)	0.955	69	52	0.650	3/5
Without C_1 and C_3 (volume and keywords only)	0.891	68	55	0.523	2/5

Note: Each row re-computes the index and re-runs the detection under the stated specification, holding everything else at the baseline. The correlation column is the Pearson correlation of the daily index with the baseline index, and the Jaccard column is the overlap of the sets of crisis days, so a value of 1 means the same days are flagged. The last column counts how many of the five episodes in Table 7 are recovered: all five are recovered under the alternative weighting schemes and the 10-article threshold, whereas the top-20 outlet subsample and the specification without both tone components recover only two. Because each weighting scheme defines a different composite, specifications are compared through these three quantities rather than through the numerical levels of the index.

Weights. The three theory-based alternatives remain close to the baseline ($r \geq 0.981$) and detect all five highest-peak episodes. A principal component analysis on the standardized components gives a first component that explains 36.5% of their variance with loadings 0.498 (C_1), −0.293 (C_2), 0.572 (C_3) and 0.582 (C_4). Volume loads with the opposite sign to the other three, so an index built from the signed scores is not a reputational pressure index at all: it correlates 0.175 with the baseline and flags a largely different set of days. Using absolute loadings hides that sign reversal, so we do not treat the PCA weights as a candidate specification and report them only as a diagnostic: the data do not support a single dominant factor behind the four components.

Detection parameters. Across the 48 combinations of window, threshold and merge gap (Table A3), Episodes 1 and 5 are detected in all configurations, Episode 3 in 94%, Episode 4 in 81% and Episode 2 in 50%. The five episodes are therefore not equally robust. Episodes 1, 3 and 5 can be read as stable detections, Episode 4 as fairly stable, and Episode 2 as a detection that depends on the parameters chosen, which should be kept in mind when reading its description in Section 4.5. The number of crisis days ranges from 16 ($w = 30$, $\tau = 3.0$) to 139 ($w = 7$, $\tau = 1.5$), and the merge gap changes only the number of episodes.

Sentiment components. Removing the BERT component leaves a series that correlates 0.949 with the baseline and recovers four of the five highest-peak episodes and 43 of the 53. Removing instead the human-coded component gives 0.955 and three of five. Removing both, so that only volume and crisis keywords remain, gives 0.891 but only two of the five highest-peak episodes and 41 of the 53: the two fraud-related episodes, whose signature is a concentration of negative articles rather than exceptional volume, are no longer flagged. The detections therefore depend on the tone components taken together, although no single component drives them: volume alone recovers two of five and keywords alone three of five.

Volume baseline. The baseline standardizes daily counts against the full-sample mean, which uses information from the end of the sample at the beginning. Rebuilding C_2 with an expanding past-only baseline gives $r = 0.932$ with 40 crisis days, and a 365-day rolling baseline gives 0.904 with 44 days; both recover three of the five highest-peak episodes. Standardizing within day of the week changes little (0.92). Normalizing by the number of active outlets, which removes the growth of the monitored universe, has the largest effect (0.574, 28 of 53 episodes): much of the late-sample rise in the level of the index reflects how many outlets the provider monitors rather than how intensely the institution is covered. Episode detection, which is local, is far less affected than the level (Table A8).

Detection rule. Replacing the conventional threshold by the empirical 97.7th percentile of the realized z-scores ($\tau = 2.75$) leaves 28 crisis days and still recovers four of the five highest-peak episodes. A robust version based on the median and the median absolute deviation flags more days (98) and recovers all five. Defining the window in calendar days rather than in valid observations changes little (68 days, Jaccard 0.795), and including the current day in the baseline window, which shrinks the score, reduces detections to 38 days. A change-point procedure (PELT with an RBF cost) is not a substitute: it identifies a single level shift in the series, on 24 October 2023, because it looks for persistent regime changes rather than the one- or two-day spikes that interest us (Table A6).

Media composition. Identical headlines published the same day by several outlets are 20.8% of articles; a semantic criterion based on character n-gram cosine similarity above 0.85 raises that share to 26.2%, and removing those near-duplicates gives an index correlated 0.952 with the baseline that recovers four of the five highest-peak episodes. Weighting each outlet equally instead of each article gives 0.988. Splitting the corpus is more informative: state and local outlets (70.2% of articles) reproduce the baseline closely (0.961), digital outlets as well (0.957), whereas national outlets (0.756) and especially print outlets (0.517, 22 of 53 episodes) behave differently. The index as constructed therefore describes the

regional and digital coverage that dominates the corpus, and a reader interested in national print coverage should build the index on that subset (Table A7).

Coverage and sources. Lowering the minimum to 10 articles per day leaves the index essentially unchanged ($r = 0.998$, five of five episodes). Without a minimum ($r = 0.997$, four episodes), the highest value of the series falls on 1 March 2020, a day with a single article, which motivates the threshold. Removing syndicated duplicates (identical headlines on the same day, 20.8% of articles) yields $r = 0.961$ and detects four episodes. An index built only from the 20 outlets with the most articles (12.3% of articles, 733 days with at least 10 articles) correlates 0.640 with the baseline and detects two episodes: the largest outlets capture the main governance events but not all volume-driven episodes, which is consistent with regional and digital outlets amplifying institutional news.

Classifier drift. As reported in Section 4.2, the BERT negative share shows no significant monthly trend ($p = 0.31$). The negative share coded by the provider's human coders, which feeds C_3, does decline (-0.22 pp per month, $p < 0.001$, $R^2 = 0.53$), from monthly values between 4.7% and 17.9% in 2020 to values below 5% in 2023–2024. This reduces the contribution of C_3 in later years. These data do not allow us to tell whether the decline reflects a change in the tone of the coverage or a change in the provider's coding practice. Because C_3 carries a weight of 0.25, the distinction matters for how the index should be read over long horizons, and we return to it in Section 5.3.

5. Discussion

5.1. Interpretation of Key Findings

The RRS summarizes the evolution of INFONAVIT's media environment over the period analyzed. Its highest-peak episodes correspond to identifiable coverage (fraud investigations, operational changes and incidents at institutional offices), and two of these five episodes are recovered under every detection setting, weighting scheme and coverage filter examined in Section 4.8, with a third recovered in 94% of them, while the remaining two depend on the specification chosen. The detections also survive replacing the sentiment classifier with Spanish-language models and reweighting the corpus by outlet, but the two fraud-related episodes are lost when both tone components are removed, so the index does rely on tone to separate adverse coverage from merely abundant coverage (Section 4.8). This is evidence that the index isolates unusual concentrations of intense or adverse coverage. It is not, by itself, evidence of external validity as a measure of reputational damage, which would require linking the index to stakeholder or institutional outcomes, and one of the five episodes is best read as a potential false alarm. The trend component describes slow changes in the institution's media profile but partly reflects the growth in coverage volume. In this way, the paper complements Fiordelisi et al. (2013) and Eckert and Gatzert (2017) by showing that a transparent composite index built from media content can monitor reputational pressure for a non-listed public institution, for which market-based measures are unavailable.

The near-zero Cohen's kappa ($\kappa = -0.0019$) between BERT and the manual labels is worth lingering on, for it admits of two readings. Most of the disagreement is concentrated in two cells of the matrix: Neutral articles that coders classified as such, but BERT classified as Negative (51.8%), and Positive articles that also changed to Negative (59.6%). One explanation here is that BERT is able to capture subtle evaluative or critical language that manual coding misses, where institutional headlines are often coded as neutral with standardized coding guidelines. If this is the case, manual annotations may not represent an unambiguous reference standard for sentiment classification in such a subjective context (Gentzkow et al., 2019). Another possibility is that the discrepancy observed may be a domain adaptation problem in which a model trained with consumer review data systemat-

ically assigns more negative sentiment to the formal and formulaic language of institutional news. We do not think the data let us settle which force dominates, and the analysis does not require us to. Because the index is constructed and interpreted primarily through its temporal dynamics rather than the absolute height of any component, an over-reporting bias that moves slowly shifts the level of the series without displacing the turning points on which our conclusions rest. Section 4.2 reports that the monthly bias shows no linear drift, which is not the same as stability: the monthly share still varies by several percentage points around its mean, and a change occurring within a few days is not absorbed by the trailing window.

A further question is how much the index depends on this component at all. Sentiment carries the largest weight (0.35), so a classifier whose agreement with the manual labels is negligible ($\kappa = -0.0019$, agreement 24.1%) is a genuine concern for the validity of the composite, and not only for the interpretation of C_1 in isolation. The sentiment-excluded specification in Table 9 speaks to it directly. Setting the sentiment weight to zero and redistributing it over the remaining three components leaves the resulting index correlated at $r = 0.949$ with the baseline, recovers four of the five highest-peak episodes and yields a Jaccard index of 0.625 over crisis days. The episodes reported in Section 4.5 are therefore not artifacts of the BERT component alone. They are not independent of tone either: removing both tone components, C_1 and C_3, leaves only two of the five (Section 4.8), so the index needs at least one measure of tone to distinguish adverse coverage from abundant coverage. What the low agreement does rule out is any reading of the absolute level of C_1, or of the RRS, as a calibrated measure of how adverse the coverage was. We make no such reading, and the index should be understood as a detector of unusual media pressure rather than as a validated measure of media tone.

Rather than invalidating the proposed framework, the weak alignment between manual tagging and the transformer model emphasizes the significance of using relative sentiment assessments instead of relying on absolute measures. In our setting, consistency over time matters more for monitoring media pressure than close agreement with subjective human labels, especially when the goal is to promptly identify unusual shifts in media activity.

A structure consistent with the operational and political context of INFONAVIT emerges from the BERTopic analysis. The prevalence of the Housing Credits topic (43,195 articles), which captures the central role of this institution in the mortgage credit market in Mexico, and that of the topics covering labor reform, fraud, and institutional crisis, indicate which subjects the coverage returned to most often over the period. The topic on elections, which emerges with more force in 2024, shows that electoral coverage came to occupy a larger share of the articles mentioning the institution. These are descriptions of the corpus. They do not establish that the political context shapes the institution's reputation, which would require evidence outside the media data. The pattern is nonetheless consistent with the theory of narrative economics (Shiller, 2017).

These findings extend earlier research based primarily on aggregate sentiment indicators by showing that thematic decomposition substantially improves interpretability. Knowing whether a rise in media pressure coincides with coverage of governance issues, labor changes, electoral discussions or housing regulations tells risk managers what the coverage was about, which is useful for deciding where to look and how to respond. It does not identify the institutional weakness behind the coverage, and we do not present the topics as a diagnosis of underlying causes.

Our geographical analysis shows that coverage intensity and negativity differ across states: among the fifteen states with the most coverage, the share of human-coded negative articles ranges from 0.9% in Oaxaca to 7.2% in Coahuila. Media-derived reputational pressure on INFONAVIT is therefore not uniform across its territory. We do not find a

simple northern-border pattern, so the link between regional negativity and the housing vacancy problems documented by Monkkonen (2019) remains a hypothesis for future work.

In general, the findings align well with prior studies that indicate media storytelling impacts financial standing (Tetlock, 2007; Deephouse, 2000; Shapiro et al., 2022). Nevertheless, in contrast to past research that concentrated on publicly traded companies or established financial systems, the current model provides evidence that similar analytical methods can be effectively applied to government-backed housing finance organizations functioning in developing markets.

These findings expand the empirical reach of NLP-based research on reputational risk, and contribute to a growing, but still limited, amount of quantitative evidence on the topic. More importantly, they suggest that reputational risk is inherently multidimensional and cannot be explained fully by sentiment polarity alone. The joint use of transformer-based sentiment analysis, topic modeling, anomaly detection and media intensity enables a more complex characterization of the reputational dynamics by capturing complementary sources of information that would be masked if each technique were used in isolation. This multidimensional framework constitutes a major methodological contribution of the study, especially for public financial institutions operating in politically sensitive contexts.

Operational uncertainty is a central channel in these results. Three of the five highest-peak episodes originate in operational matters: fraud against the institution (Episodes 1 and 5) and a process change affecting borrowers (Episode 2). This pattern is consistent with the view that reputational risk amplifies operational risk events (Fiordelisi et al., 2014; Gillet et al., 2010; Sturm, 2013). It also shows the limits of a media-based measure: Hurricane Otis, a major operational shock for borrowers in Guerrero, generated the largest coverage peak in the sample but little adverse framing, so the index did not flag it.

Table 10 places the RRS among existing approaches to measuring reputational risk. Its main advantages are daily frequency, an institution-specific focus and an open specification that does not require listed securities; its main weaknesses are that it measures media pressure rather than losses and inherits the limitations of the underlying classifiers.

Table 10. Comparison of the RRS with existing approaches to measuring reputational risk.

Approach	Data and Examples	Strengths	Limitations for a Public Housing Lender
Market-based event studies	Abnormal returns around operational loss announcements (Perry & de Fontnouvelle, 2005; Gillet et al., 2010; Fiordelisi et al., 2013; Sturm, 2013)	Monetary estimate of reputational loss	Require listed equity and identified events; ex post
Surveys and reputation rankings	Stakeholder perception surveys (Fombrun & Shanley, 1990)	Direct measure of perceptions	Costly; low frequency; not timely
Media-screening incident databases	Commercial ESG and reputational incident data (Kölbel et al., 2017; Gantchev et al., 2022)	Broad coverage; event-level detail	Proprietary scoring; incident-based; limited transparency
News-based text indices	Policy uncertainty, sentiment and climate news indices (Baker et al., 2016; Engle et al., 2020; Shapiro et al., 2022; Bybee et al., 2024)	High frequency; replicable	Designed for macro or market aggregates, not a single institution
RRS (this study)	138,789 Spanish-language headlines on INFONAVIT; BERT sentiment, volume, human tone labels and crisis keywords	Daily; institution-specific; open specification; documented robustness	Measures media pressure, not losses; classifier domain shift; headlines only

Note: The comparison is qualitative and rests on the cited studies as published; we did not re-estimate any of the alternative measures on our data. The row describing the RRS refers to an indicator of media-derived reputational pressure, which is not the same object as the reputational losses estimated by the market-based studies, so the rows are not alternative estimates of a common quantity.

To distinguish the strength of the evidence behind each conclusion, Table 11 classifies the main findings into results supported by our evidence, implications that are consistent with the literature but not tested here, and propositions for future research.

Table 11. Status of the main findings.

Finding or Implication	Status	Basis
Episodes 1 and 5 are detected under all 48 detection settings, Episode 3 in 94% of them and Episodes 4 and 2 in 81% and 50%; detections are similar across weighting schemes and coverage thresholds, while the top-20 outlet subsample recovers two of the five episodes and the removal of syndicated duplicates four of five	Supported by our evidence	Tables 9 and A3
The highest-peak episodes correspond to fraud investigations, operational changes and institutional incidents	Supported by our evidence (ex post content audit)	Table 7; Section 4.5
BERT's negativity bias is large and shows no linear drift, although its monthly level varies	Supported by our evidence	Section 4.2
The provider's human-coded negative share declines over the sample (−0.22 pp per month), which may reflect coverage or coding practice	Supported by our evidence; source of the decline not identified	Sections 4.8 and 5.3
Adverse media coverage affects institutional standing and financial outcomes	Consistent with the literature; not tested here	Deephouse (2000); Tetlock (2007); Kölbel et al. (2017)
Reputational damage amplifies operational risk events	Consistent with the literature; not tested here	Gillet et al. (2010); Sturm (2013); Fiordelisi et al. (2014)
P1: RRS spikes precede changes in institutional outcomes (complaints, loan origination, voluntary contributions)	Proposition for future testing	Requires outcome data
P2: Fraud-related episodes have more persistent effects than volume-driven episodes	Proposition for future testing	Requires outcome data and a longer series
P3: Regional differences in negative coverage reflect regional housing conditions such as vacancy	Proposition for future testing	Monkkonen (2019); Figure 8

Note: The table separates what the analysis establishes from what it does not. "Supported by our evidence" means supported within the limits stated in Section 5.3, the classifier's domain shift and the decline in the human-coded negative share. "Consistent with the literature" marks statements we did not test on our data. Propositions P1 to P3 are stated for future testing and are not results of this study.

The framework can be transferred to other institutions because it requires only a continuous media-monitoring feed, a sentiment model for the local language and a lexicon adapted to local usage. In Mexico, the same specification could be applied to FOVISSSTE, the housing fund for public-sector workers, allowing comparisons within the sector. In other emerging economies, public or provident-fund housing lenders, such as Caixa Econômica Federal as operator of FGTS-funded housing programs in Brazil, combine political exposure and fraud risk in a similar way. In developed economies, government-sponsored enterprises such as Fannie Mae and Freddie Mac in the United States, or public promotional banks, could use the index as a complement to market-based measures, which do not capture daily

shifts in public scrutiny. In each case, the lexicon, the minimum-coverage threshold and the sentiment model should be re-validated using the checks in Sections 4.2 and 4.8.

5.2. Practical Implications

Beyond its methodological contribution, the framework has specific uses for public housing finance institutions. We state them as operating procedures rather than generic recommendations:

1. Monitoring and escalation. The daily RRS and its trailing z-score can feed a key risk indicator with predefined escalation levels; for example, $z > 2$ triggers a same-day review of headlines by the communications and risk areas, and $z > 3$ or two consecutive flagged days escalate the case to the risk committee. Because Episode 3 shows that alerts can be driven by coverage volume without adverse content, the protocol should require a human content check before any external response. These thresholds are proposed management conventions derived from the distribution of the index. They have not been validated as early-warning thresholds, and this study does not test whether the index anticipates institutional events.
2. Fraud and service communication. Fraud-related coverage (Episodes 1 and 5 and the Fraud and Alerts topic) calls for coordinated responses: public alerts on fraudulent intermediaries, foreclosure-sale scams and identity theft, and timely disclosure about investigations. Operational changes that affect borrowers (Episode 2) should come with communication plans that anticipate deadlines and demand at service offices.
3. Regional prioritization. Negative coverage differs markedly across states (from 0.9% to 7.2% among the fifteen states with the most coverage), so regional offices can use state-level indicators to allocate outreach and service capacity.
4. Board and supervisory reporting. A monthly summary of RRS levels, flagged episodes and their dominant topics can be added to risk-committee reporting and linked to the operational risk event database, so that each flagged episode is matched with any associated operational loss.
5. Trend monitoring. The STL trend and the yearly means of the components help distinguish structural changes in coverage, such as growth in volume, from a genuine deterioration in tone, which avoids over-reacting to level shifts.

Reputational pressure does not affect all stakeholders in the same way. Following stakeholder salience theory (Mitchell et al., 1997; Wood et al., 2021), Table 12 maps INFONAVIT's main stakeholders to the concerns that drive their perceptions, the RRS signals that capture those concerns, the potential consequences and the corresponding management responses. The mapping is a proposal derived from the topics and episodes in this study rather than an estimated result.

The results also have implications for regulators and housing policy. First, supervisors of public financial institutions could ask that reputational risk be monitored with documented and reproducible indicators, reporting their method, thresholds and robustness, instead of qualitative statements alone, in line with the Basel Committee on Banking Supervision (Basel Committee on Banking Supervision [BCBS], 2009) view of reputational risk as a risk behind other risks. Second, because fraud-related coverage dominates the most severe adverse episodes, housing authorities and the institution could coordinate anti-fraud information campaigns with law-enforcement announcements. Third, media-based indicators can support the evaluation of changes in housing policy, such as the conversion of loans to pesos, by showing how their implementation is perceived. Fourth, an open specification allows the same indicator to be compared across public housing lenders, such as INFONAVIT and FOVISSSTE, which supports sector-level oversight. These uses require

care: the index measures media pressure, not losses, and should complement rather than replace supervisory judgment.

Table 12. Stakeholder mapping of reputational risk for INFONAVIT.

Stakeholder	Main Concern	RRS Signal	Potential Consequence	Management Response
Affiliated workers and borrowers	Fraud, loan conditions, service quality	Peaks in C_3 and C_4 in Fraud and Alerts or Housing Credits coverage	Loss of trust; complaints; lower voluntary contributions	Fraud alerts; clear communication of loan conditions; service-office capacity
Employers	Contribution rules; labor and outsourcing reforms	Volume peaks (C_2) in Labor Reform coverage	Resistance to or non-compliance with contribution rules	Guidance on rules; outreach through business chambers
Federal government and Congress	Governance; corruption cases; political use of the institution	Peaks in C_3 and C_4 in Institutional Crisis and election coverage	Loss of political support; governance or legal reforms	Transparent disclosure; timely reporting to oversight bodies
Unions and tripartite governing bodies	Representation; labor relations	Coverage in the Union Activity topic	Internal conflict; governance disputes	Dialogue within tripartite bodies; timely internal communication
Supervisors and auditors	Internal control; fraud; operational failures	Flagged episodes ($z > 2$) linked to fraud or operational events	Supervisory scrutiny; audits; sanctions	Link episodes to the operational risk event database; report to the risk committee
Investors and funding counterparties	Credit quality; governance	Sustained rise in the RRS level and in C_3	Higher funding costs; reduced market access	Investor communication; disclosure of risk indicators
Housing developers and local governments	Housing vacancy; urban development	State-level negativity in Urban Development coverage	Lower participation in housing programs	Regional coordination; vacancy-reduction programs

Note: The table is a management framework proposed by the authors and not an empirical result. The RRS signal column describes the index behavior a given stakeholder would be expected to react to, and the response column states suggested practice. Neither the mapping nor the responses were tested against institutional outcomes.

5.3. Limitations

It is important to note some limitations. Our analysis was based on headlines only and did not consider the text of the full article. Headlines are intended to capture the gist or main thrust of an article, and as such are the result of editorial decision-making; nevertheless, they do not tell the full story of how journalists have reported on the events we studied. Even though headlines encapsulate the editorial perspective of each piece and thus serve as a suitable element for tracking media, including the entire text of the article might offer deeper contextual insights about the severity, duration, and representation of events that affect reputation.

The sentiment model was trained on consumer reviews and applied off the shelf to institutional news, so it carries a domain shift that inflates the negative class; for that reason, we caution against reading the absolute level of the sentiment component, or the raw negative share, as a literal measure of how adverse coverage was. What the framework does support is comparison over time, since a bias that changes slowly is largely absorbed once each day is compared with its own recent history. Changes in the bias over a few days, for instance when coverage turns to a topic the model reads poorly, are not absorbed. We also do not claim that the classifier is valid for this domain; fine-tuning it on annotated INFONAVIT headlines would be needed for that. Two data features also matter. The share of articles labeled negative by the provider declines over the sample, which reduces the contribution of C_3 in later years, and the June–November 2020 delivery had to be excluded, so the index is silent for that period.

The temporal behavior of the manually coded component deserves separate emphasis, because it bears on what the index can support over long horizons. The share of articles that the provider's coders label negative falls from monthly values between 4.7% and 17.9% in 2020 to values below 5% in 2023 and 2024, a decline of 0.22 percentage points per month ($p < 0.001$, $R^2 = 0.53$). We cannot distinguish a genuine improvement in the tone of coverage from a change in the provider's coding practice over those five years, and the second reading is at least as plausible as the first, given that the corpus was assembled from successive deliveries by the same monitoring service. Part of the long-run movement of the RRS may therefore reflect the coding process rather than media pressure. Two consequences follow. Levels of the index are not comparable across distant periods, and the trend component of the STL decomposition should not be read as a change in the institution's media environment, which reinforces the caution already stated in Section 4.4.1. Crisis detection is less exposed, because each day is compared with its own trailing 14-day window and a drift of this size is absorbed within that window, but a component whose measurement process is not stable across the sample limits what the composite can be asked to do over long horizons.

Future work could improve the proposed framework in several ways. First, fine-tuning the model on a manually annotated dataset of INFONAVIT headlines may enhance the reliability of the sentiment estimates. Second, the current RRS weighting scheme is theory-driven rather than empirically optimized, making it worthwhile to evaluate alternative approaches, such as weights calibrated against institutional outcomes, beyond the equal and PCA-based schemes examined in Section 4.8.

The crisis keyword lexicon was manually developed, and future work could investigate automatic or data-driven approaches for the construction and updating of this component. Because matching is by substring, generic terms such as deuda (debt) also appear in routine credit news, which explains the volume-driven alert in Episode 3. In addition, social media data (Twitter/X, Facebook) are not included, although they constitute an increasingly relevant part of the public discourse. All media are treated equally, without adjustments for audience reach or editorial influence (Deephouse, 2000). The lack of a causal identification strategy constrains our capacity to make causal claims about the link between media-derived reputational pressure and institutional outcomes.

The proposed framework could be improved by future studies that include full news articles, custom transformer models for specific institutions, multilingual models for financial language and external validation against indicators of operational, financial or market performance of institutions. These additions would help assess not just changes in media pressure, but also whether those changes have consequences for broader financial results. Finally, the episode descriptions in Section 4.5 are ex post. A stronger test would compare the index with a record of institutional events compiled independently and before inspecting the index, and with outcomes such as complaints, loan origination or voluntary contributions.

6. Conclusions

We developed a holistic NLP framework for monitoring media-derived reputational pressure in public housing finance and implemented it on INFONAVIT. We constructed a Reputational Risk Score based on 138,789 news articles published by 9495 media outlets over approximately five years. The score incorporates BERT-based sentiment analysis, media-volume dynamics, negativity concentration, and crisis-keyword intensity.

Besides the presentation of a new indicator, the proposed framework shows how media coverage of an institution can be described along several dimensions at once: emotions, themes, volume and temporal irregularities. This holistic perspective is a good complement to existing approaches, which tend to rely on a single measure of media tone.

We show that transformer-based models can be applied at scale to Spanish-language media, while making clear that the model used here is not validated for this domain: its agreement with human coding is negligible, and what the framework supports is the detection of relative changes in coverage rather than the measurement of sentiment itself, contributing to the largely English-language literature on text-as-data methods in finance (Gentzkow et al., 2019; Loughran & McDonald, 2011). The four-component composite index combines four signals of media-derived reputational pressure and detects episodes whose content (fraud investigations, operational changes and institutional incidents) can be audited; two of them are detected under every detection setting tested and a third in 94% of them.

From a methodological point of view, the near-total disagreement between the BERT predictions and the manually assigned sentiment labels ($\kappa = -0.0019$) is an informative result. Several factors may explain this discrepancy. On the one hand, it may reflect a domain-adaptation problem arising from applying a model trained on consumer review data to institutional news. On the other hand, the model might detect subtle evaluative language that human coders, following standardized annotation criteria, would classify as neutral. Since both explanations are plausible, the available data do not allow a definitive determination of the relative contributions of the two to the disagreement. Either way, the exercise cautions against treating manual sentiment as a gold standard for media coverage and, in our case, justifies building the index on the relative movement of sentiment rather than its absolute level. This points to a practical guideline for media-based monitoring: when a sentiment tool is used outside the context it was trained for, relative changes over time are more informative than absolute sentiment levels for detecting unusual media pressure, although they do not make the tool valid for the new domain.

The methodology is applicable to other institutions. Our analysis is presented for the case of the Mexican housing lender, INFONAVIT, but the general NLP pipeline, formulation of the RRS, and the time-series analysis are directly usable in similar financial, government-backed housing, and regulatory organizations in other Latin American countries and developing markets.

Further work is needed to fine-tune the BERT model on Mexican institutional text, consider social media data for the RRS indicator, empirically estimate the RRS weights based on an ex-ante evaluation, and consider a causal approach to study reputational shocks' impact on institutions' outcomes.

The contribution of this work is, in the end, twofold. Methodologically, it shows that a transparent and reproducible composite index can be built entirely from routinely collected media-monitoring data, giving a public financial institution a quantitative reputational signal where before there was only qualitative judgment. Substantively, it turns nearly five years of coverage of INFONAVIT into a dated record of when coverage of the institution intensified and what it was about, the kind of evidence a risk or communications team can act on. Because every step of the pipeline is documented and the code is available, we hope other public lenders across the region will adapt the approach to their own media environments and, in doing so, help move reputational risk for these institutions from an intuition into a measured quantity.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing—original draft preparation, and writing—review and editing, A.L.J.-P., J.C.T.-G., M.Á.M.-G., and F.V.-M. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets presented in this article are not readily available because the raw media monitoring data cannot be shared publicly due to contractual restrictions with the data provider. The code that reproduces all results and the daily aggregated series are openly available at https://github.com/AnaJZP/infonavit-sentiment (accessed on 21 October 2026) (scripts 06–08 reproduce the analysis, and scripts 09–12 the label recovery, the Spanish-model comparison and the additional robustness analyses).

Acknowledgments: The authors thank the anonymous reviewers for their comments and suggestions.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

Abbreviation	Definition
BCBS	Basel Committee on Banking Supervision
BERT	Bidirectional Encoder Representations from Transformers
CDF	Cumulative Distribution Function
c-TF-IDF	Class-based Term Frequency-Inverse Document Frequency
ESG	Environmental, Social, and Governance
FGR	Fiscalía General de la República (Attorney General's Office)
FGTS	Fundo de Garantia do Tempo de Serviço (Brazil)
FOVISSSTE	Fondo de la Vivienda del Instituto de Seguridad y Servicios Sociales de los Trabajadores del Estado
GDP	Gross Domestic Product
GSE	Government-Sponsored Enterprise
HDBSCAN	Hierarchical Density-Based Spatial Clustering of Applications with Noise
INFONAVIT	Instituto del Fondo Nacional de la Vivienda para los Trabajadores
LDA	Latent Dirichlet Allocation
MAD	Median Absolute Deviation
NLP	Natural Language Processing
NPMI	Normalized Pointwise Mutual Information
PCA	Principal Component Analysis
PELT	Pruned Exact Linear Time
RBF	Radial Basis Function
RRS	Reputational Risk Score
STL	Seasonal-Trend Decomposition based on Loess
UIF	Unidad de Inteligencia Financiera (Financial Intelligence Unit)
UMA	Unidad de Medida y Actualización (Mexican reference unit)
UMAP	Uniform Manifold Approximation and Projection
VSM	Veces Salario Mínimo (minimum-wage units)

Appendix A. Crisis Keyword Lexicon

Table A1. Crisis keyword lexicon (K).

Spanish Term	English Translation
fraude	fraud
corrupción	corruption
corrupto	corrupt
escándalo	scandal
denuncia	complaint/accusation
demanda judicial	lawsuit
demandado	sued/defendant
demandan	(they) sue/demand
irregularidad	irregularity

Table A1. *Cont.*

Spanish Term	English Translation
desvío	diversion of funds
robo	theft
abuso	abuse
nepotismo	nepotism
conflicto	conflict
protesta	protest
queja	complaint
desalojo	eviction
embargo	seizure/garnishment
crisis	crisis
colapso	collapse
impunidad	impunity
investigación	investigation
ilegal	illegal
sanción	sanction/penalty
multa	fine
pérdida	loss
deuda	debt
morosidad	delinquency
cartera vencida	non-performing loans
problema	problem

Note: Matching is case-insensitive substring matching on headlines, which is why generic terms such as deuda also match routine credit news; this is the mechanism behind the volume-driven alert in Episode 3. The term demanda was excluded because in this corpus it refers mostly to demand for housing or credit, and was replaced by demanda judicial, demandado and demandan. The lexicon was compiled by the authors and is not derived from the data.

Appendix B. Additional Figures

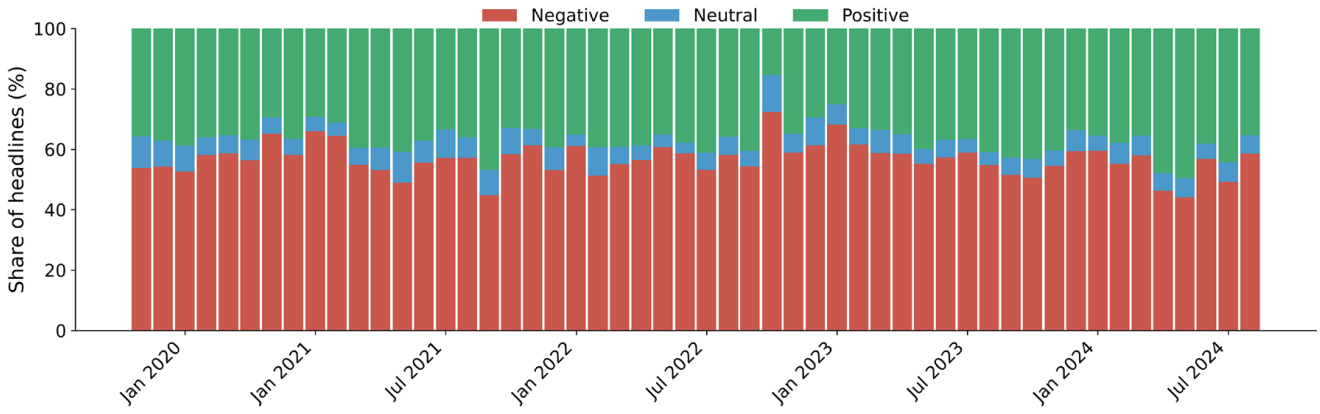


Figure A1. Monthly distribution of BERT sentiment classifications (Negative, Neutral, Positive) from November 2019 to July 2024, excluding the June–November 2020 delivery. The high share of negative classifications in every month reflects BERT's evaluative sensitivity to institutional media language.

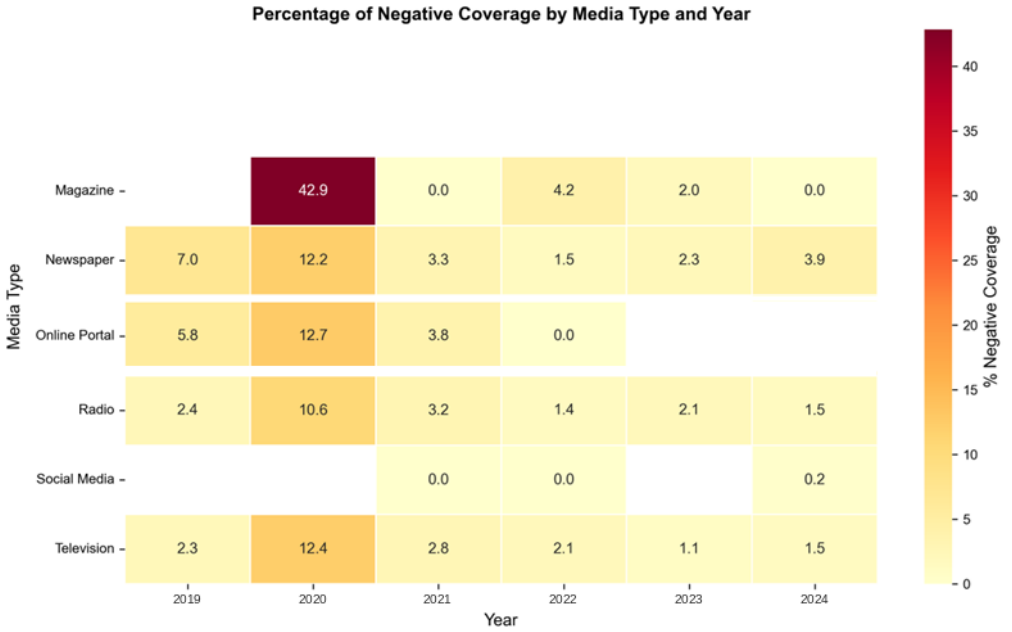


Figure A2. Annual percentage of articles labeled negative by the monitoring service, by media type and year (heatmap). Magazine outlets had the greatest share of negative coverage in 2020 (42.9%), while social media channels consistently had the lowest share of negative coverage throughout the period of observation.

Appendix C. Robustness of RRS Weights

Table A2. Sensitivity of the RRS to alternative weight configurations (days with at least 30 articles). The PCA row uses the normalized absolute loadings on the first principal component; as Section 4.8 explains, the volume component loads with the opposite sign, so this row is reported as a diagnostic and not as a candidate specification.

Configuration	(Alpha, Beta, Gamma, Delta)	Mean	Std. Dev.	Crises (z > 2)	Corr. w/Baseline
Baseline	(0.35, 0.25, 0.25, 0.15)	33.56	6.77	63	1.000
Equal weights	(0.25, 0.25, 0.25, 0.25)	28.57	6.76	64	0.993
Sentiment-heavy	(0.50, 0.20, 0.20, 0.10)	39.12	6.05	69	0.981
Volume-heavy	(0.25, 0.35, 0.25, 0.15)	32.55	8.79	69	0.981
Sentiment-excluded	(0.00, 0.40, 0.35, 0.25)	21.92	9.86	67	0.949
PCA-based	(0.30, 0.09, 0.28, 0.33)	24.86	5.18	70	0.922

Note: Each configuration re-computes the index over the days with at least 30 articles. Means and standard deviations lie on the fixed 0 to 100 scale but are not comparable in level across configurations, since each set of weights defines a different composite; the correlation with the baseline and the number of detected crises are the quantities that support comparison. PCA-based weights are the normalized absolute loadings on the first principal component of the four components, reported only as a diagnostic (Section 4.8).

Table A3. Crisis detection under alternative detection parameters (merge gap g = 3 days; the number of crisis days does not depend on g).

Window w (Days)	Threshold τ	Crisis Days	Episodes	Table 7 Episodes Detected
7	1.5	139	97	5/5
7	2.0	81	66	4/5
7	2.5	56	49	4/5
7	3.0	37	35	4/5

Table A3. *Cont.*

Window w (Days)	Threshold τ	Crisis Days	Episodes	Table 7 Episodes Detected
14	1.5	115	75	5/5
14	2.0	63	53	5/5
14	2.5	36	33	4/5
14	3.0	20	19	4/5
21	1.5	113	75	5/5
21	2.0	56	47	5/5
21	2.5	30	27	4/5
21	3.0	18	17	3/5
30	1.5	97	65	5/5
30	2.0	50	40	5/5
30	2.5	28	25	4/5
30	3.0	16	15	2/5

Appendix D. Tone Label Audit

Table A4. Records whose tone field does not contain a valid tone label, by year of delivery.

Year	Records Delivered	Valid Tone Label	Not a Tone Label	Share Not a Tone Label
2019	1534	1534	0	0.0%
2020	12,376	12,376	0	0.0%
2021	21,486	10,709	10,777	50.2%
2022	28,502	22,979	5523	19.4%
2023	41,099	41,099	0	0.0%
2024	38,621	24,520	14,101	36.5%

Table A5. Sentiment classifiers compared on the full corpus. The index column rebuilds C_1 with each model and leaves the rest of the specification unchanged.

Model	Negative Share	Kappa vs. Human Coders	Agreement with Multilingual BERT	r of the Index with Baseline	Top-5 Episodes
nlptown multilingual BERT (baseline)	55.8%	−0.0019	—	1.000	5/5
BETO (Spanish)	15.2%	0.025	25.2%	0.941	5/5
RoBERTuito (Spanish)	25.5%	−0.0034	31.9%	0.952	5/5
Human coders (monitoring service)	4.7%	—	24.1%	—	—

Table A6. Alternative detection rules (baseline: trailing $z > 2$ over 14 valid observations).

Rule	Crisis Days	Episodes	Jaccard with Baseline	Top-5 Episodes
Empirical 97.7th percentile of z ($\tau = 2.75$)	28	26	0.444	4/5
Robust z (median/MAD), $\tau = 2$	98	64	0.533	5/5
14 calendar days (not valid observations)	68	56	0.795	5/5
Current day included in the baseline window	38	35	0.578	4/5

Table A7. Media composition: de-duplication, outlet weighting and subsets of the corpus.

Specification	Share of Articles	r with Baseline	Crisis Days	Jaccard	Top-5 Episodes
Semantic de-duplication (cos ≥ 0.85)	—	0.952	56	0.653	4/5
Outlet-weighted index (equal weight per outlet)	—	0.988	59	0.768	5/5
National coverage only (≥10 articles/day)	28.2%	0.756	70	0.462	4/5
State and local coverage (≥10 articles/day)	70.2%	0.961	62	0.543	5/5
Print outlets (≥10 articles/day)	17.1%	0.517	43	0.191	3/5
Digital outlets (≥10 articles/day)	75.9%	0.957	77	0.556	4/5

Table A8. Alternative constructions of the volume component C_2.

Specification	r with Baseline	Crisis Days	Episodes	Jaccard	Top-5 Episodes
C_2 past-only (expanding)	0.932	40	34	0.515	3/5
C_2 past-only (365-day)	0.904	44	37	0.551	3/5
C_2 day-of-week adjusted	0.920	75	57	0.683	4/5
C_2 articles per active outlet	0.574	62	51	0.276	4/5

Appendix E. All Detected Episodes

Table A9. All 53 detected episodes (trailing $z > 2$, 14-day window, 3-day merge gap), ordered by peak RRS. The dominant topic is the most frequent BERTopic cluster among the episode's articles and the headline is the most repeated one on the peak day; neither enters the index; headlines longer than 70 characters are truncated.

Episode	Period	Days	Peak RRS	Articles	Negative (human)	Dominant Topic	Most Repeated Headline
1	4 March 2020	1	58.68	120	59	3_fraude_alerta_alertan_advierte	UIF investiga probable fraude al Infonavit por más de 5 mmdp
2	10 January 2023	1	53.16	218	43	0_pesos_crédito_créditos_mil	Abarrotan Infonavit para cambiar las VSM a pesos
3	7 July 2021	1	53.10	215	5	0_pesos_crédito_créditos_mil	Descuenta Infonavit 50% de deuda por liquidación anticipada
4	4 June 2024	1	51.39	304	2	10_comunidad_sindicato_líder_sindical	Rompen banderas LGBTIQ+ en oficinas del Infonavit
5	14–15 January 2021	2	51.11	138	91	0_pesos_crédito_créditos_mil	Detiene FGR por desfalco a abogados de Infonavit
6	31 March 2022	1	50.70	119	3	0_pesos_crédito_créditos_mil	Ráfagas|Canaco amanece con nuevo presidente
7	14–15 August 2023	2	50.46	428	2	0_pesos_crédito_créditos_mil	Crédito hipotecario ha sido resiliente: Infonavit
8	11–12 December 2023	2	49.81	541	0	0_pesos_crédito_créditos_mil	Nombres, Nombres Y... Nombres//Bemoles en la liquidación de seguros en
9	5 April 2023	1	49.80	219	0	0_pesos_crédito_créditos_mil	Faltan 1.6 millones de familias por convertir sus créditos a pesos: In
10	2 January 2023	1	49.77	316	12	0_pesos_crédito_créditos_mil	Infonavit extiende conversión de créditos a pesos hasta el 10 de enero
11	16 February 2023	1	49.63	216	2	9_presidencial_quiebra_amlo_operacione	Aeromar deja de volar por no pagar su deuda
12	22 November 2023	1	49.31	240	1	0_pesos_crédito_créditos_mil	Fondo en EU quiere rescatar la siderúrgica AHMSA: AMLO
13	18–19 September 2023	2	48.40	447	6	0_pesos_crédito_créditos_mil	Toma Delfina Gómez Protesta a titulares de Secretaría de Finanzas, de

Table A9. *Cont.*

Episode	Period	Days	Peak RRS	Articles	Negative (human)	Dominant Topic	Most Repeated Headline
14	7 November 2022	1	48.37	357	1	0_pesos_crédito_créditos_mil	Fonacot listo para lo que soliciten los trabajadores
15	23 August 2022	1	48.20	150	31	0_pesos_crédito_créditos_mil	Duplica Infonavit morosos
16	21–22 February 2022	2	48.00	248	65	0_pesos_crédito_créditos_mil	Se complica megafraude de 60 mil viviendas en Infonavit
17	10–11 July 2022	2	47.40	281	6	0_pesos_crédito_créditos_mil	Infonavit entrega el primer “Crediterreno” en Yucatán
18	20 October 2021	1	47.36	136	11	0_pesos_crédito_créditos_mil	INFONAVIT RELANZA EL DESCUENTO DE 50% DE LA DEUDA A QUIEN LIQUIDE SU C
19	8 June 2021	1	46.85	80	49	0_pesos_crédito_créditos_mil	Infonavit otorga 16 contratos directos y sin justificación
20	12 May 2022	1	46.79	140	13	0_pesos_crédito_créditos_mil	Convierte tu crédito de VSM a pesos en mi cuenta Infonavit
21	4 October 2023	1	46.43	199	2	0_pesos_crédito_créditos_mil	Infonavit ha dispersado 1.1 billones de pesos en créditos en gobierno
22	3 December 2020	1	45.92	145	29	0_pesos_crédito_créditos_mil	Achaca AMLO a corrupción abandono de viviendas
23	23 June 2023	1	45.65	213	2	0_pesos_crédito_créditos_mil	Eres mujer Tu crédito Infonavit será más fácil que nunca
24	9 February 2023	1	45.32	215	0	0_pesos_crédito_créditos_mil	Infonavit reabre la ventanilla para cambiar créditos a pesos
25	4 August 2021	1	44.98	207	18	0_pesos_crédito_créditos_mil	Capitanes
26	23–24 August 2021	2	44.68	301	23	0_pesos_crédito_créditos_mil	Sufren 3.1 millones para pagar su casa
27	20 December 2022	1	44.57	213	3	0_pesos_crédito_créditos_mil	Construir con un crédito Infonavit ahora es ás fácil
28	26–27 December 2021	2	44.41	326	5	4_imss_hogar_real_ante	Incumplen Ley de subcontratación
29	17 May 2021	1	44.22	192	2	0_pesos_crédito_créditos_mil	Despertador//¿Es de izquierda la 4a Transformación?

Table A9. *Cont.*

Episode	Period	Days	Peak RRS	Articles	Negative (human)	Dominant Topic	Most Repeated Headline
30	2 December 2022	1	44.15	181	1	0_pesos_crédito_créditos_mil	Piden cambiar a pesos las hipotecas
31	8 February 2022	1	44.12	133	32	0_pesos_crédito_créditos_mil	Nombres, Nombres y... Nombres|Batalla vivienda con inflación, Canade
32	12 February 2020	1	44.02	81	31	0_pesos_crédito_créditos_mil	Paró Infonavit revisión a lo recuperado por Gertz
33	8–9 August 2023	2	43.57	328	24	0_pesos_crédito_créditos_mil	Más de 1 millón 700 mil créditos Infonavit todavía pueden convertirse
34	11 March 2022	1	43.43	107	22	0_pesos_crédito_créditos_mil	Rafael Zaga Tawil denuncia amenazas
35	4 June 2021	1	43.43	153	17	2_reforma_empresas_ilegal_ley	Los fraudes con outsourcing. "maléficos" contra empleados
36	9 December 2022	1	43.26	165	8	0_pesos_crédito_créditos_mil	Acude director general del IMSS a la 126 Asamblea General del INFONAVI
37	18 May 2020	1	43.11	76	28	0_pesos_crédito_créditos_mil	Pide FGR captura de ligados a fraude al Infonavit
38	10 August 2022	1	43.04	156	3	0_pesos_crédito_créditos_mil	Infonavit hace justicia apoyando a más de 3.3 millones de acreditados
39	3 August 2022	1	42.98	152	6	0_pesos_crédito_créditos_mil	Tu dinero en Infonavit está protegido contra la inflación
40	4 March 2021	1	42.88	151	4	0_pesos_crédito_créditos_mil	Larrea Impone "Esclavismo Disfrazado": Gómez Urrutia
41	28 March 2023	1	41.98	155	1	0_pesos_crédito_créditos_mil	Nuevas opciones para pagar tu crédito Infonavit desde Estados Unidos
42	21 September 2022	1	41.85	179	0	0_pesos_crédito_créditos_mil	Infonavit instalará módulo de atención para trabajadores afectados por
43	15 June 2023	1	41.79	156	2	0_pesos_crédito_créditos_mil	Invertirán Gobierno de Sonora e Infonavit mil 500 millones de pesos pa
44	19 January 2022	1	41.55	181	3	0_pesos_crédito_créditos_mil	Liberan cuentas de Rafael Zaga
45	12 May 2023	1	40.85	136	8	0_pesos_crédito_créditos_mil	Más de cinco mil personas en Oaxaca han convertido su crédito Infonavi

Table A9. *Cont.*

Episode	Period	Days	Peak RRS	Articles	Negative (human)	Dominant Topic	Most Repeated Headline
46	20–21 June 2022	2	39.53	257	7	0_pesos_crédito_créditos_mil	Infonavit inicia entrega de pagos por seguro de daños a afectados por
47	13 July 2023	1	39.43	135	1	0_pesos_crédito_créditos_mil	Tribunal tira órdenes de aprehensión contra Rafael Zaga Tawil y su hij
48	13 December 2019	1	39.26	35	15	0_pesos_crédito_créditos_mil	Paro nacional de trabajadores del Infonavit por desacuerdo con Directo
49	15 January 2020	1	39.09	67	7	0_pesos_crédito_créditos_mil	Tendrán hipotecas menor alza en 4 años
50	2 May 2023	1	38.79	157	3	0_pesos_crédito_créditos_mil	Infonavit debe crear programas para jóvenes
51	16 November 2021	1	38.66	106	11	14_quién_pasa_caso_qué	Infonavit congela cobro a los desaparecidos
52	17 December 2021	1	37.74	93	15	0_pesos_crédito_créditos_mil	Aguinaldos, deudas, e inflación
53	20 April 2020	1	36.96	112	5	0_pesos_crédito_créditos_mil	Jesús Padilla “No queremos ver gente robando un pan, por no atender la

Note: The 16 combinations of window and threshold (48 with the three merge gaps) are run with the merge gap fixed at 3 days; the merge gap changes how flagged days are grouped into episodes but not how many days are flagged. The last column counts the episodes from Table 7 that are recovered: Episodes 1 and 5 in every combination, Episode 3 in 94%, Episode 4 in 81% and Episode 2 in 50%.

References

Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv*, arXiv:1908.10063. [CrossRef]

Ashwin, J., Kalamara, E., & Saiz, L. (2024). Nowcasting euro area GDP with news sentiment: A tale of two crises. *Journal of Applied Econometrics*, *39*(5), 887–905. [CrossRef]

Aula, P., & Mantere, S. (2020). *Strategic reputation management: Towards a company of good*. Routledge. [CrossRef]

Baker, S. R., Bloom, N., & Davis, S. J. (2016). Measuring economic policy uncertainty. *The Quarterly Journal of Economics*, *131*(4), 1593–1636. [CrossRef]

Basel Committee on Banking Supervision. (2009). *Enhancements to the Basel II framework*. Bank for International Settlements. Available online: https://www.bis.org/publ/bcbs157.htm (accessed on 21 October 2026).

Bholat, D., Hansen, S., Santos, P., & Sheridan-Sherlock, C. (2015). *Text mining for central banks* (Bank of England Centre for Central Banking Studies Handbook, No. 33). Available online: https://researchonline.lse.ac.uk/id/eprint/62548 (accessed on 21 October 2026).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022. Available online: https://www.jmlr.org/papers/v3/blei03a.html (accessed on 21 October 2026).

Busuioc, E. M., & Lodge, M. (2016). The reputational basis of public accountability. *Governance*, *29*(2), 247–263. [CrossRef]

Busuioc, M., & Lodge, M. (2017). Reputation and accountability relationships: Managing accountability expectations through reputation. *Public Administration Review*, *77*(1), 91–100. [CrossRef]

Bybee, L., Kelly, B., Manela, A., & Xiu, D. (2024). Business news and business cycles. *The Journal of Finance*, *79*(5), 3105–3147. [CrossRef]

Cañete, J., Chaperon, G., Fuentes, R., Ho, J.-H., Kang, H., & Pérez, J. (2020). Spanish pre-trained BERT model and evaluation data (BETO). *arXiv*, arXiv:2308.02976. [CrossRef]

Carpenter, D. P., & Krause, G. A. (2012). Reputation and public administration. *Public Administration Review*, *72*(1), 26–32. [CrossRef]

Carretta, A., Farina, V., Martelli, D., Fiordelisi, F., & Schwizer, P. (2011). The impact of corporate governance press news on stock market returns. *European Financial Management*, *17*(1), 100–119. [CrossRef]

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. [CrossRef]

Cummins, J. D., Lewis, C. M., & Wei, R. (2006). The market value impact of operational loss events for US banks and insurers. *Journal of Banking & Finance*, *30*(10), 2605–2634. [CrossRef]

Deephouse, D. L. (2000). Media reputation as a strategic resource: An integration of mass communication and resource-based theories. *Journal of Management*, *26*(6), 1091–1112. [CrossRef]

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019* (pp. 4171–4186). Association for Computational Linguistics. [CrossRef]

Eckert, C., & Gatzert, N. (2017). Modeling operational risk incorporating reputation risk: An integrated analysis for financial firms. *Insurance: Mathematics and Economics*, *72*, 122–137. [CrossRef]

Engle, R. F., Giglio, S., Kelly, B., Lee, H., & Stroebel, J. (2020). Hedging climate change news. *The Review of Financial Studies*, *33*(3), 1184–1216. [CrossRef]

Engle, R. F., & Ruan, T. (2019). Measuring the probability of a financial crisis. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(37), 18341–18346. [CrossRef]

Etter, M., Ravasi, D., & Colleoni, E. (2019). Social media and the formation of organizational reputation. *Academy of Management Review*, *44*(1), 28–52. [CrossRef]

Fiordelisi, F., Soana, M.-G., & Schwizer, P. (2013). The determinants of reputational risk in the banking sector. *Journal of Banking & Finance*, *37*(5), 1359–1371. [CrossRef]

Fiordelisi, F., Soana, M.-G., & Schwizer, P. (2014). Reputational losses and operational risk in banking. *The European Journal of Finance*, *20*(2), 105–124. [CrossRef]

Fombrun, C., & Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Academy of Management Journal*, *33*(2), 233–258. [CrossRef]

Gantchev, N., Giannetti, M., & Li, R. (2022). Does money talk? Divestitures and corporate environmental and social policies. *Review of Finance*, *26*(6), 1469–1508. [CrossRef]

Gatzert, N. (2015). The impact of corporate reputation and reputation damaging events on financial performance: Empirical evidence from the literature. *European Management Journal*, *33*(6), 485–499. [CrossRef]

Gatzert, N., & Schmit, J. (2016). Supporting strategic success through enterprise-wide reputation risk management. *The Journal of Risk Finance*, *17*(1), 26–45. [CrossRef]

Gentzkow, M., Kelly, B., & Taddy, M. (2019). Text as data. *Journal of Economic Literature*, *57*(3), 535–574. [CrossRef]

Gillet, R., Hübner, G., & Plunus, S. (2010). Operational risk and reputation in the financial industry. *Journal of Banking & Finance*, *34*(1), 224–235. [CrossRef]

Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv*, arXiv:2203.05794. [CrossRef]
Heidinger, D., & Gatzert, N. (2018). Awareness, determinants and value of reputation risk management: Empirical evidence from the banking and insurance industry. *Journal of Banking & Finance*, *91*, 106–118. [CrossRef]
Huang, A. H., Wang, H., & Yang, Y. (2023). FinBERT: A large language model for extracting information from financial text. *Contemporary Accounting Research*, *40*(2), 806–841. [CrossRef]
Jiménez-Preciado, A. L., Martínez-García, M. Á., Trejo-García, J. C., & Venegas-Martínez, F. (2025). Short- and long-term assessments of ESG risk in Mexican mortgage institutions: Combining expert surveys, radar plot visualization, and cluster analysis. *Sustainability*, *17*, 5616. [CrossRef]
Kalamara, E., Turrell, A., Redl, C., Kapetanios, G., & Kapadia, S. (2022). Making text count: Economic forecasting using newspaper text. *Journal of Applied Econometrics*, *37*(5), 896–919. [CrossRef]
Kölbel, J. F., Busch, T., & Jancso, L. M. (2017). How media coverage of corporate social irresponsibility increases financial risk. *Strategic Management Journal*, *38*(11), 2266–2284. [CrossRef]
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174. [CrossRef]
Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. [CrossRef]
Loughran, T., & McDonald, B. (2016). Textual analysis in accounting and finance: A survey. *Journal of Accounting Research*, *54*(4), 1187–1230. [CrossRef]
McInnes, L., Healy, J., & Astels, S. (2017). HDBSCAN: Hierarchical density based clustering. *Journal of Open Source Software*, *2*(11), 205. [CrossRef]
McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, arXiv:1802.03426. [CrossRef]
Mitchell, R. K., Agle, B. R., & Wood, D. J. (1997). Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of Management Review*, *22*(4), 853–886. [CrossRef]
Monkkonen, P. (2011). The housing transition in Mexico: Expanding access to housing finance. *Urban Affairs Review*, *47*(5), 672–695. [CrossRef]
Monkkonen, P. (2012). Housing finance reform and increasing socioeconomic segregation in Mexico. *International Journal of Urban and Regional Research*, *36*(4), 757–772. [CrossRef]
Monkkonen, P. (2019). Empty houses across North America: Housing finance and Mexico's vacancy crisis. *Urban Studies*, *56*(10), 2075–2091. [CrossRef]
Perry, J., & de Fontnouvelle, P. (2005). *Measuring reputational risk: The market reaction to operational loss announcements*. Federal Reserve Bank of Boston Working Paper. [CrossRef]
Pérez, J. M., Furman, D. A., Alonso Alemany, L., & Luque, F. (2022). RoBERTuito: A pre-trained language model for social media text in Spanish. In *Proceedings of LREC 2022* (pp. 7235–7243). European Language Resources Association. Available online: https://aclanthology.org/2022.lrec-1.785 (accessed on 21 October 2026).
Pérez, J. M., Rajngewerc, M., Giudici, J. C., Furman, D. A., Luque, F., Alonso Alemany, L., & Martínez, M. V. (2021). pysentimiento: A Python toolkit for opinion mining and social NLP tasks. *arXiv*, arXiv:2106.09462. [CrossRef]
Reyes, A. (2021). Mexico's housing crisis: Vacancy, limited access & deaf policy responses. *International Journal of Urban Sciences*, *25*(Supp. 1), 167–194. [CrossRef]
Serafeim, G., & Yoon, A. (2023). Stock price reactions to ESG news: The role of ESG ratings and disagreement. *Review of Accounting Studies*, *28*(3), 1500–1530. [CrossRef]
Shapiro, A. H., Sudhof, M., & Wilson, D. J. (2022). Measuring news sentiment. *Journal of Econometrics*, *228*(2), 221–243. [CrossRef]
Shiller, R. J. (2017). Narrative economics. *American Economic Review*, *107*(4), 967–1004. [CrossRef]
Sturm, P. (2013). Operational and reputational risk in the European banking industry: The market reaction to operational risk events. *Journal of Economic Behavior & Organization*, *85*, 191–206. [CrossRef]
Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of Finance*, *62*(3), 1139–1168. [CrossRef]
Tobback, E., Naudts, H., Daelemans, W., de Fortuny, E. J., & Martens, D. (2018). Belgian economic policy uncertainty index: Improvement through text mining. *International Journal of Forecasting*, *34*(2), 355–365. [CrossRef]
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 (NeurIPS)* (Vol. 30, pp. 5998–6008). Curran Associates, Inc. Available online: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 21 October 2026).

Witzany, J. (2017). *Credit risk management: Pricing, measurement, and modeling*. Springer.
Wood, D. J., Mitchell, R. K., Agle, B. R., & Bryan, L. M. (2021). Stakeholder identification and salience after 20 years: Progress, problems, and prospects. *Business & Society*, *60*(1), 196–245. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.