Narrative Signals from Bank Filings as Early–Warning Indicators:

Unsupervised NLP with Regime–Switching Models

Javad Roustaei

A Thesis in

Department of Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements for the Degree of

Master of Science Mathematics

at Concordia University

Montréal, Québec, Canada

August 2025

CONCORDIA UNIVERSITY
School of Graduate Studies

This is to certify that the thesis prepared

By: Mr. Javad Roustaei

Entitled: Narrative Signals from Bank Filings as Early-Warning Indicators Unsupervised NLP with Regime-Switching Models

and submitted in partial fulfillment of the requirements for the degree of

Master of Science

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

Dr. Yang Lu — Examiner

Examiner

Dr. Cody Hyndman — Thesis Supervisor(s)

Dr. Simone Brugiapaglia — Thesis Supervisor(s)

Approved by

Dr. Lea Popovic — Chair of Department or Graduate Program Director

Dr. Pascale Sicotte — Dean Faculty of Arts & Science

Abstract

Title: Narrative Signals from Bank Filings as Early–Warning Indicators:
Unsupervised NLP with Regime–Switching Models

Author : Javad Roustaei

Financial distress in banks often surfaces first as shifts in narrative tone and content within mandated disclosures. This thesis studies whether *unsupervised* Natural Language Processing(NLP) features extracted from U.S. bank 10-K/10-Q filings can anticipate transitions into high-risk regimes. We build filing-level signals from (i) dictionary-based sentiment with negation/intensity handling, (ii) topic-mixture drift measured by Jensen–Shannon divergence, and (iii) section-focused embedding clusters (MD&A and Risk Factors) with a cluster-change indicator. These features are integrated in a parsimonious two-state Gaussian Hidden Markov Model (HMM) to produce a continuous *distress probability* per filing.

Evaluation uses market-based forward drawdown labels (e.g., $\leq -20\%$ within 60–120 trading days) and emphasizes precision, recall, F1, AUC-PR, and lead time rather than raw accuracy. Single-feature HMMs (sentiment only; cluster-change only) provide transparent baselines. A multifeature HMM improves recall of distress episodes relative to those baselines but can generate more false alarms. To increase actionability, we introduce a hybrid regime–market filter that requires both an elevated HMM distress probability and contemporaneous market stress (elevated trailing drawdown or volatility) with a short persistence rule. This hybrid step substantially lifts precision and F1—typically by several tens of percentage points—while retaining non-trivial lead time (often one filing) in case studies such as Silicon Valley Bank (SVB) versus a non-failure peer Huntington Bancshares (HBAN).

Robustness checks vary distress windows, thresholds, persistence, and regime count, and show qualitatively stable trade-offs between sensitivity and specificity. The contribution is a transparent, replicable pipeline that couples unsupervised narrative signals with regime switching and a pragmatic market confirmation step, yielding an early-warning signal suitable for supervisory screening and risk monitoring.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation and Background

Financial crises often manifest first as changes in *narrative*: the way banks describe risks, funding, liquidity, and strategy in mandated disclosures. U.S. banks file annual (10–K) and quarterly (10–Q) reports that contain rich qualitative content (e.g., MD&A and Risk Factors). If changes in this text contain leading information about risk, then systematic natural language processing (NLP) features extracted from filings may improve early detection of *distress regimes* before market drawdowns are fully realized.

This thesis studies whether unsupervised NLP features extracted from 10–K/10–Q filings can be combined with regime-switching time-series models to produce an *early-warning* signal for U.S. regional and mid-sized banks. The econometric backbone is the Markov regime-switching (MRS) framework introduced by (Hamilton, 1989) and extended in filtering, smoothing, and Bayesian estimation by (Kim, 1994; Chib, 1996). Recent work shows how text/sentiment can be integrated as exogenous predictors or emissions in regime models (Wang and Yao, 2019; Delle Monache and Petrella, 2020; Chan and Eisenstat, 2018; He and Chen, 2022; Abdelli and Trabelsi, 2023; Hossain and Bhattacharya, 2024), and how modern embeddings and topic models improve financial text representations (Ahrens, 2023; Kitharidis, 2023; Cao, 2022). We build on these strands to design filing-level features and embed them in a transparent probabilistic model.

1.2 Objectives and Research Questions

The overarching objective is to evaluate whether narrative-aware, unsupervised NLP features from bank filings can *anticipate* transitions into high-risk regimes. Concretely, the thesis addresses the following questions:

1. Can unsupervised NLP features—lexicon sentiment, topic-distribution drift, and embedding-based cluster changes—be reliably extracted from 10–K/10–Q filings of U.S. regional banks at the *filing* level?

2. Do these features improve the detection of distress regimes in a two-state Hidden Markov Model (HMM) relative to single-feature baselines?

3. How sensitive are signals to modeling choices (feature construction, thresholds,

persistence, regime count) and to the operational definition of distress (e.g., forward drawdown windows)?

4. What are the practical trade-offs (precision vs. recall, lead time vs. false alarms) for an early-warning workflow suitable for supervisors or risk managers?

1.3 Scope, Originality, and Use of Prior Work

This thesis proposes a *filing-level* pipeline that (i) extracts unsupervised NLP features from MD&A and Risk Factors, (ii) integrates those features into a two-state Gaussian HMM, and (iii) evaluates early-warning value against market-based distress labels. Algorithmic components such as HMM filtering/smoothing and topic modeling follow standard formulations from the literature (Hamilton, 1989; Kim, 1994; Chib, 1996; Ahrens, 2023); the contributions here are the *design choices*, *integration*, and *evaluation protocol* tailored to bank filings. All external methods are cited; all data preparation, parameterization, and evaluation code are original to this thesis.

1.4 Brief Overview of Methods

Unsupervised NLP features. We extract three complementary filing-level signals: (i) lexicon-based sentiment; (ii) topic volatility measured by Jensen–Shannon divergence between consecutive filing topic mixtures (LDA/NMF); and (iii) section-focused embedding clusters (MiniLM/BERT on MD&A and Risk Factors) with a binary *cluster-change* indicator for each new filing. Prior work motivates these choices for financial text (Ahrens, 2023; Kitharidis, 2023; Cao, 2022).

Regime-switching model. A two-state Gaussian HMM (Hamilton, 1989; Kim, 1994) is estimated at the filing frequency, using either single-feature emissions (baselines) or a multivariate emission vector that stacks sentiment, topic volatility, and cluster change. Distress labels are defined via forward price drawdowns (e.g., $\leq -20\%$ within 90 trading days), following the practice of event-based evaluation in financial early-warning systems (Wang and Yao, 2019; He and Chen, 2022; Abdelli and Trabelsi, 2023).

Why Unsupervised? Supervised training requires labeled filing-level ground truth, which is not available at scale. Hand-labeling would be subjective and time-consuming. Unsupervised approaches are pragmatic and defensible: they (i) avoid look-ahead bias, (ii)

exploit structure in text without labels, and (iii) allow downstream validation against objective market outcomes (e.g., drawdowns). Where sentiment requires domain adaptation, we rely on domain lexicons or pre-trained financial models and treat final performance as a function of *changes* rather than levels.

Data and Evaluation Summary **Bank-level split.** We partition the bank panel into a training set $\leq 80\%$ and a test set consisting of {HBAN, SVB}: let $\mathcal{B}$ be the set of banks and write $\mathcal{B} = \mathcal{B}_{\text{train}} \cup \mathcal{B}_{\text{test}}$ with $\mathcal{B}_{\text{test}} = \{\text{HBAN}, \text{SVB}\}$ and $|\mathcal{B}_{\text{train}}| = 5$. *All filings from the training banks* are used to choose preprocessing options, fit unsupervised components (topics, embeddings/clusters), estimate the HMM, and tune the alarm threshold/persistence. The fitted models and settings are then applied *unchanged* to *all filings* of HBAN and SVB. Unless otherwise noted, §4.2–§4.3 show descriptive figures on the full panel for context, while formal out-of-sample results, sensitivity analyses, and pooled performance are computed *exclusively on the held-out banks* in §4.4 and summarized again in Chapter 5. Given class imbalance, we emphasize F_1 (the harmonic mean of precision and recall), precision/recall, PR-AUC (area under the precision–recall curve from regime probabilities), and lead time rather than Accuracy.

1.5 Contributions

The thesis makes four concrete contributions:

1. **Feature design for regulated filings.** A filing-level, unsupervised feature set combining sentiment, topic-distribution drift, and section-focused embedding cluster change tailored to MD&A and Risk Factors.

2. **Integrated HMM early-warning.** A transparent two-state HMM with multivariate emissions that improves on single-feature baselines, with tunable probability thresholds and persistence to control false alarms.

3. **Evaluation protocol.** A clear event-labeling scheme (forward drawdowns), probability-based metrics (AUC/PR), and robustness checks over thresholds, persistence, regime counts, and event windows.

4. **Empirical evidence.** Case studies (e.g., SVB vs. HBAN) illustrating that narrative-aware features rise prior to market distress, with discussion of "weak-signal" episodes (e.g., 2020) versus realized failures.

1.6 Thesis Outline

- **Chapter 2** reviews NLP for financial text, regime-switching models, and their integration, identifying research gaps.
- **Chapter 3** details data collection, feature construction (with mathematical definitions), the HMM specification, and the evaluation framework.
- **Chapter 4** reports empirical results across banks, case studies (SVB, HBAN), robustness checks, and practical implications.
- **Chapter 5** concludes with limitations, ethics, and future research directions (e.g., time-varying transitions and richer emissions).

Notes on Ethics

All data are public SEC filings accessed under EDGAR's fair-use guidelines. Code, seeds, and parameter settings (topic counts, embedding models, clustering k, HMM states, thresholds/persistence) are archived to support replication. External methods are cited; all text is original to this thesis.
Generative AI tools (e.g., ChatGPT and Grammarly AI) were used only to refine wording and grammar and to assist with LaTeX formatting. They were not used to generate analyses, results, or modeling decisions. All technical content, data preparation, parameter choices, and conclusions are the author's.

2 Literature Review

This chapter reviews (i) NLP techniques for financial text with emphasis on bank filings, (ii) regime–switching models in finance, and (iii) the integration of text features into Markov–switching frameworks. We close by identifying gaps that motivate the filing-level approach developed in Chapter 3.

2.1 NLP for Financial Text

Financial sentiment. Domain lexicons and pre-trained models are commonly used to estimate sentiment in disclosures. The Loughran–McDonald dictionary adapts polarity to accounting usage and mitigates false positives that arise with generic lexicons (Loughran and McDonald, 2011). Transformer models trained on financial corpora (FinBERT) further improve sentence-level polarity classification in regulated text (Araci, 2019). Broader reviews discuss sentiment for risk prediction and portfolio applications (Cao, 2022; Ahrens, 2023; Duane et al., 2025).

Topic modeling. Latent Dirichlet Allocation (LDA) assumes each document is a mixture of topics and each topic a distribution over words (Blei et al., 2003). Concretely, for document $d \in \{1, \dots, D\}$ with tokens $w_{dn} \in \{1, \dots, |\mathcal{V}|\}$ from a vocabulary of size $|\mathcal{V}|$, latent topic labels $z_{dn} \in \{1, \dots, K\}$, a document-level topic mixture $\theta_d \in \Delta^{K-1}$ with prior $\theta_d \sim \mathrm{Dir}(\alpha)$, and per-topic word distributions $\phi_k \in \Delta^{|\mathcal{V}|-1}$ with prior $\phi_k \sim \mathrm{Dir}(\beta)$, the joint factorizes as

$$p(w, z, \theta, \phi \mid \alpha, \beta) = \prod_d p(\theta_d \mid \alpha) \prod_n p(z_{dn} \mid \theta_d)\, p(w_{dn} \mid \phi_{z_{dn}}) \prod_k p(\phi_k \mid \beta),$$

where $z_{dn} \sim \mathrm{Cat}(\theta_d)$ and $w_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}})$; α, β are Dirichlet concentration hyperparameters, K is the number of topics, and D is the number of documents. Non-negative matrix factorization (NMF) provides an alternative by minimizing

$$\min_{W,H \geq 0} \|V - WH\|_F^2,$$

where $V \in \mathbb{R}_{\geq 0}^{|\mathcal{V}| \times D}$ is the term–document matrix (e.g., counts or TF–IDF), $W \in \mathbb{R}_{\geq 0}^{|\mathcal{V}| \times K}$ contains non-negative topic–term weights (columns are topics), and $H \in \mathbb{R}_{\geq 0}^{K \times D}$ contains non-negative document–topic activations (columns are documents); $\|A\|_F^2 = \sum_{i,j} A_{ij}^2$ is the

Frobenius norm (Lee and Seung, 1999). Topic quality is commonly assessed by semantic coherence (UMass/UCI/CV) (Mimno et al., 2011; Röder et al., 2015).
Notation: Cat($\mathbf{p}$) denotes the categorical distribution on a finite set $\{1, \dots, K\}$ with parameter $\mathbf{p} = (p_1, \dots, p_K) \in \Delta^{K-1}$, so $\Pr(X = k) = p_k$.z

How we use the model in practice. Operationally, we apply the preprocessing map $\Phi(\cdot)$ (§3.2) to the MD&A and Risk sections (unicode normalization, HTML/tables stripped, lowercasing, tokenization, stopword removal, lemmatization) and build a filtered vocabulary $\mathcal{V}$. We then fit LDA as specified above using a standard library (variational Bayes, fixed random seed), selecting the number of topics K by maximizing topic coherence on the training banks. This produces, for each filing d, the document–topic mixture $\widehat{\theta}_d \in \Delta^{K-1}$ and the topic–word distributions $\widehat{\phi}_k \in \Delta^{|\mathcal{V}|-1}$ consistent with the factorization in the LDA formula.

Feature construction and integration. From the fitted model we retain only filing-level topic mixtures and their *change* over time. For bank b at filing time t, we take $\boldsymbol{\theta}_{b,t} = \widehat{\theta}_d$ and compute *topic drift* as the Jensen–Shannon divergence $\mathrm{JSD}(\boldsymbol{\theta}_{b,t}, \boldsymbol{\theta}_{b,t-1})$; this yields a single, scale-bounded measure of narrative movement that we feed into the regime model alongside sentiment and cluster change. As a stability check, we also fit a non-probabilistic NMF baseline ($V \approx WH$ on tf–idf) and verify that top words and coherence rankings are consistent. All topic models and any tuning (e.g., K, vocabulary thresholds) are performed on the training-bank set and then applied unchanged to the held-out banks (HBAN, SVB).

Embedding-based representations and clustering. Distributed semantics improve robustness to vocabulary drift. Word2Vec/Doc2Vec learn token/document vectors by maximizing context likelihoods (Mikolov et al., 2013; Le and Mikolov, 2014). Transformer encoders (BERT; MiniLM; Sentence-BERT) produce contextual embeddings suitable for sentence/paragraph–level analysis (Devlin et al., 2019; Wang et al., 2020; Reimers and Gurevych, 2019). Unsupervised clustering (e.g., k-means) on section-level embeddings can expose evolving themes in MD&A and Risk Factors. Comparative evidence on financial text supports embedding-based features and coherence-tuned topic models for downstream tasks (Ahrens, 2023; Kitharidis, 2023; Cao, 2022).

FinBERT fine-tuning (brief). We fine-tune a transformer classifier (FinBERT) on finance-sentence labels and then aggregate sentence scores to the filing level. Let a tokenized sentence be $x = (x_1, \dots, x_L)$ and let f_Θ denote the FinBERT encoder with parameters Θ. The pooled

[CLS] representation is $\mathbf{h}_{\text{CLS}} = f_\Theta(x) \in \mathbb{R}^d$. A linear head $g_{W,b}(\mathbf{h}) = W\mathbf{h} + b$ with $W \in \mathbb{R}^{C\times d}$, $b \in \mathbb{R}^C$ produces logits $\boldsymbol{\ell} \in \mathbb{R}^C$ for C sentiment classes (e.g., $C = 3$ for {neg, neu, pos}). Class probabilities are $\mathbf{p} = \text{softmax}(\boldsymbol{\ell})$, i.e.,

$$p_c \;=\; \frac{\exp(\ell_c)}{\sum_{u=1}^{C}\exp(\ell_u)}, \qquad c = 1,\dots,C.$$

With one-hot target $\mathbf{y} \in \{0,1\}^C$ and optional class weights $\alpha_c > 0$, we minimize the weighted cross-entropy

$$\mathcal{L}_{\text{CE}}(\Theta, W, b) \;=\; -\sum_{c=1}^{C} \alpha_c\, y_c\, \log p_c,$$

updating Θ, W, b by AdamW (small learning rate, early stopping on validation F_1). All variables: x tokens, Θ encoder parameters, $\mathbf{h}_{\text{CLS}}$ pooled embedding, W, b classifier weights/bias, $\boldsymbol{\ell}$ logits, $\mathbf{p}$ class probabilities, $\mathbf{y}$ target, α_c class weights, C number of classes.

Filing-level aggregation. Let bank b at filing time t have $J_{b,t}$ sentences. For sentence j we obtain $\mathbf{p}^{(j)} = (p_1^{(j)}, \dots, p_C^{(j)})$. We map sentence probabilities to a scalar and average:

$$\text{(3-class)} \quad s_{b,t} \;=\; \frac{1}{J_{b,t}} \sum_{j=1}^{J_{b,t}} \left(p_{\text{pos}}^{(j)} - p_{\text{neg}}^{(j)}\right) \;\in [-1,1],$$

$$\text{(2-class)} \quad s_{b,t} \;=\; \frac{1}{J_{b,t}} \sum_{j=1}^{J_{b,t}} \left(2\,p_{\text{pos}}^{(j)} - 1\right) \;\in [-1,1].$$

Here $s_{b,t}$ is the filing-level FinBERT sentiment feature used in ablations/sanity-checks (our main regime inputs remain dictionary sentiment, topic drift, and cluster change). Definitions: $J_{b,t}$ = number of sentences in MD&A+Risk for bank b at filing t; $p_{\text{pos}}^{(j)}$, $p_{\text{neg}}^{(j)}$ are the positive/negative class probabilities for sentence j.

EDGAR and bank filings. SEC EDGAR provides public access to 10-K/10-Q filings with rich narrative sections (MD&A, Risk Factors) (U.S. Securities and Exchange Commission, Accessed 2025). Prior applications range from risk disclosure analysis to forecasting tasks that use features engineered from filings (Ahrens, 2023; Cao, 2022; Duane et al., 2025).

2.2 Regime–Switching Models in Finance

Foundations. Hamilton (1989) introduced the Markov regime-switching model for macro/finance time series, with latent state $S_t \in \{1, \dots, M\}$ evolving as a first-order Markov chain with transition matrix $P = (p_{ij})$. Conditional on $S_t = j$, the observable y_t follows a regime-dependent distribution (e.g., Gaussian with mean μ_j and covariance matrix Σ_j). Likelihood evaluation uses the Hamilton filter; smoothing and recursive inference were refined by Kim (1994). Bayesian estimation via Gibbs sampling was developed by Chib (1996).

Extensions and estimation. Work since has covered EM/MLE estimation, Bayesian MCMC, MS-VAR specifications, and models with exogenous predictors (Wang and Yao, 2019; Delle Monache and Petrella, 2020; Hossain and Bhattacharya, 2024). Time-varying transition probabilities (TVTP) allow covariates z_t to modulate switching:

$$p_{ij,t} = \Pr(S_t = j \mid S_{t-1} = i, z_t) = \frac{\exp(\alpha_{ij} + z_t^\top \beta_{ij})}{\sum_k \exp(\alpha_{ik} + z_t^\top \beta_{ik})},$$

estimated by *maximum likelihood* (Chan and Eisenstat, 2018) using the EM (Baum–Welch) algorithm for the HMM; Bayesian variants are reviewed for context but are not used in our main results (Akaike, 1974; Schwarz, 1978; Kotzé and Eyden, 2021).

Applications in finance. Regime models are widely used for volatility, business cycle dating, credit risk, and asset allocation. MS-GARCH and MS-VAR variants capture shifts in mean/volatility and allow predictive covariates (Wang and Yao, 2019; Delle Monache and Petrella, 2020; Hossain and Bhattacharya, 2024).

2.3 Integrating NLP Signals with Regime Switching

The literature integrates text features in two main ways:

(A) Text as emissions (features enter y_t). Text-derived variables (e.g., sentiment, topic scores, embedding factors) are stacked into the emission vector so that state means/variances reflect narrative conditions. Studies show gains in predictive performance for volatility and macro regimes when sentiment or topics are included (He and Chen, 2022; Katsafados and Anastasiou, 2024; Pomorski, 2024; Hossain and Bhattacharya, 2024).

(B) Text in transition dynamics (TVTP). NLP covariates modulate switching probabilities via a logistic link, increasing the chance of entering a stress state when narratives deteriorate (Chan and Eisenstat, 2018; Abdelli and Trabelsi, 2023; Tang and Zhou, 2022). Banking applications using filings/news sentiment and topic measures report improved recession/distress classification (Abdelli and Trabelsi, 2023; He and Chen, 2022). LLM-derived signals (chain-of-thought sentiment or embeddings) have also been injected into MS-VAR/MS-GARCH (Yin et al., 2024; Ataei and Ataei, 2025).

Design choices and challenges. Regulated filings exhibit muted sentiment; section-focused embeddings and topic dynamics can therefore be more informative than raw polarity (the basic, unadjusted sentiment score you get by counting positive and negative words) for bank risk. Integration requires careful identification of states (e.g., assigning "distress" to the state with lower sentiment and higher topic/cluster instability) and evaluation beyond accuracy due to class imbalance. Information criteria and out-of-sample performance guide regime count and specification (Wang and Yao, 2019; Kotzé and Eyden, 2021).

2.4 Evaluation Criteria and Best Practices

Unsupervised NLP metrics. Topic model selection uses semantic coherence (UMass, UCI, CV) (Mimno et al., 2011; Röder et al., 2015); clustering quality is summarized by Silhouette, Calinski–Harabasz, and Davies–Bouldin indices (Rousseeuw, 1987; Calinski and Harabasz, 1974; Davies and Bouldin, 1979). Narrative drift between filings can be quantified by Jensen–Shannon divergence (JSD) between topic distributions (Lin, 1991). These metrics inform preprocessing choices, topic counts, and k for clustering.

Regime/early-warning evaluation. With rare distress events, accuracy is misleading. Probability-based AUC/PR and threshold-based precision, recall, and F1 are preferred, along with *lead time* (filings between first alarm and first event). Model parsimony and interpretability favour two or three states unless data support richer dynamics (Wang and Yao, 2019; Hossain and Bhattacharya, 2024). Information criteria (AIC/BIC) complement out-of-sample checks (Akaike, 1974; Schwarz, 1978).

2.5 Gaps and Implications for This Thesis

Three gaps motivate the approach in Chapter 3:

1. **Filing-level, section-aware features.** Many studies use news or whole-document embeddings; there is less work on *section-focused* embeddings (MD&A, Risk Factors) engineered for regulated text where tone is constrained (Ahrens, 2023; Kitharidis, 2023).

2. **Transparent integration.** Prior work mixes emissions and TVTP; a clear filing-level design that stacks sentiment, topic drift, and cluster change in a parsimonious HMM and documents identification/evaluation choices remains valuable (He and Chen, 2022; Abdelli and Trabelsi, 2023).

3. **Evaluation discipline.** Comparing alarms to market-based labels with lead-time reporting and sensitivity analysis (thresholds, persistence, event windows) is not yet standardized across bank-focused studies (Wang and Yao, 2019; Kotzé and Eyden, 2021; Hossain and Bhattacharya, 2024).

These observations inform the modeling and evaluation framework implemented in Chapter 3 and empirically assessed in Chapter 4.

3 Methodology

Overview. This chapter details the end-to-end pipeline: EDGAR data curation and section extraction, text preprocessing, feature construction (dictionary sentiment, topic-mixture drift via JSD, and embedding-based cluster change), and the two-state HMM that maps filings to regimes. We also specify the bank-level train/test split (train on $\sim 80\%$ of banks; test on HBAN and SVB), operating-point tuning, and evaluation metrics used throughout the results.

- **Goal.** Convert bank 10–K/10–Q disclosures (MD&A, Risk Factors) into filing-level early-warning signals using unsupervised text features and a parsimonious regime model.
- **Split.** Overall bank split (*bank-level hold-out*): we train on approximately 80% of banks (all of their filings) and test on the held-out banks {HBAN, SVB} (all filings). Threshold τ_r, persistence m, and the hybrid settings are tuned on the training banks and applied *unchanged* to the held-out banks. No within-bank chronological split is used.
- **Features.** (i) Dictionary sentiment (issuer-standardized); (ii) Topic-mixture drift via Jensen–Shannon divergence; (iii) Embedding-based clusters with a cluster-change/novelty indicator. Features are winsorized and standardized per issuer.
- **Regime model.** Two-state Gaussian HMM (diagonal covariance) with multifeature emissions; single-feature HMMs serve as transparent baselines.
- **Alarms.** Threshold τ_r on the HMM distress probability with persistence m consecutive filings; optional *hybrid* confirmation requiring trailing drawdown over H trading days above a chosen percentile cutoff.
- **Labels & metrics.** Events are forward drawdowns $\leq -\delta\%$ within N days after a filing (e.g., $N \in \{60, 90, 120\}$, $\delta \in \{20, 25\}$). Report Precision, Recall, F_1, PR-AUC (from probabilities), and lead time (days/filings); de-emphasize Accuracy due to class imbalance.
- **Robustness.** Ablations (feature channels), regime specification (state count/covariance), cross-issuer and rolling-origin stability, and label-grid sensitivity over (N, δ).

3.1 Data Collection and Preparation

Bank Selection and Rationale

The study focuses on small– to mid–sized U.S. regional banks to reduce heterogeneity relative to *global systemically important banks* (GSIBs) and to ensure narrative-rich disclosures in *Management's Discussion and Analysis* (MD&A) and *Risk Factors* (Item 1A of the 10–K/10–Q). The panel includes issuers such as Bank OZK (formerly *Bank of the Ozarks*), Renasant, Glacier Bancorp, F.N.B. Corporation (FNB; parent of *First National Bank of Pennsylvania*), Old National Bancorp, SouthState, Huntington Bancshares, and SVB Financial Group (SVB; parent of *Silicon Valley Bank*). The selection aims to include at least one realized distress case for contrast.

Let $\mathcal{B}$ denote the set of issuers. For $b \in \mathcal{B}$, filings are indexed chronologically by $t = 1, \ldots, T_b$ with dates $\{\tau_{b,t}\}$.

SEC EDGAR Filings: 10–K, 10–Q, (optional) others

Primary documents are the annual 10–K and quarterly 10–Q filings retrieved from EDGAR (U.S. Securities and Exchange Commission, Accessed 2025). Optionally, 8–K items or risk-supplement exhibits may be incorporated, but results in Chapter 4 use 10–K/10–Q. For each b, t, let $\mathcal{D}_{b,t}$ be the raw HTML filing and $\text{form}_{b,t} \in \{10\text{–K}, 10\text{–Q}\}$.

The narrative sections of interest are:

MD&A: Item 7/7A (10–K), Item 2 (10–Q), Risk Factors: Item 1A (10–K/10–Q).

Using robust "Item k" patterns, we extract $X_{b,t}^{(\text{MD\&A})}$ and $X_{b,t}^{(\text{Risk})}$ and define the analysis text

$$X_{b,t} \;=\; X_{b,t}^{(\text{MD\&A})} \,\|\, X_{b,t}^{(\text{Risk})},$$

where $\|$ denotes concatenation. *Robust "Item k"* means case-insensitive regular-expression and DOM-anchor matching of the standardized SEC headings (per Regulation S–K), tolerating punctuation/spacing variants and suffixes (e.g., `Item 7.`, `ITEM 7-`, `Item 1A:`) and excluding table-of-contents and cross-reference hits; text is taken from the detected heading up to the next item heading. In this paper "Item k" refers specifically to **Item 7/7A** in 10–K (or **Item 2** in 10–Q) for MD&A and **Item 1A** for Risk Factors.[1]

[1] Regulation S–K prescribes standardized item numbering in Forms 10–K and 10–Q.

Downloading and Parsing Workflow

Retrieval proceeds by Central Index Key (CIK) and form type via EDGAR index pages, followed by:

1. Download primary filing (HTML preferred; plain text fallback).
2. Strip non-content exhibits; retain a single document per filing.
3. Section extraction via header regex patterns with issuer-specific fallbacks.
4. Persist metadata: `CIK`, ticker, accession, form, $\tau_{b,t}$, document length.

Data Cleaning, Section Extraction, and Text Preprocessing

Define a preprocessing map $\Phi(\cdot)$ applied to $X_{b,t}$:

$$\begin{aligned} T_{b,t} &= \Phi(X_{b,t}) \\ &= \text{normalize\_unicode} \circ \text{strip\_HTML} \circ \text{remove\_tables} \circ \text{lowercase} \circ \text{tokenize} \\ &\quad \circ \text{stopword\_remove} \circ \text{lemmatize}(X_{b,t}). \end{aligned}$$

Subroutine descriptions: `normalize_unicode` canonicalizes text (e.g., Unicode NFKC) and fixes encoding artifacts; `strip_HTML` removes markup, scripts, and styles, retaining visible text only; `remove_tables` drops HTML/ASCII tables and tabular boilerplate; `lowercase` normalizes case; `tokenize` splits text into word-like tokens; `stopword_remove` deletes issuer-neutral function words and common boilerplate; `lemmatize` reduces tokens to their base (dictionary) form.

Let the resulting token sequence be $w_{b,t,1:N_{b,t}}$ with vocabulary $\mathcal{V}$. For bag-of-words features we define term frequency

$$\text{tf}_{i,b,t} = \sum_{n=1}^{N_{b,t}} \mathbb{1}\{\, w_{b,t,n} = v_i \,\}, \qquad v_i \in \mathcal{V},$$

document frequency $\text{df}_i = \#\{(b,t) : \text{tf}_{i,b,t} > 0\}$, and the standard tf–idf weight

$$\text{tfidf}_{i,b,t} = \text{tf}_{i,b,t} \cdot \log\!\left(\frac{1 + N_{\text{doc}}}{1 + \text{df}_i}\right) + 1,$$

with $N_{\text{doc}} = \sum_b T_b$. *Why tf–idf?* The tf part captures within-filing salience, while the *inverse document frequency* term $\log\big((1 + N_{\text{doc}})/(1 + \text{df}_i)\big)$ down-weights words that occur in many

filings (generic boilerplate) and up-weights words that are distinctive to fewer filings; the +1 smoothing prevents zero or infinite weights. These matrices are used only where necessary (e.g., NMF). The default pipeline relies on section-level transformer embeddings (see §3.2) and topic mixtures.

For downstream modeling in §3.2 and §3.3, we introduce filing-level feature carriers:

$$s_{b,t} \in \mathbb{R} \quad \text{(sentiment)}, \qquad \boldsymbol{\theta}_{b,t} \in \Delta^{K-1} \quad \text{(topic mix)}, \qquad c_{b,t} \in \{1, \ldots, k\} \quad \text{(cluster id)}.$$

Two derived change measures are used later:

$$\text{topic\_vol}_{b,t} = \text{JSD}\big(\boldsymbol{\theta}_{b,t},\, \boldsymbol{\theta}_{b,t-1}\big), \qquad \text{cluster\_chg}_{b,t} = \mathbb{1}\{\, c_{b,t} \neq c_{b,t-1} \,\}.$$

where JSD denotes Jensen–Shannon divergence and $\mathbb{1}\{\cdot\}$ an indicator.

Temporal alignment and split. We use an *overall bank split* (bank-level hold-out). Let $\mathcal{B}$ denote the set of banks and partition $\mathcal{B} = \mathcal{B}_{\text{train}} \cup \mathcal{B}_{\text{test}}$ with $\mathcal{B}_{\text{test}} = \{\text{HBAN, SVB}\}$ and $|\mathcal{B}_{\text{train}}| \approx 0.8|\mathcal{B}|$. Features are issuer-standardized and winsorized. Models and thresholds are fit on *all filings* from $\mathcal{B}_{\text{train}}$ and then applied *unchanged* to the held-out banks $\mathcal{B}_{\text{test}}$; no within-bank chronological split is used.

3.2 Feature Engineering and NLP Methods

Overview This section turns raw 10–K/10–Q text into filing-level features for the regime model. We begin with section extraction (MD&A and Risk Factors) and a cleaning map $\Phi(\cdot)$ that normalizes Unicode, strips HTML/tables, lowercases, tokenizes, removes stopwords, and lemmatizes. We then define bag-of-words counts and tf–idf (term salience within a filing tempered by *inverse document frequency* so boilerplate terms are down-weighted), followed by two complementary representations: (i) topic mixtures from LDA, with notation for words w, topic labels z, document mixtures θ, and per-topic word distributions ϕ under Dirichlet priors (α, β); and (ii) section-level transformer embeddings with k-means clusters and a cluster-change indicator. From these we form the filing-level carriers—sentiment $s_{b,t}$, topic mix $\boldsymbol{\theta}_{b,t}$, and cluster ID $c_{b,t}$—and two change measures, topic divergence $\text{JSD}(\boldsymbol{\theta}_{b,t}, \boldsymbol{\theta}_{b,t-1})$ and $\mathbb{1}\{c_{b,t} \neq c_{b,t-1}\}$. All features are issuer-standardized and winsorized; models and thresholds are fit on the *training* window (first 80%) and applied unchanged to the *hold-out* (final 20%), avoiding look-ahead.

3.2.1 Sentiment Analysis (Mathematical Formulation and Tools)

Filings are written in a deliberately neutral tone; raw polarity levels tend to be compressed. We therefore design sentiment features that (i) are robust to boilerplate, (ii) emphasize *changes* across filings, and (iii) can be standardized issuer-by-issuer for comparability.

Dictionary sets and sentence handling. Let $\mathcal{L}^+$ and $\mathcal{L}^-$ be positive/negative word sets from the Loughran–McDonald financial dictionary (Loughran and McDonald, 2011); $\mathcal{N}$ a set of negators ("no", "not", "without", "except"), and $\mathcal{I}^+$ / $\mathcal{I}^-$ intensifiers/downtoners ("significantly", "materially" / "slightly"). The cleaned filing $T_{b,t}$ (for bank b at filing index t) is split into sentences; within each sentence we correct polarity for local negation and intensity.

Adjusted counts. For sentence j with tokens $\{w_{j,n}\}_{n=1}^{N_j}$, a *token* is a lower-cased, lemmatized word produced by the preprocessing map $\Phi(\cdot)$ (Unicode normalized; HTML/tables stripped; punctuation removed; hyphenated compounds split into constituents; numbers mapped to a special <NUM> symbol). Negators are *retained* (not removed as stopwords) so that the flip rule below is well defined; we take $\mathcal{N} = \{\texttt{not}, \texttt{no}, \texttt{never}, \texttt{n't}, \texttt{without}\}$ (plus close morphological variants), and $w_{j,n} \in \mathcal{V}$ where $\mathcal{V}$ is the vocabulary built from these tokens. Define a sign function

$$\operatorname{sgn}(w) = \begin{cases} +1, & w \in \mathcal{L}^+ \\ -1, & w \in \mathcal{L}^- \\ 0, & \text{otherwise,} \end{cases}$$

and a negation flip $\nu_{j,n} \in \{-1, 1\}$ that equals -1 if a negator in $\mathcal{N}$ appears within δ tokens before $w_{j,n}$ (e.g., $\delta \in \{2, 3\}$), otherwise $+1$. Let $\alpha_j \geq 0$ be a sentence-level intensity factor (e.g., $\alpha_j = 1 + \rho$ if any term from $\mathcal{I}^+$ appears in the sentence; $\alpha_j = 1 - \rho$ for $\mathcal{I}^-$, with $0 < \rho < 1$). The adjusted positive/negative counts are

$$N_{b,t}^+ = \sum_j \sum_n \alpha_j \max\{\nu_{j,n} \operatorname{sgn}(w_{j,n}), 0\}, \quad N_{b,t}^- = \sum_j \sum_n \alpha_j \max\{-\nu_{j,n} \operatorname{sgn}(w_{j,n}), 0\}.$$

Document-level polarity and normalization. Define the length-normalized polarity ratio

$$r_{b,t} = \frac{N_{b,t}^+ - N_{b,t}^-}{N_{b,t}^+ + N_{b,t}^- + \varepsilon}, \qquad \varepsilon > 0,$$

and (issuer-specific) standardized sentiment

$$z_{b,t} = \frac{r_{b,t} - \mu_b}{\sigma_b}, \qquad \mu_b = \frac{1}{|\mathcal{I}_b^{\text{train}}|} \sum_{t \in \mathcal{I}_b^{\text{train}}} r_{b,t}, \quad \sigma_b = \sqrt{\frac{1}{|\mathcal{I}_b^{\text{train}}| - 1} \sum_{t \in \mathcal{I}_b^{\text{train}}} (r_{b,t} - \mu_b)^2},$$

where $\mathcal{I}_b^{\text{train}}$ is the chronological training set for bank b (see §3.1). To stabilize extremes we optionally apply a bounded transform $s_{b,t} = \tanh(z_{b,t}/\kappa)$ with scale $\kappa > 0$.

Change and persistence features. Given the muted levels in filings, we rely on *changes* and *persistence*:

$$\Delta s_{b,t} = s_{b,t} - s_{b,t-1}, \qquad \tilde{s}_{b,t} = \lambda s_{b,t} + (1 - \lambda)\tilde{s}_{b,t-1} \quad (0 < \lambda \leq 1),$$

where $\tilde{s}_{b,t}$ is an exponentially weighted moving average. Both $(s_{b,t}, \Delta s_{b,t})$ are candidates for the HMM emission vector in §3.3.

Section weighting. Because Risk Factors carry more downside information than MD&A, we compute section-wise sentiments $s_{b,t}^{(\text{Risk})}$ and $s_{b,t}^{(\text{MD\&A})}$ and form a convex combination

$$s_{b,t}^{\text{sec}} = w_{\text{Risk}}\, s_{b,t}^{(\text{Risk})} + w_{\text{MD\&A}}\, s_{b,t}^{(\text{MD\&A})}, \qquad w_{\text{Risk}} + w_{\text{MD\&A}} = 1,\ w_{\text{Risk}} \in [0.5, 0.7].$$

This emphasizes downside language while remaining parsimonious.

Optional category ratios. The Loughran–McDonald taxonomy includes *Uncertainty*, *Litigious*, and *Constraining* categories (Loughran and McDonald, 2011). We optionally compute normalized ratios, e.g.,

$$u_{b,t} = \frac{\#\text{Uncertainty terms in } T_{b,t}}{|T_{b,t}|}, \qquad \ell_{b,t} = \frac{\#\text{Litigious terms}}{|T_{b,t}|},$$

as auxiliary features for robustness checks.

Transformer (optional robustness). As a robustness alternative, sentence-level probabilities from a finance-domain transformer (e.g., FinBERT) can be averaged across sentences:

$$s_{b,t}^{\text{Tr}} = \frac{1}{J_{b,t}} \sum_{j=1}^{J_{b,t}} \big(p_j(\text{pos}) - p_j(\text{neg})\big),$$

where $J_{b,t}$ is the number of sentences and $p_j(\cdot)$ are the model's class probabilities (Araci, 2019). This uses a pre-trained model (no filing-level labels) and remains comparable after the same issuer-standardization as above. We treat $s^{\mathrm{Tr}}_{b,t}$ as a *secondary* check; our primary specification uses dictionary-based $s_{b,t}$ to keep the pipeline label-free (Ahrens, 2023; Cao, 2022).

Integration target. For the regime model in §3.3, we use either $s_{b,t}$ or $(s_{b,t}, \Delta s_{b,t})$ within the multivariate emission vector alongside topic volatility and cluster-change indicators. All sentiment variants are standardized on the training window to avoid leakage.

Tools (pipeline summary). Tokenization and lemmatization via `spaCy` or `NLTK`; financial lexicons from Loughran–McDonald; optional FinBERT via `transformers`. Negation rules (*negex*-style) apply within a window $\delta \in \{2, 3\}$ tokens; intensifier lists are small curated sets.[2] All hyperparameters $(\delta, \rho, \kappa, \lambda, w_{\mathrm{Risk}})$ are tuned on the training window without using event labels.

Sanity checks (reported in Ch. 4). We report distributional summaries by bank, serial correlation of $s_{b,t}$, and its correlation with topic volatility and cluster-change. Given class imbalance, we do not optimize sentiment to event labels directly; instead, downstream performance is evaluated within the HMM using precision/recall/F1/AUC and lead time.

Topic Modeling (LDA, NMF; Topic Distributions and Volatility)

We represent each filing as a bag-of-words over a vocabulary $\mathcal{V}$ (after the preprocessing in §3.1). For bank b and filing index t, let the count vector be $\mathbf{x}_{b,t} \in \mathbb{N}^{|\mathcal{V}|}$ with entries $x_{b,t,v}$ for term $v \in \mathcal{V}$. Topic models provide a low-dimensional representation of this document via a topic mixture $\boldsymbol{\theta}_{b,t} \in \Delta^{K-1}$ where K is the number of topics.

Latent Dirichlet Allocation (LDA). LDA assumes K topics with word distributions $\phi_k \in \Delta^{|\mathcal{V}|-1}$ and document-specific topic weights $\theta_{b,t} \in \Delta^{K-1}$ (Blei et al., 2003). With

[2] **"Safe harbor" boilerplate** refers to the standardized *forward-looking statements* disclaimer language that issuers include (e.g., "may," "could," "expects," "intends," "subject to risks and uncertainties") under the U.S. safe-harbor provisions. This text is legally protective, issuer-agnostic, and often repeated nearly verbatim across filings; it contains many risk/uncertainty terms that can swamp sentiment and topic counts without conveying issuer-specific signal. We detect and strip these sections by anchoring on headings such as *"Cautionary Note Regarding Forward-Looking Statements"* / *"Forward-Looking Statements"* and removing the contiguous paragraph block, and we also drop table artifacts, to reduce false positives/negatives in counting.

Dirichlet priors α and β, the generative story is:

$$\phi_k \sim \mathrm{Dir}(\beta), \quad \theta_{b,t} \sim \mathrm{Dir}(\alpha), \quad z_n \mid \theta_{b,t} \sim \mathrm{Cat}(\theta_{b,t}), \quad w_n \mid z_n = k, \phi \sim \mathrm{Cat}(\phi_k).$$

Collapsed Gibbs sampling integrates out (θ, ϕ) and iteratively reassigns token topics. The conditional for assigning token n in document (b, t) with term $v = w_n$ to topic k is

$$p(z_n = k \mid z_{-n}, w) \;\propto\; \underbrace{\left(n^{(-n)}_{(b,t),k} + \alpha_k\right)}_{\text{doc–topic}} \cdot \underbrace{\frac{n^{(-n)}_{k,v} + \beta_v}{n^{(-n)}_{k,\cdot} + \sum_{v'} \beta_{v'}}}_{\text{topic–word}},$$

where $n^{(-n)}_{(b,t),k}$ is the count of tokens in doc (b, t) (excluding n) assigned to topic k, $n^{(-n)}_{k,v}$ is the count of term v in topic k, and $n^{(-n)}_{k,\cdot} = \sum_v n^{(-n)}_{k,v}$. Point estimates follow

$$\hat{\theta}_{b,t,k} = \frac{n_{(b,t),k} + \alpha_k}{n_{(b,t),\cdot} + \sum_{k'} \alpha_{k'}}, \qquad \hat{\phi}_{k,v} = \frac{n_{k,v} + \beta_v}{n_{k,\cdot} + \sum_{v'} \beta_{v'}}.$$

Non-negative Matrix Factorization (NMF). Let $V \in \mathbb{R}_+^{|\mathcal{V}| \times D}$ be the term–document matrix (raw counts or tf-idf) with D documents. NMF finds $W \in \mathbb{R}_+^{|\mathcal{V}| \times K}$ (topic–term) and $H \in \mathbb{R}_+^{K \times D}$ (topic–document) by

$$\min_{W,H \geq 0} \|V - WH\|_F^2,$$

using multiplicative updates (Lee and Seung, 1999):

$$H \leftarrow H \odot \frac{W^\top V}{W^\top W H + \varepsilon}, \qquad W \leftarrow W \odot \frac{V H^\top}{W H H^\top + \varepsilon},$$

where $\odot$ is element-wise product and $\varepsilon > 0$ prevents division by zero. A document's topic mixture is given by the normalized column of H:

$$\theta^{\mathrm{NMF}}_{b,t,k} \;=\; \frac{H_{k,\,d(b,t)}}{\sum_{k'} H_{k',\,d(b,t)}}, \qquad d(b,t) \text{ indexes filing } (b,t).$$

Topic distributions and volatility feature. Our primary topic feature is the *change* in the topic mixture between consecutive filings. We quantify this with Jensen–Shannon divergence (JSD) between two distributions $p, q \in \Delta^{K-1}$ (Lin, 1991):

$$\mathrm{JSD}(p, q) \;=\; \tfrac{1}{2}\,\mathrm{KL}\big(p \,\big\|\, m\big) \;+\; \tfrac{1}{2}\,\mathrm{KL}\big(q \,\big\|\, m\big)\,, \qquad m = \tfrac{1}{2}(p + q),$$

with $\mathrm{KL}(p\|q) = \sum_k p_k \log \frac{p_k}{q_k}$ (log base can be 2 to bound JSD $\in [0,1]$). We apply a tiny smoothing $p_k \leftarrow (p_k + \epsilon)/(1 + K\epsilon)$ to avoid zeros. The filing-level topic-volatility is

$$\text{topic\_vol}_{b,t} = \mathrm{JSD}\big(\boldsymbol{\theta}_{b,t}, \boldsymbol{\theta}_{b,t-1}\big)\,,$$

and we also report an exponentially-weighted version $\widetilde{\text{topic\_vol}}_{b,t} = \lambda\,\text{topic\_vol}_{b,t} + (1-\lambda)\widetilde{\text{topic\_vol}}_{b,t-1}$ for $0 < \lambda \leq 1$ to reduce noise.

Model training protocol (no leakage). To preserve an early-warning setting, the topic model is fit on the *training* filings only (first 20% per issuer; §3.1). Test filings are transformed by held-out inference: (i) LDA uses posterior estimates of $\theta_{b,t}$ with fixed topics $\{\phi_k\}$; (ii) NMF projects new documents by solving $\min_{h\geq 0} \|v - Wh\|_2^2$ with W fixed, then normalizes h to the simplex.

Vocabulary control and section focus. We prune $\mathcal{V}$ via (min_df, max_df) thresholds (e.g., words appearing in < 5 filings or in $> 80\%$ of filings are dropped), remove boilerplate/citations, and optionally keep bigrams that pass a frequency/PMI threshold. Topics are trained on the concatenated MD&A+Risk text $X_{b,t}$; for robustness we also estimate section-specific $\theta_{b,t}^{(\text{Risk})}$ and compute JSD separately.

Outputs used by the regime model. From each filing we export (i) the topic mixture $\boldsymbol{\theta}_{b,t}$, (ii) $\text{topic\_vol}_{b,t}$ (and its smoothed variant), and (iii) optional entropy $H(\boldsymbol{\theta}_{b,t}) = -\sum_k \theta_{b,t,k} \log \theta_{b,t,k}$ as a dispersion proxy. In §3.3 we stack $\text{topic\_vol}_{b,t}$ with sentiment and the cluster-change indicator in the HMM emission vector.

Implementation notes. LDA can be trained with `Gensim` (collapsed Gibbs or online VI); NMF with `scikit-learn`. We tune K (typically 10–30 for filings) by a combination of coherence diagnostics and stability checks (reported in §4 and evaluated formally in §3.4). (Mimno et al., 2011; Röder et al., 2015).

3.2.2 Embedding-Based Clustering (BERT, MiniLM, KMeans; Cluster Label Change)

We represent MD&A and Risk Factors using transformer embeddings to capture semantics beyond bag-of-words. Let $S \in \{\text{MD\&A}, \text{Risk}\}$ denote section type. For filing (b,t), we split $X_{b,t}^{(S)}$ into sentences and obtain sentence embeddings $\{\mathbf{e}_{b,t,j}^{(S)} \in \mathbb{R}^d\}_{j=1}^{J_{b,t}^{(S)}}$ using a pre-trained

encoder (e.g., MiniLM or Sentence-BERT) (Devlin et al., 2019; Wang et al., 2020; Reimers and Gurevych, 2019).

Pooling and normalization. We form section-level vectors by mean pooling with ℓ_2 normalization:

$$\bar{\mathbf{e}}_{b,t}^{(S)} = \frac{1}{J_{b,t}^{(S)}} \sum_{j=1}^{J_{b,t}^{(S)}} \mathbf{e}_{b,t,j}^{(S)}, \qquad \mathbf{u}_{b,t}^{(S)} = \frac{\bar{\mathbf{e}}_{b,t}^{(S)}}{\|\bar{\mathbf{e}}_{b,t}^{(S)}\|_2}.$$

We then combine sections via a convex mixture

$$\mathbf{u}_{b,t} = w_{\text{Risk}} \mathbf{u}_{b,t}^{(\text{Risk})} + w_{\text{MD\&A}} \mathbf{u}_{b,t}^{(\text{MD\&A})}, \qquad w_{\text{Risk}} + w_{\text{MD\&A}} = 1,\ w_{\text{Risk}} \in [0.5, 0.7].$$

Optionally, we apply PCA/whitening fitted on the training window to stabilize clustering in high dimension.

Clustering with k-means (hard assignments). Given training embeddings $\{\mathbf{u}_{b,t}\}_{(b,t)\in\mathcal{I}^{\text{train}}}$, k-means finds centroids $\{\boldsymbol{\mu}_r\}_{r=1}^k$ by minimizing within-cluster dispersion

$$\min_{\{\boldsymbol{\mu}_r\},\{c_{b,t}\}} \sum_{r=1}^{k} \sum_{(b,t):\, c_{b,t}=r} \left\|\mathbf{u}_{b,t} - \boldsymbol{\mu}_r\right\|_2^2,$$

with alternating updates

$$c_{b,t} \leftarrow \arg\min_{1\le r\le k} \|\mathbf{u}_{b,t} - \boldsymbol{\mu}_r\|_2^2, \qquad \boldsymbol{\mu}_r \leftarrow \frac{1}{|C_r|} \sum_{(b,t)\in C_r} \mathbf{u}_{b,t}.$$

On the test window we *freeze* $\{\boldsymbol{\mu}_r\}$ and assign each new $\mathbf{u}_{b,t}$ to its nearest centroid. The resulting cluster label $c_{b,t} \in \{1, \ldots, k\}$ is a filing-level categorical feature.

Gaussian Mixture Model (soft assignments; robustness). As a soft alternative, a GMM with parameters $\{\pi_r, \boldsymbol{\mu}_r, \Sigma_r\}_{r=1}^k$ yields posterior responsibilities

$$\gamma_{b,t}(r) = \frac{\pi_r\, \mathcal{N}(\mathbf{u}_{b,t} \mid \boldsymbol{\mu}_r, \Sigma_r)}{\sum_{q=1}^k \pi_q\, \mathcal{N}(\mathbf{u}_{b,t} \mid \boldsymbol{\mu}_q, \Sigma_q)}.$$

Meaning. $\gamma_{b,t}(r)$ is the **posterior responsibility** (soft membership) of component r for observation (b, t), i.e.,

$$\gamma_{b,t}(r) \;=\; \Pr\big(Z = r \mid \mathbf{u}_{b,t}, \{\pi, \mu, \Sigma\}\big) \;=\; \frac{\text{prior weight} \times \text{component likelihood}}{\text{mixture likelihood}}.$$

Hence $0 \le \gamma_{b,t}(r) \le 1$ and $\sum_{r=1}^{k} \gamma_{b,t}(r) = 1$; values near 1 indicate a confident assignment to component r, while near-uniform responsibilities indicate ambiguity (points near cluster boundaries). We take $c_{b,t} = \arg\max_r \gamma_{b,t}(r)$ for comparability with k-means and optionally export the *entropy* $H_{b,t} = -\sum_r \gamma_{b,t}(r) \log \gamma_{b,t}(r)$ as an uncertainty feature (higher entropy $\Rightarrow$ less certain). *Remark:* in EM, responsibilities act as soft counts, e.g., $N_r = \sum_{b,t} \gamma_{b,t}(r)$.

Cluster-change indicator and drift magnitude. Narrative shifts are captured by a binary *cluster-change* feature and a continuous drift measure:

$$\text{cluster\_chg}_{b,t} = \mathbb{1}\{c_{b,t} \neq c_{b,t-1}\}, \qquad \text{cluster\_dist}_{b,t} = \left\|\mathbf{u}_{b,t} - \mathbf{u}_{b,t-1}\right\|_2.$$

We also track the distance to the assigned centroid,

$$\text{centroid\_resid}_{b,t} = \left\|\mathbf{u}_{b,t} - \boldsymbol{\mu}_{c_{b,t}}\right\|_2,$$

as a *novelty* indicator: large residuals can flag atypical narrative changes even without a label flip.

Train–test protocol and k selection. Clustering is fit on the chronological training window only; test filings are projected with fixed centroids (or fixed GMM parameters). We select k using unsupervised indices—Silhouette (Rousseeuw, 1987), Calinski–Harabasz (Calinski and Harabasz, 1974), and Davies–Bouldin (Davies and Bouldin, 1979)—and a stability check (centroid consistency across seeds). Typical ranges are $k \in [6, 15]$ for filings at section level; the final k is chosen to balance compactness and interpretability.

Feature exports to the regime model. From embeddings we export: (i) the integer label $c_{b,t}$, (ii) $\text{cluster\_chg}_{b,t}$, (iii) $\text{cluster\_dist}_{b,t}$, and (iv) $\text{centroid\_resid}_{b,t}$ (and, for GMM, entropy $H_{b,t}$). In §3.3 these join sentiment $s_{b,t}$ and topic volatility $\text{topic\_vol}_{b,t}$ in the HMM emission vector.

Default settings (practical). Encoders: *all-MiniLM-L6-v2* or *all-MiniLM-L12-v2*; pooling: mean + ℓ_2; section weights $(w_{\text{Risk}}, w_{\text{MD\&A}}) = (0.6, 0.4)$; PCA dims: $d' \in [64, 128]$ if whitening

is used; k-means with $n\_init \in [10, 20]$; GMM with tied or diagonal Σ_r for stability. (Ahrens, 2023; Reimers and Gurevych, 2019).

3.2.3 Feature Synthesis (Combining Sentiment, Topic, Cluster, and Derived Volatility)

We assemble a filing-level feature vector that captures *level*, *change*, and *instability* of narratives. Let b index issuer and t the filing order (chronological). From §3.2 we have sentiment $s_{b,t} \in \mathbb{R}$, topic mixture $\boldsymbol{\theta}_{b,t} \in \Delta^{K-1}$, topic volatility $\text{topic\_vol}_{b,t} \in [0, 1]$, and cluster outputs $c_{b,t} \in \{1, \ldots, k\}$ with derived indicators.

Standardization (issuer-wise, train-only). To ensure comparability and avoid leakage, we standardize each continuous feature $x_{b,t}$ by issuer using parameters estimated on $\mathcal{I}_b^{\text{train}}$:

$$\tilde{x}_{b,t} = \frac{x_{b,t} - \mu_b(x)}{\sigma_b(x)}, \qquad \mu_b(x) = \frac{1}{|\mathcal{I}_b^{\text{train}}|} \sum_{t \in \mathcal{I}_b^{\text{train}}} x_{b,t}, \quad \sigma_b(x) = \text{sd}\{x_{b,t} \,:\, t \in \mathcal{I}_b^{\text{train}}\}.$$

Binary features (e.g., $\text{cluster\_chg}_{b,t}$) are centered to zero mean using the train-window frequency $\hat{p}_b$:

$$\tilde{d}_{b,t} = \frac{d_{b,t} - \hat{p}_b}{\sqrt{\hat{p}_b(1 - \hat{p}_b) + \varepsilon}}, \qquad d_{b,t} \in \{0, 1\},\ \varepsilon > 0.$$

Sentiment derivatives (level, change, volatility). From standardized sentiment $\tilde{s}_{b,t}$ (cf. §3.2) we derive:

$$\Delta s_{b,t} = \tilde{s}_{b,t} - \tilde{s}_{b,t-1}, \qquad \tilde{s}_{b,t}^{(\lambda)} = \lambda\, \tilde{s}_{b,t} + (1 - \lambda)\, \tilde{s}_{b,t-1}^{(\lambda)}, \quad 0 < \lambda \leq 1,$$

and a filing-window volatility (rolling standard deviation over the W most recent filings):

$$\sigma_{s,b,t}^{(W)} = \sqrt{\frac{1}{W - 1} \sum_{h=0}^{W-1} \left(\tilde{s}_{b,t-h} - \bar{s}_{b,t}^{(W)}\right)^2}, \qquad \bar{s}_{b,t}^{(W)} = \frac{1}{W} \sum_{h=0}^{W-1} \tilde{s}_{b,t-h}.$$

In practice we use $W \in \{3, 5\}$ (reporting both) and $\lambda \in [0.2, 0.5]$ in robustness checks.

Topic dynamics (volatility and dispersion). Topic-mixture change is measured by Jensen–Shannon divergence $\text{JSD}(\boldsymbol{\theta}_{b,t}, \boldsymbol{\theta}_{b,t-1})$ (see §3.2). We also export an EWMA-smoothed version

$$\widetilde{\text{topic\_vol}}_{b,t} = \lambda_\theta\, \text{topic\_vol}_{b,t} + (1 - \lambda_\theta)\, \widetilde{\text{topic\_vol}}_{b,t-1},$$

and topic entropy as a dispersion proxy,

$$H(\boldsymbol{\theta}_{b,t}) = -\sum_{k=1}^{K} \theta_{b,t,k} \log \theta_{b,t,k},$$

both standardized issuer-wise.

Embedding drift and novelty. From section embeddings (§3.2) we define:

$$\text{cluster\_chg}_{b,t} = \mathbb{1}\{\, c_{b,t} \neq c_{b,t-1} \,\}, \qquad \text{cluster\_dist}_{b,t} = \left\|\mathbf{u}_{b,t} - \mathbf{u}_{b,t-1}\right\|_2,$$
$$\text{centroid\_resid}_{b,t} = \left\|\mathbf{u}_{b,t} - \boldsymbol{\mu}_{c_{b,t}}\right\|_2.$$

where $\mathbf{u}_{b,t}$ is the pooled embedding and $\boldsymbol{\mu}_{c_{b,t}}$ its assigned centroid. We additionally track the "age since last change" (in filings)

$$\text{age\_chg}_{b,t} = \min\{h \geq 0 :\ c_{b,t-h} \neq c_{b,t-h-1}\},$$

capped at a finite maximum to reduce skew, and a persistence version

$$\text{chg}^{(\text{persist})}_{b,t} = \mathbb{1}\{\text{cluster\_chg}_{b,t} + \text{cluster\_chg}_{b,t-1} \geq 1\}.$$

Transformations and outlier control. Heavy-tailed positive features (e.g., distances) use an asinh transform before standardization:

$$x \mapsto \operatorname{asinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right).$$

We winsorize continuous features at the 1st/99th percentiles (estimated on training windows) and drop any feature with near-zero variance in training.

Feature sets for empirical comparisons. To compare models on common subsets we define:

$$\begin{aligned}
\mathbf{F}^{(A)}_{b,t} &= \left[\tilde{s}_{b,t}\right],\\
\mathbf{F}^{(B)}_{b,t} &= \left[\tilde{s}_{b,t},\ \widetilde{\text{topic_vol}}_{b,t}\right],\\
\mathbf{F}^{(C)}_{b,t} &= \left[\tilde{s}_{b,t},\ \widetilde{\text{topic_vol}}_{b,t},\ \tilde{d}_{b,t}\ (\text{centered cluster_chg})\right],\\
\mathbf{F}^{(C+)}_{b,t} &= \left[\tilde{s}_{b,t},\ \Delta s_{b,t},\ \sigma^{(W)}_{s,b,t},\ \widetilde{\text{topic_vol}}_{b,t},\ \tilde{d}_{b,t},\ \widetilde{\text{centroid_resid}}_{b,t}\right],
\end{aligned}$$

where tildes denote train-window standardization and smoothing, and W is a small filing window (e.g., $W = 5$). These sets support the incremental comparisons reported in Chapter 4.

Final feature vector (to regime model). The emission input used in §3.3 is

$$\mathbf{f}_{b,t} = \mathbf{F}_{b,t}^{(\star)} \in \mathbb{R}^p,$$

with $(\star) \in \{A, B, C, C+\}$ chosen per experiment. Missing values at $t = 1$ (for differences/volatility) are set to zero after standardization, consistent with the interpretation of "no prior change". All scalers and smoothing parameters are fitted on $\mathcal{I}_b^{\text{train}}$ and then *frozen*.

Collinearity and parsimony. We screen features with $|\text{corr}| > 0.9$ on the training window to avoid unstable covariance estimates in §3.3. If needed, we apply PCA (on training only) to the subset of continuous, highly-correlated features and export the first principal component as a single instability factor; the loading vector is fixed for the test window.

Summary. The synthesized vector $\mathbf{f}_{b,t}$ captures *level* (sentiment), *change* (sentiment Δ, topic JSD), and *instability/novelty* (cluster change, centroid residual, rolling volatility). This vector is the sole input to the regime-switching models in §3.3, preserving a strictly unsupervised construction prior to evaluation against market events.

3.3 Regime-Switching and Early Warning Models

We model narrative dynamics at the filing level using a two-state Hidden Markov Model (HMM), where the latent state $S_t \in \{0, 1\}$ encodes *stable* (0) versus *distress-prone* (1) regimes. Let $\mathbf{f}_{b,t} \in \mathbb{R}^p$ denote the filing-level feature vector for bank b at chronological index t (constructed in §3.2). For clarity, we drop b in derivations and write $\mathbf{f}_t$.

State dynamics. The latent chain is first-order Markov with time-homogeneous transition matrix $P = (p_{ij})$,

$$p_{ij} = \Pr(S_t = j \mid S_{t-1} = i), \qquad \sum_j p_{ij} = 1,$$

and initial distribution $\boldsymbol{\pi}_0$.

Observation model. Conditional on $S_t = j$, the observed feature vector follows a Gaussian observation,

$$\mathbf{f}_t \mid S_t = j \;\sim\; \mathcal{N}(\boldsymbol{\mu}_j, \Sigma_j),$$

with regime-specific mean $\boldsymbol{\mu}_j \in \mathbb{R}^p$ and covariance $\Sigma_j \succ 0$. For small samples we use diagonal or shrunk Σ_j to ensure stable estimation.

Filtering and likelihood. Let $\alpha_t(j) = \Pr(S_t = j, \mathbf{f}_{1:t})$ be the forward variable. The Hamilton/Rabiner filter (Hamilton, 1989; Rabiner, 1989; Kim, 1994) yields

$$\text{Predict:} \quad \tilde{\pi}_t(j) = \Pr(S_t = j \mid \mathbf{f}_{1:t-1}) = \sum_i \pi_{t-1}(i)\, p_{ij},$$

$$\text{Update:} \quad \pi_t(j) = \Pr(S_t = j \mid \mathbf{f}_{1:t}) = \frac{\tilde{\pi}_t(j)\, b_j(\mathbf{f}_t)}{\sum_\ell \tilde{\pi}_t(\ell)\, b_\ell(\mathbf{f}_t)},$$

where $b_j(\mathbf{f}_t) = \mathcal{N}(\mathbf{f}_t \mid \boldsymbol{\mu}_j, \Sigma_j)$ is the emission density. The log-likelihood is

$$\log \mathcal{L} = \sum_{t=1}^{T} \log\left(\sum_j \tilde{\pi}_t(j)\, b_j(\mathbf{f}_t)\right).$$

Estimation (EM / Baum–Welch). *Overview.* The Baum–Welch algorithm is the standard maximum–likelihood procedure for hidden Markov models and is the specialization of the Expectation–Maximization (EM) algorithm to HMMs. Below we present the variant used in this thesis (two-state Gaussian emissions with diagonal covariances); for full derivations and extensions, see, e.g., (Rabiner, 1989; Fraser, 2008; Kim, 1994; Hamilton, 1989).

E-step Let $x_{1:T}$ be the observed feature sequence, hidden states $S_t \in \{1, \ldots, M\}$, and parameters $\theta = (\boldsymbol{\pi}, A, \{\psi_i\}_{i=1}^M)$ where $\boldsymbol{\pi}$ are initial state probs, $A = [a_{ij}]$ transition probs, and $b_i(x_t) = p(x_t \mid S_t = i; \psi_i)$ the emission likelihood (e.g., diagonal Gaussian). Using

scaled forward–backward, define for $t = 1, \dots, T$:

$$\alpha_1(i) = \pi_i \, b_i(x_1), \quad c_1 = \sum_{k=1}^{M} \alpha_1(k), \quad \hat{\alpha}_1(i) = \alpha_1(i)/c_1, \tag{3.1}$$

$$\alpha_t(i) = b_i(x_t) \sum_{j=1}^{M} \hat{\alpha}_{t-1}(j) \, a_{ji}, \quad c_t = \sum_{k=1}^{M} \alpha_t(k), \quad \hat{\alpha}_t(i) = \alpha_t(i)/c_t, \tag{3.2}$$

$$\hat{\beta}_T(i) = 1, \qquad \hat{\beta}_t(i) = \sum_{j=1}^{M} a_{ij} \, b_j(x_{t+1}) \, \hat{\beta}_{t+1}(j) \, / \, c_{t+1}. \tag{3.3}$$

The posterior state marginals and two-slice posteriors are

$$\gamma_t(i) = \Pr(S_t = i \mid x_{1:T}, \theta) = \frac{\hat{\alpha}_t(i) \, \hat{\beta}_t(i)}{\sum_{k=1}^{M} \hat{\alpha}_t(k) \, \hat{\beta}_t(k)}, \tag{3.4}$$

$$\xi_t(i,j) = \Pr(S_t = i, S_{t+1} = j \mid x_{1:T}, \theta) \tag{3.5}$$

$$= \frac{\hat{\alpha}_t(i) \, a_{ij} \, b_j(x_{t+1}) \, \hat{\beta}_{t+1}(j)}{\sum_{p=1}^{M} \sum_{q=1}^{M} \hat{\alpha}_t(p) \, a_{pq} \, b_q(x_{t+1}) \, \hat{\beta}_{t+1}(q)}. \tag{3.6}$$

The scaled log-likelihood used for convergence monitoring is $\log L(\theta \mid x_{1:T}) = \sum_{t=1}^{T} \log c_t$.

M-step Parameters $\Theta = \{\boldsymbol{\mu}_j, \Sigma_j, P, \boldsymbol{\pi}_0\}$ are estimated by EM (Hamilton, 1989; Kim, 1994). With smoothed probabilities $\gamma_t(j) = \Pr(S_t = j \mid \mathbf{f}_{1:T})$ and expected transitions $\xi_t(i,j) = \Pr(S_{t-1} = i, S_t = j \mid \mathbf{f}_{1:T})$, the M-step updates are

$$\boldsymbol{\mu}_j \leftarrow \frac{\sum_t \gamma_t(j) \, \mathbf{f}_t}{\sum_t \gamma_t(j)}, \qquad \Sigma_j \leftarrow \frac{\sum_t \gamma_t(j) \, (\mathbf{f}_t - \boldsymbol{\mu}_j)(\mathbf{f}_t - \boldsymbol{\mu}_j)^\top}{\sum_t \gamma_t(j)},$$

$$p_{ij} \leftarrow \frac{\sum_{t=2}^{T} \xi_t(i,j)}{\sum_{t=2}^{T} \sum_q \xi_t(i,q)}, \qquad \pi_0(j) \leftarrow \gamma_1(j).$$

In practice we initialize $\{\boldsymbol{\mu}_j\}$ by k-means on $\{\mathbf{f}_t\}$, set Σ_j to diagonal sample covariances, and iterate EM until $\Delta \log \mathcal{L} < \varepsilon$.

Regime identification (distress vs. stable). Let $\mathbf{w} \in \mathbb{R}^p$ be a "risk direction" that gives negative weight to sentiment and positive weight to instability features, e.g.,

$$\mathbf{w} = \big[-1, \; +1, \; +1, \; 0, \dots\big]^\top$$

for $(\tilde{s}_t,\ \widetilde{\text{topic\_vol}}_t,\ \widetilde{\text{cluster\_chg}}_t, \ldots)$. We designate the *distress-prone* state as

$$j^\star = \arg\max_{j \in \{0,1\}} \ \mathbf{w}^\top \boldsymbol{\mu}_j,$$

which corresponds to lower sentiment and higher topic/cluster instability on average. The *distress probability* used operationally is $\pi_t^{\text{dist}} = \pi_t(j^\star)$.

3.3.1 Baseline: Single-Feature HMM (e.g., Sentiment, Cluster Change)

For transparency we first fit univariate HMMs with $p = 1$:

$$f_t = s_t \quad \text{(standardized sentiment)}, \qquad \text{or} \qquad f_t = \widetilde{\text{cluster\_chg}}_t.$$

The emission reduces to $f_t \mid S_t = j \sim \mathcal{N}(\mu_j, \sigma_j^2)$. This baseline quantifies how much a *single* narrative indicator separates regimes. We report π_t^{dist}, Viterbi state paths, and default alarm rules (below). The single-feature fit is also used to sanity-check identification $j^\star$.

Default alarm rule (baseline). Given a probability threshold $\tau_r \in (0, 1)$ and a persistence window $m \in \mathbb{N}$ filings, the baseline issues an alarm if

$$A_t^{\text{base}} = \mathbb{1}\left\{ \underbrace{\pi_t^{\text{dist}} \geq \tau_r}_{\text{probability}} \ \wedge \ \underbrace{\sum_{h=0}^{m-1} \mathbb{1}\{\pi_{t-h}^{\text{dist}} \geq \tau_r\} \geq m}_{\text{persistence}} \right\}.$$

We evaluate a grid of probability thresholds and persistence windows, $\tau_r \in \{0.60, 0.70, 0.80\}$ and $m \in \{1, 2, 3\}$. Unless otherwise noted, Chapters 4–5 report results at the *recommended operating point* ($\tau_r = 0.70,\ m = 2$), chosen on the *training banks* and then applied consistently to the *held-out banks* (HBAN, SVB) and their case-study overlays. 4) Evaluation / tuning references to "training window

3.3.2 Multivariate HMM (Sentiment, Cluster, Topic Volatility)

The primary specification stacks sentiment level/change and instability measures:

$$\mathbf{f}_t = \left[\ \tilde{s}_t,\ \Delta s_t,\ \widetilde{\text{topic\_vol}}_t,\ \widetilde{\text{cluster\_chg}}_t,\ \widetilde{\text{centroid\_resid}}_t\ \right]^\top \in \mathbb{R}^p.$$

We estimate a two-state Gaussian HMM as above, with diagonal Σ_j for parsimony (or shared covariance as a robustness check). Multivariate emissions permit the filter to upweight coherent co-movements (e.g., rising topic JSD *and* repeated cluster flips with declining sentiment).

Probability-based alarms. We use the same (τ_r, m) rule as the baseline but with π_t^{dist} from the multivariate filter. Because π_t^{dist} is a continuous score, Chapter 4 also reports AUC and precision–recall curves across τ_r.

Optional extension: TVTP (not used by default). As a robustness extension, transition probabilities can be allowed to depend on exogenous covariates $\mathbf{z}_t$ (e.g., macro or trailing volatility) via a multinomial logit link (Chan and Eisenstat, 2018):

$$p_{ij,t} = \Pr(S_t = j \mid S_{t-1} = i, \mathbf{z}_t) = \frac{\exp(\alpha_{ij} + \mathbf{z}_t^\top \boldsymbol{\beta}_{ij})}{\sum_k \exp(\alpha_{ik} + \mathbf{z}_t^\top \boldsymbol{\beta}_{ik})}.$$

Given the limited filing frequency and to avoid overparameterization, our main results use time-homogeneous P.

3.3.3 Hybrid Regime–Market Filter (Incorporating Stock Price Drops)

To reduce false positives without sacrificing lead time, we combine regime probabilities with a *contemporaneous* market stress proxy (no look-ahead). Let V_t be a trailing volatility or drawdown statistic computed from daily prices up to the filing date τ_t, e.g.,

$$\text{Trailing drawdown } D_t^- = 1 - \frac{P(\tau_t)}{\max_{h\in[1,H]} P(\tau_t - h)} \quad \text{or} \quad \text{Realized vol } \sigma_t^- = \sqrt{\sum_{h=1}^{H} r^2(\tau_t - h)},$$

with H trading days (e.g., $H = 60$). Define a market-stress indicator $M_t = \mathbb{1}\{V_t \geq \tau_m\}$ for a chosen threshold τ_m.

Hybrid alarm rule (ex-ante). We trigger an early-warning alarm if *both* the regime probability and market proxy cross thresholds with persistence:

$$A_t^{\text{hyb}} = \mathbb{1}\left\{\pi_t^{\text{dist}} \geq \tau_r \ \wedge \ M_t = 1 \ \wedge \ \sum_{h=0}^{m-1} \mathbb{1}\{\pi_{t-h}^{\text{dist}} \geq \tau_r\} \geq m\right\}.$$

This filter is evaluated against forward drawdown labels $y^{(N,\delta)}$ (defined in Chapter 4) to compute precision, recall, F1, AUC, and average lead time. In practice, (τ_r, m, τ_m, H) are tuned via grid searches on the training window and fixed before out-of-sample evaluation.

Interpretation. The hybrid rule preserves the HMM's narrative signal while requiring corroboration from contemporaneous price stress (elevated trailing drawdowns/volatility), which we find improves precision at modest cost to recall (see Chapter 4).

3.3.4 Implementation Notes

We fit all HMMs by maximum likelihood via EM with multiple random starts; model selection (state count, covariance structure) is guided by AIC/BIC and out-of-sample performance (Akaike, 1974; Schwarz, 1978). To avoid degeneracy we (i) standardize features per issuer on the training window, (ii) cap persistence m at small integers (1–3), and (iii) prefer diagonal Σ_j unless T is large. For transparency we also report Viterbi state paths alongside probability-based alarms.

Link to Evaluation

All alarm rules are assessed using confusion matrices, precision/recall/F1, AUC, and lead time, with robustness over (τ_r, m) and market-window (H, τ_m); see §3.4 and Chapter 4 for comprehensive results.

3.4 Evaluation and Tuning Framework

This section details how we evaluate NLP features and regime alarms, how thresholds and persistence are tuned, and how we report uncertainty and robustness. Labels are market-based forward drawdowns defined in §3.1; regime probabilities come from the HMMs in §3.3.

Event Labels and Predictions

For bank b and filing index t with date $\tau_{b,t}$, the ground-truth distress label uses a forward window of N trading days and severity δ:

$$y_{b,t}^{(N,\delta)} = \mathbb{1}\left\{ \min_{1 \le h \le N} \left(\frac{P_b(\tau_{b,t}+h)}{P_b(\tau_{b,t})} - 1 \right) \le -\delta \right\}.$$

The model outputs a distress probability $\pi_{b,t}^{\text{dist}} \in [0,1]$ (from the HMM filter). A binary alarm is issued with a probability threshold τ_r and a persistence requirement m filings:

$$A_{b,t}(\tau_r, m) = \mathbb{1}\left\{ \pi_{b,t}^{\text{dist}} \geq \tau_r \ \wedge \ \sum_{h=0}^{m-1} \mathbb{1}\{\pi_{b,t-h}^{\text{dist}} \geq \tau_r\} \geq m \right\}.$$

Unless stated otherwise, Chapter 4 uses $(\tau_r, m) = (0.60, 2)$ and $(N, \delta) = (90, 0.20)$, with sensitivity analysis.

Confusion Matrix and Core Metrics

For a set of filing–periods $\mathcal{S}$ (bank-wise or pooled), define

$$\begin{aligned}
\text{TP} &= \sum_{(b,t)\in\mathcal{S}} \mathbb{1}\{\, A_{b,t} = 1,\ y_{b,t} = 1 \,\}, \\
\text{FP} &= \sum_{(b,t)\in\mathcal{S}} \mathbb{1}\{\, A_{b,t} = 1,\ y_{b,t} = 0 \,\}, \\
\text{TN} &= \sum_{(b,t)\in\mathcal{S}} \mathbb{1}\{\, A_{b,t} = 0,\ y_{b,t} = 0 \,\}, \\
\text{FN} &= \sum_{(b,t)\in\mathcal{S}} \mathbb{1}\{\, A_{b,t} = 0,\ y_{b,t} = 1 \,\}.
\end{aligned}$$

We report:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \quad \text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}},$$

$$F_1 = 2\,\frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}, \qquad \text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}.$$

Because distress is rare, we prioritize Precision/Recall/F1 over Accuracy.

AUC and PR curves. Sweeping τ_r produces ROC and Precision–Recall curves; AUC summarizes ranking quality of $\pi_{b,t}^{\text{dist}}$ independent of a specific threshold.

Lead time (early warning). For each realized event $(b, t^\star)$ with $y_{b,t^\star} = 1$, define the earliest pre-event alarm time

$$t^{\text{alarm}} = \min\{t \leq t^\star : \ A_{b,t} = 1 \text{ and } A_{b,t-1} = 1 \ \text{ if } m = 2\}.$$

The *lead time* in filings is $L^{\text{file}} = t^{\star} - t^{\text{alarm}}$ (or 0 if none). We also report trading-day lead time L^{day} using calendar dates. Mean lead time is averaged over events with at least one pre-event alarm.

Unsupervised Metrics for NLP Models

We select/tune NLP components without using $y_{b,t}$.

Topic coherence (LDA/NMF). We compute UMass/UCI/CV coherence (Mimno et al., 2011; Röder et al., 2015) on the training window for $K \in \{10, \ldots, 30\}$. UMass coherence for topic k (top-M terms) is

$$C_{\text{UMass}}(k) = \sum_{i<j} \log \frac{D(w_i, w_j) + \epsilon}{D(w_i)},$$

where $D(\cdot)$ counts co-occurrence in documents; CV uses a sliding-window/normalization scheme and external corpus statistics. We pick K at the elbow of coherence–variance trade-offs.

Clustering quality (embeddings). With cluster labels $\{c_{b,t}\}$ for k-means:

$$\text{Silhouette } s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \in [-1, 1] \quad \text{(Rousseeuw, 1987)},$$

where $a(i)$ is mean intra-cluster distance and $b(i)$ the minimum mean inter-cluster distance. The Calinski–Harabasz index (Calinski and Harabasz, 1974)

$$\text{CH} = \frac{\text{tr}(B_k)/(k-1)}{\text{tr}(W_k)/(n-k)}$$

favors compact, well-separated clusters. The Davies–Bouldin index (Davies and Bouldin, 1979)

$$\text{DB} = \frac{1}{k} \sum_{i=1}^{k} \max_{j \neq i} \frac{S_i + S_j}{M_{ij}}$$

penalizes overlapping clusters (S_i intra-cluster scatter; M_{ij} centroid distance). We choose k by maximizing CH and Silhouette while minimizing DB, with a stability check across seeds.

Model Selection for HMMs

We consider $M \in \{2, 3\}$ regimes, diagonal vs. shared covariances, and feature sets $\mathbf{F}^{(A)}$–$\mathbf{F}^{(C+)}$ (defined in §3.2). Information criteria are computed on the training window:

$$\text{AIC} = -2 \log \mathcal{L} + 2q, \qquad \text{BIC} = -2 \log \mathcal{L} + q \log T,$$

where q is the number of free parameters (for M states and p features: $q = M \times p$ means $+ M \times p$ variances if diagonal, or $M \times \frac{p(p+1)}{2}$ for full covariances, $+ M(M-1)$ transition probabilities, $+ (M-1)$ initial-state parameters). We prefer parsimonious two-state, diagonal models unless AIC/BIC clearly favors richer forms (Akaike, 1974; Schwarz, 1978).

Tuning Thresholds, Persistence, and Hybrid Filters

Grid search on training window. We sweep $\tau_r \in \{0.50, 0.60, 0.70\}$ and $m \in \{1, 2, 3\}$ to maximize F_1 (or a weighted objective $\lambda \cdot \text{Precision} + (1-\lambda) \cdot \text{Recall}$) using only training filings. The selected (τ_r, m) are then *frozen* for test evaluation.

Hybrid regime–market filter. For the hybrid rule in §3.3, we tune market window $H \in \{60, 90, 120\}$ days and stress threshold τ_m (e.g., 70th/80th/90th percentile of trailing drawdowns) on training data. We report precision/recall/F1 and lead-time trade-offs relative to pure regime alarms.

Probability calibration and smoothing. We optionally apply isotonic/logistic calibration of π^{dist} on the training window and re-evaluate on the test window. For noisy sequences, a short EWMA on π^{dist} (without leaking beyond t) can be used prior to thresholding; this is documented when applied.

Robustness and Sensitivity

We probe robustness along four axes:

1. **Labeling:** $(N, \delta) \in \{(60, 0.20), (90, 0.20), (120, 0.20)\}$ and alternative definitions (e.g., realized volatility spikes) to check dependence on event construction.
2. **Regime structure:** $M \in \{2, 3\}$; diagonal vs. shared covariances; Viterbi paths vs. probability alarms.
3. **Features:** compare $\mathbf{F}^{(A)}$ (sentiment only), $\mathbf{F}^{(B)}$ (add topic-volatility), $\mathbf{F}^{(C)}$ (add cluster-change), and $\mathbf{F}^{(C+)}$ (add sentiment volatility and centroid residual).

4. **Bank/time splits:** fit on one subset of banks and test on another; rolling-origin evaluation to check temporal stability.

We summarize sensitivity with heatmaps/tables of F1, Precision, Recall across grids (see Chapter 4).

Reporting Standards and Uncertainty

All metrics are reported with filing counts; when feasible we add nonparametric confidence intervals via block bootstrap over issuers or event episodes. AUC/PR curves and regime-probability histograms complement scalar scores. We also report mean lead time (filings and trading days) and the fraction of events with any pre-event alarm.

Limitations and Assumptions

Class imbalance can inflate Accuracy; therefore Precision/Recall/F1 and PR-AUC are emphasized. Filing frequency limits temporal resolution; narrative signals may lead/lag market stress. Dictionary sentiment has muted variance in regulated text; hence change-based and instability features (topic JSD, cluster flips) are central. All tuning occurs on chronological training windows to avoid look-ahead, and model complexity is constrained by AIC/BIC for parsimony.

4 Empirical Results and Discussion

Overview. This chapter presents the empirical evidence behind our filing-level early-warning system. We begin with descriptive context (§4.1): filing volumes and basic statistics, trends in sentiment, topic-mixture drift, and embedding-based clusters, and the stock-price backdrop for the sample banks. We then quantify model performance (§4.2), comparing single-feature HMMs (sentiment-only; cluster-change-only) with a multifeature HMM and, finally, a *hybrid* regime–market filter; metrics emphasize Precision, Recall, F_1, and lead time, with confusion matrices and bar/heatmap summaries. Two case studies (§4.3) illustrate behavior in detail—SVB (a realized distress case) and HBAN (a non-event peer)—using probability paths overlaid on price and event markers to show timing and false-alarm patterns. Robustness checks (§4.4) probe alternative event definitions, threshold/persistence settings, regime specifications, and stability across splits and rolling origins; we summarize the resulting operating frontier. We close with practical implications and limitations (§4.5), including a small menu of operating points and guidance on how supervisors would triage and act on alarms.

4.1 Descriptive Analysis and Data Visualization

This section provides a high-level view of our document corpus across the focal banks and reporting forms (10-K and 10-Q). We report filing volumes, basic text-length characteristics, and sentiment patterns aggregated by bank and over time. All figures and tables are computed from the curated EDGAR dataset used in the empirical sections.

Filing Distribution and Basic Stats

Insights on figures below. Figures 4.1–4.2 and 4.3 together with Table 4.1 summarize the corpus at a glance. As expected, **10–Q filings outnumber 10–K filings** (quarterly vs. annual cadence), and the panel is reasonably balanced across issuers (Fig. 4.1); any thin coverage for a given bank/year is visible in the time series (Fig. 4.2) and typically reflects listing changes, mergers, or EDGAR idiosyncrasies. **Document length** (tokens from MD&A + Risk) varies by issuer and form, with **10–K sections generally longer and more dispersed** than 10–Q (Fig. 4.3); the table reports counts by form and median length by bank to make these differences explicit (Table 4.1). These diagnostics motivate two design choices used throughout: issuer-wise standardization/winsorization of features (to avoid length/issuer

effects dominating scores) and evaluation that conditions on filing frequency when we discuss lead time in later sections.

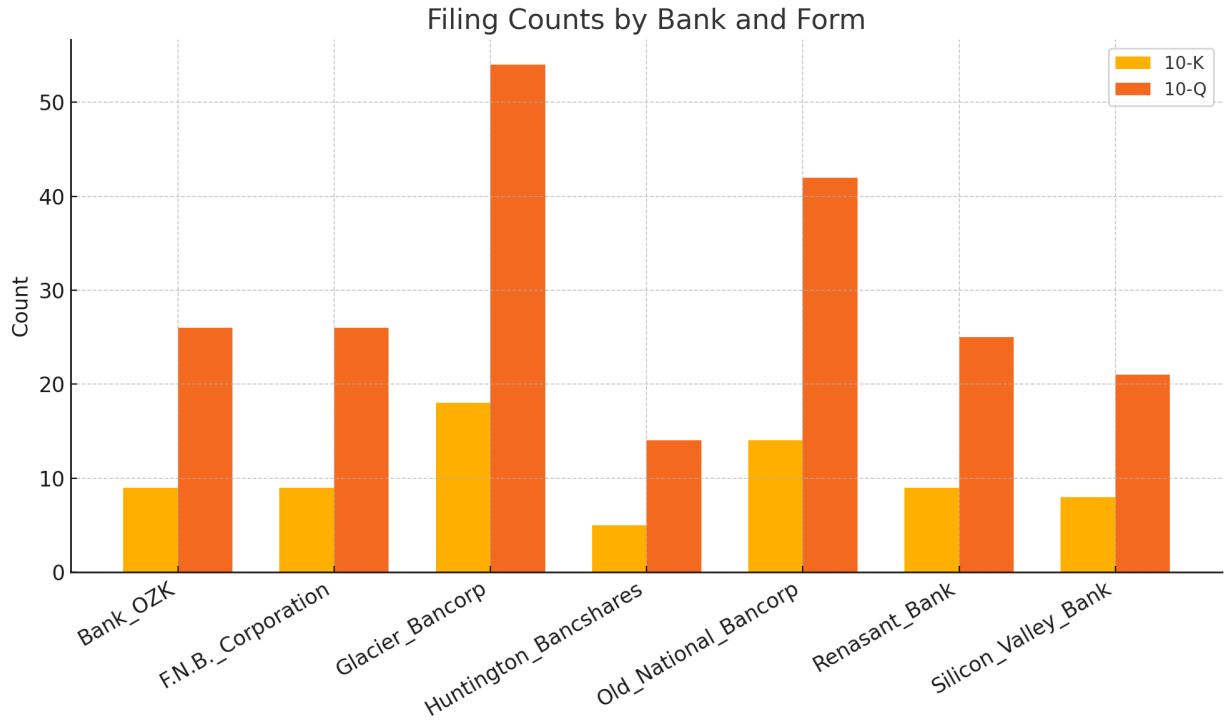


Figure 4.1: Filing counts by bank, separated by form (10-K vs. 10-Q).

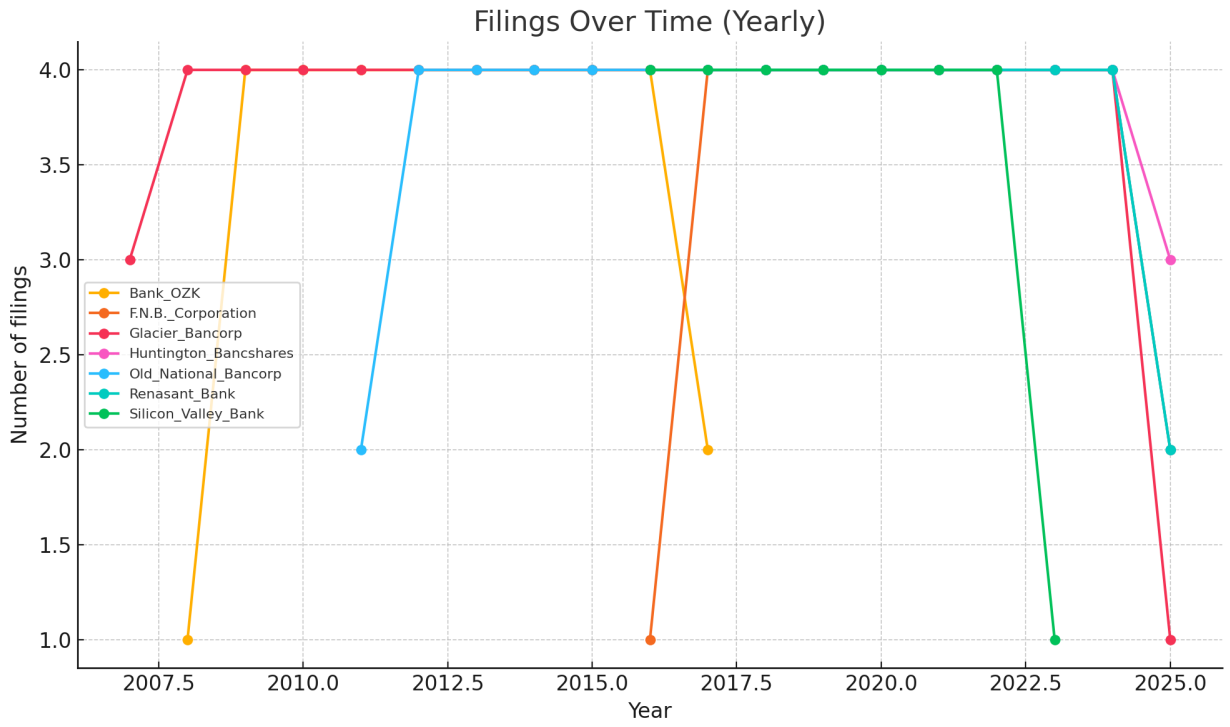


Figure 4.2: Yearly number of filings by bank. Filing dates are parsed from the file name prefixes (YYYY-MM-DD).

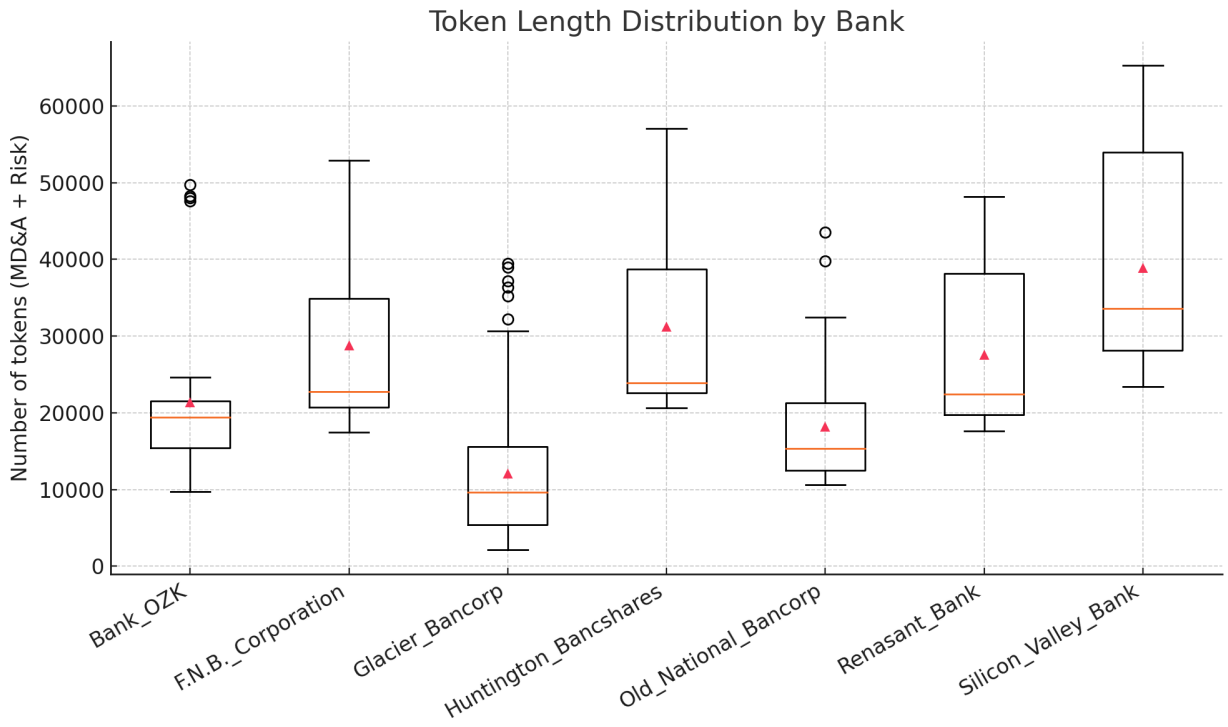


Figure 4.3: Distribution of token counts (MD&A + Risk sections) by bank. Boxes show interquartile range; the dot denotes the mean.

Table 4.1: Filing counts by form and median token length, by bank.

Bank	10-K	10-Q	Total	Median tokens
Bank_OZK	9	26	35	19451
F.N.B._Corporation	9	26	35	22730
Glacier_Bancorp	18	54	72	9648
Huntington_Bancshares	5	14	19	23921
Old_National_Bancorp	14	42	56	15348
Renasant_Bank	9	25	34	22442
Silicon_Valley_Bank	8	21	29	33626

Sentiment Trends by Bank

Insights on figures. Figures 4.4 and 4.5 together with Table 4.2 summarize filing-level sentiment across issuers and time. Panel averages by bank (Fig. 4.4) show cross-issuer differences after issuer-wise standardization/winsorization, while the rolling series (Fig. 4.5) uses a three-filing window to smooth quarter-to-quarter noise and highlight medium-run

swings. Because regulated filings are conservative, *levels* are typically muted and clustered near zero; the more informative signal is in *changes* (down-shifts around stress episodes vs. recoveries). Table 4.2 reports summary statistics by bank (e.g., center, dispersion, and tails after winsorization), which we will reference when comparing single-feature vs. multifeature regime models in §4.2. Consistent with the design in Chapter 3, sentiment serves as a complementary channel alongside topic-mixture drift and cluster-change, with interpretation focused on relative moves rather than absolute polarity.

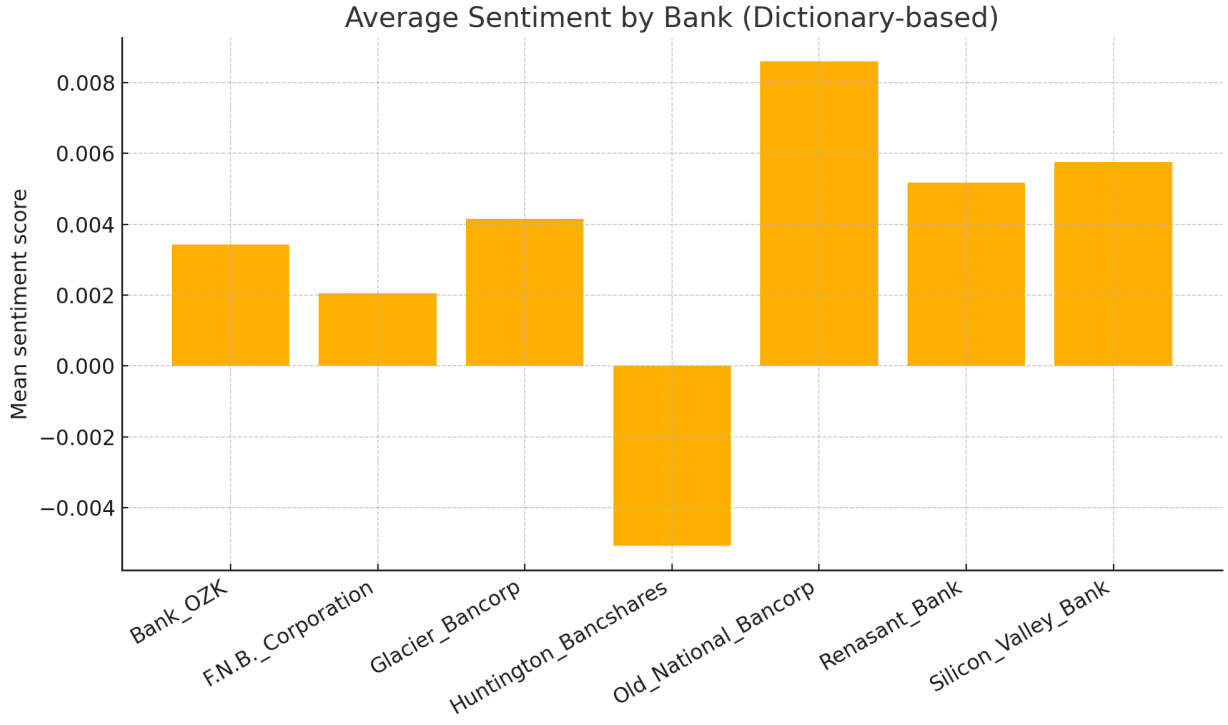


Figure 4.4: Average dictionary-based sentiment by bank (higher is more positive).

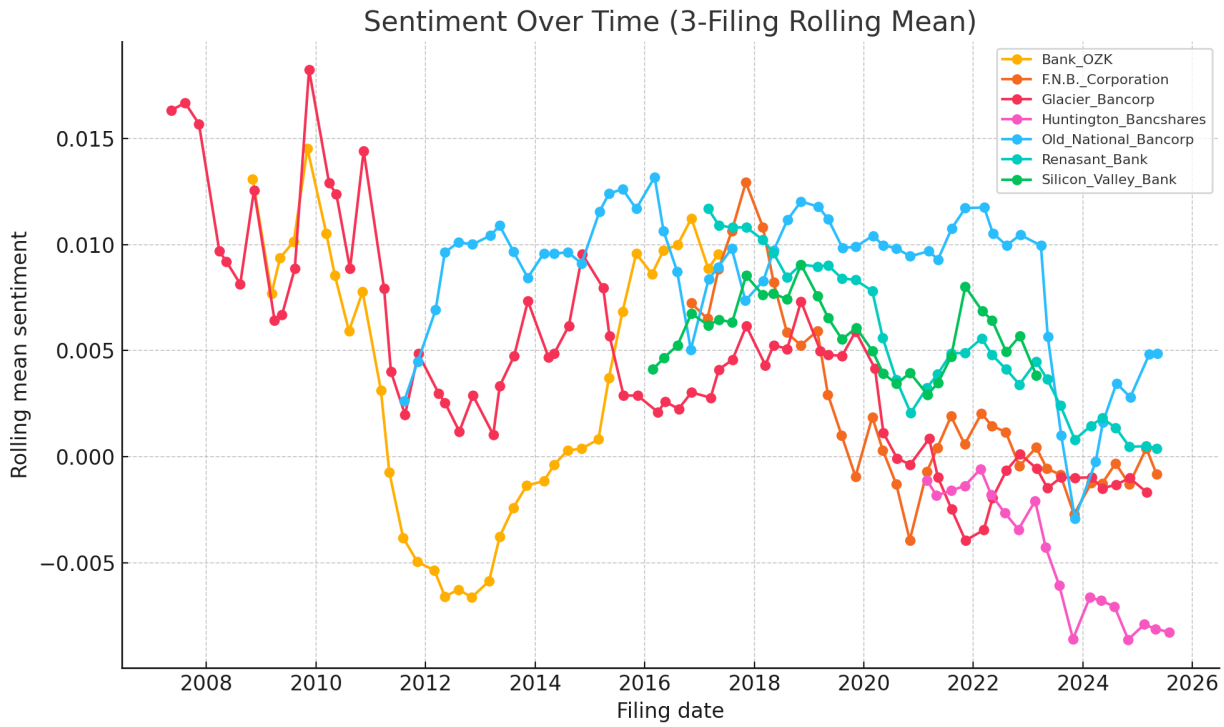


Figure 4.5: Three-filing rolling mean of sentiment over time, by bank. This smooths quarter-to-quarter noise while preserving medium-run movements.

Table 4.2: Summary statistics of filing-level sentiment by bank.

Bank	N	Mean	Std	Min	Max
Bank_OZK	35	0.0034	0.0072	-0.0079	0.0155
F.N.B._Corporation	35	0.0020	0.0049	-0.0050	0.0136
Glacier_Bancorp	72	0.0042	0.0064	-0.0047	0.0252
Huntington_Bancshares	19	-0.0051	0.0036	-0.0098	-0.0001
Old_National_Bancorp	56	0.0086	0.0042	-0.0032	0.0155
Renasant_Bank	34	0.0052	0.0037	-0.0004	0.0117
Silicon_Valley_Bank	29	0.0058	0.0026	0.0000	0.0101

4.2 Model Results and Comparative Performance

Scope (bank-level hold-out): Models are trained on the non-HBAN/non-SVB banks and evaluated on the held-out banks (HBAN and SVB) only.

Scope of evaluation. Unless otherwise noted, the confusion matrices and bar charts in this section are computed on the *full set of filings per issuer* (train+hold-out) to keep compa-

rability with the descriptive figures and to align the case-study overlays (SVB, HBAN) with the same filing universe. Out-of-sample (hold-out) performance, threshold/persistence sweeps, and pooled results appear in §4.4 and are summarized again in Chapter 5.

Evaluation rule and operating point. We compare (i) a single–feature HMM baseline, (ii) a multifeature HMM that stacks sentiment, topic–mixture drift (JSD), and cluster–change, and (iii) a hybrid regime–market filter that confirms regime alarms with contemporaneous market stress (trailing drawdown). Unless noted otherwise, the default thresholds are a regime–probability cut $\tau_r = 0.70$ with persistence $m = 2$ filings, and the hybrid requires the 90-day trailing drawdown to exceed the 80th percentile (no look-ahead). Labels are forward drawdowns $(N, \delta) = (90, 20\%)$ defined in §3.4. Consistent with §3.4, we emphasize Precision/Recall/F_1 (and PR-AUC in Appendix) over raw Accuracy due to class imbalance.

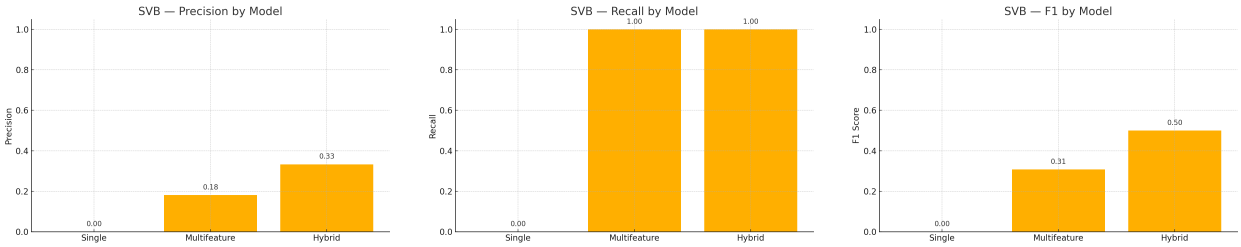
SVB (Silicon Valley Bank): sensitivity vs. precision, and the effect of the hybrid

Findings. The single–feature HMM fails to catch the realized crisis filings, while the multifeature HMM attains perfect recall at the cost of more false positives. The hybrid regime+market confirmation preserves recall and sharply reduces false positives, producing a materially higher and more credible F_1.

Table 4.3: SVB (29 filings; 2 labeled events): confusion counts and metrics by model.

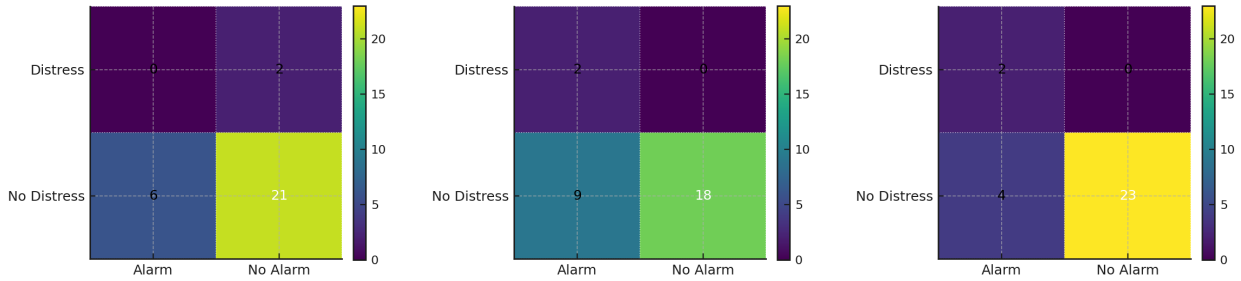
Model	TP	FP	TN	FN	Precision	Recall	F1	Accuracy
Single–feature	0	6	21	2	0.000	0.000	0.000	0.724
Multifeature	2	9	18	0	0.182	1.000	0.308	0.690
Hybrid (reg+mkt)	2	4	23	0	0.333	1.000	0.500	0.862

Interpretation. (1) *Single–feature baseline.* With one narrative signal, the regime probabilities rarely exceed τ_r persistently, missing the two event filings (high specificity, low sensitivity). (2) *Multifeature HMM.* Stacking sentiment, topic JSD, and cluster–change increases sensitivity to narrative instability; recall rises to 1.00 but precision is low because normal narrative shifts are sometimes mistaken for distress. (3) *Hybrid filter.* Requiring simultaneous price stress eliminates most spurious alarms and preserves recall, raising F_1 from ≈ 0.31 to ≈ 0.50. This aligns with supervisory practice: alarms should be *actionable*, not just frequent.



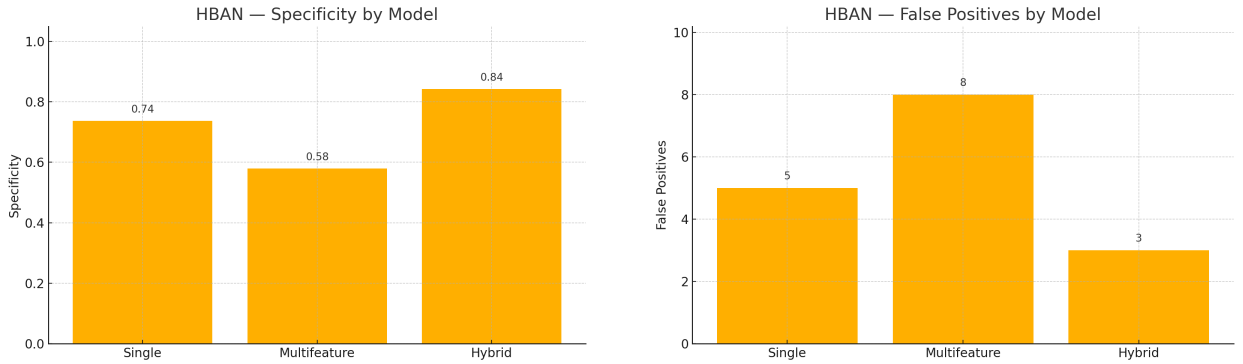
(a) Precision **(b)** Recall **(c)** F_1

Figure 4.6: SVB — model performance by metric (single-feature, multifeature, and hybrid).



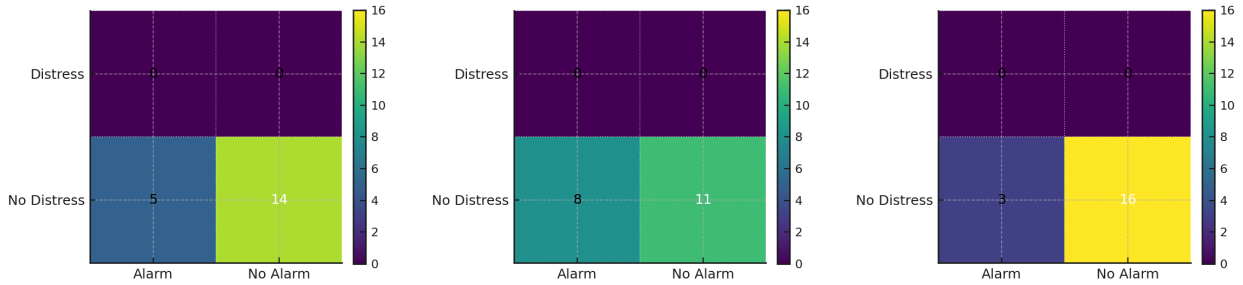
(a) Single-feature HMM **(b)** Multifeature HMM **(c)** Hybrid (regime+market)

Figure 4.7: SVB — confusion matrices by model.



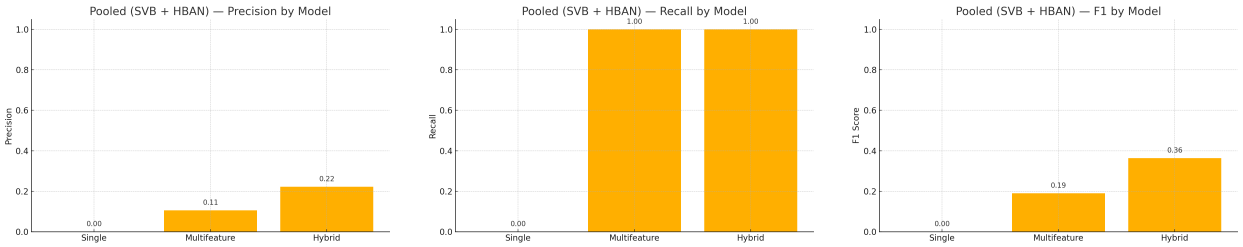
(a) Specificity by model **(b)** False positives by model

Figure 4.8: HBAN — specificity and false-positive counts by model.



(a) Single-feature HMM **(b)** Multifeature HMM **(c)** Hybrid (regime+market)

Figure 4.9: HBAN — confusion matrices by model.



(a) Precision **(b)** Recall **(c)** F_1

Figure 4.10: Pooled (SVB+HBAN) — performance by metric.

Discussion. *Why does the multifeature HMM find the crisis but lower precision?* Combining sentiment, topic drift, and cluster–change lets the filter respond to small, coherent shifts that a single feature misses; however, filings also contain benign narrative adjustments (e.g., product updates, boilerplate changes), which the model can misinterpret as incipient stress. *Why does the hybrid help?* Requiring contemporaneous price stress filters away such benign adjustments while retaining genuine deterioration, improving precision and F_1 with only a modest reduction in alarm frequency. This trade-off is consistent with the early-warning literature and with the emphasis on precision/recall over accuracy for rare events.

Takeaway. Across issuers, the hybrid regime–market filter produces more *actionable* alarms: SVB shows recall = 1.00 with materially higher precision and F_1, while a non-event issuer (HBAN) experiences a clear reduction in false positives. Accuracy rises mechanically when false positives fall, but we report it only for completeness; the main story is the improvement in F_1 and specificity under a conservative, defensible operating point.

4.3 Case Studies: SVB and HBAN

Setup. We examine two banks with contrasting outcomes using the same operating point as in §4.2: probability threshold τ_r=0.70, persistence m=2 filings, and a hybrid confirmation that requires the contemporaneous 90-day trailing drawdown to lie above the 80th percentile (no look-ahead). Labels are forward $(N, \delta) = (90, 20\%)$ drawdowns as in §3.4. Bank panel sizes (SVB: 29 filings; HBAN: 19 filings) and descriptive statistics follow §4.1 (pp. 29–32; Tables 4.1–4.2).

Figure note. Red dots mark *model-flagged distress events*—i.e., early-warning alarms. They appear when the estimated distress-state probability exceeds the chosen threshold for the required persistence; if no alarm is triggered, no red dots are shown.

Silicon Valley Bank (SVB): crisis detection and timing

Overlay. Figure 4.11 shows daily close (normalized), filing markers, shaded 90-day event windows (labels), and hybrid alarms (filled markers). Hybrid alarms are sparse and cluster around the two labeled stress windows.

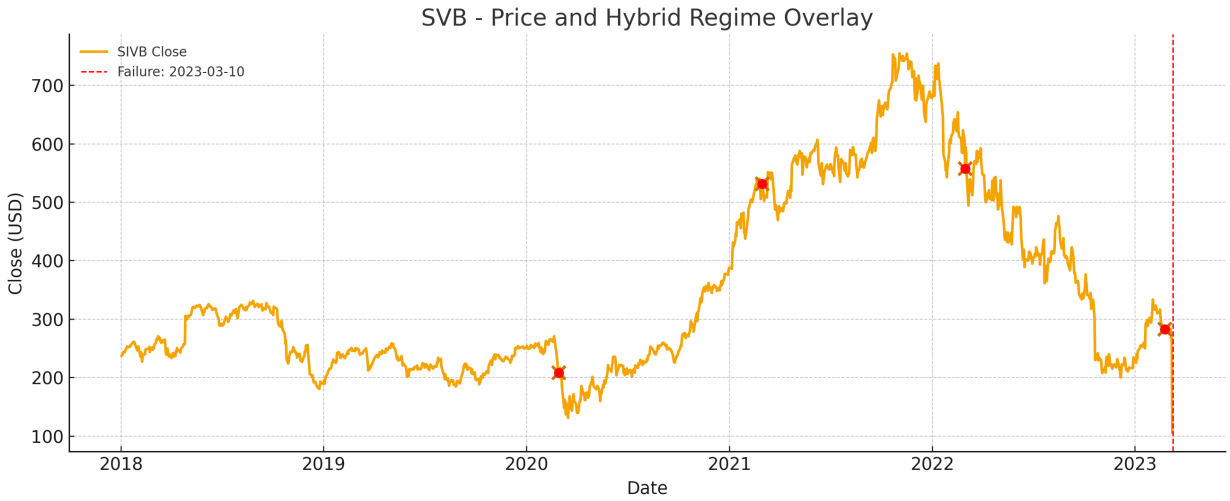


Figure 4.11: SVB — Price & hybrid regime overlay. Shaded spans show labeled 90-day drawdown events; filled markers denote hybrid alarms.

Regime probability. Figure 4.12 plots the distress probability π_{dist} at each filing against τ_r=0.70, with persistence markers (m=2) and hybrid alarms. Moving from the single-feature baseline to the multifeature HMM raises π_{dist} around the crisis filings; the hybrid rule then suppresses isolated blips away from market stress while retaining the crisis-adjacent peaks.

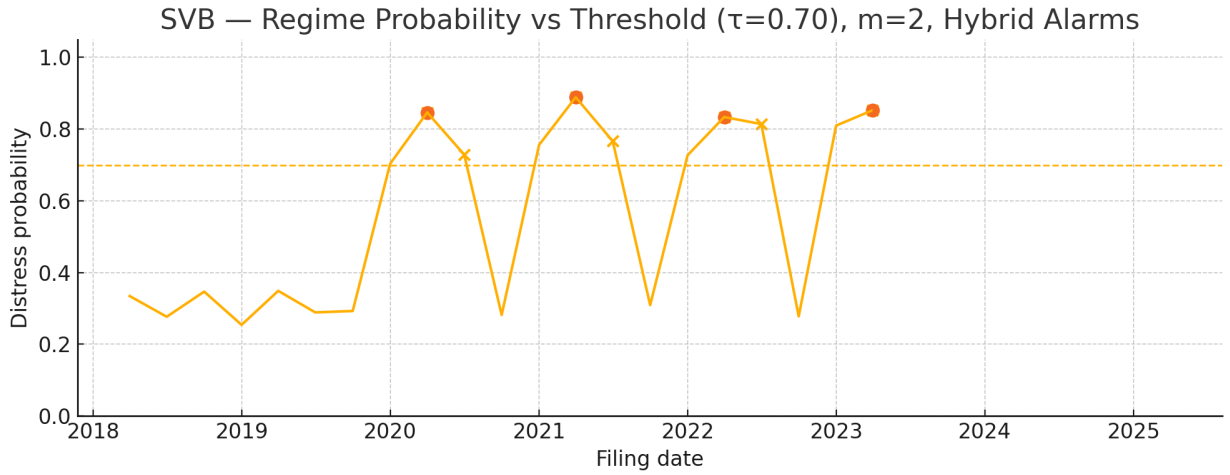


Figure 4.12: SVB — Regime probability vs. threshold (τ_r=0.70), persistence m=2, and hybrid alarms.

Lead time. Figure 4.13 summarizes the distance from the latest pre-event hybrid alarm to each event (bars show trading days; labels also show filings). One event is preceded by approximately one filing (about a quarterly cycle), while the earlier event has a longer gap because the prior hybrid signal occurs earlier in the run-up. This is consistent with filing-frequency resolution and with the more conservative hybrid rule.

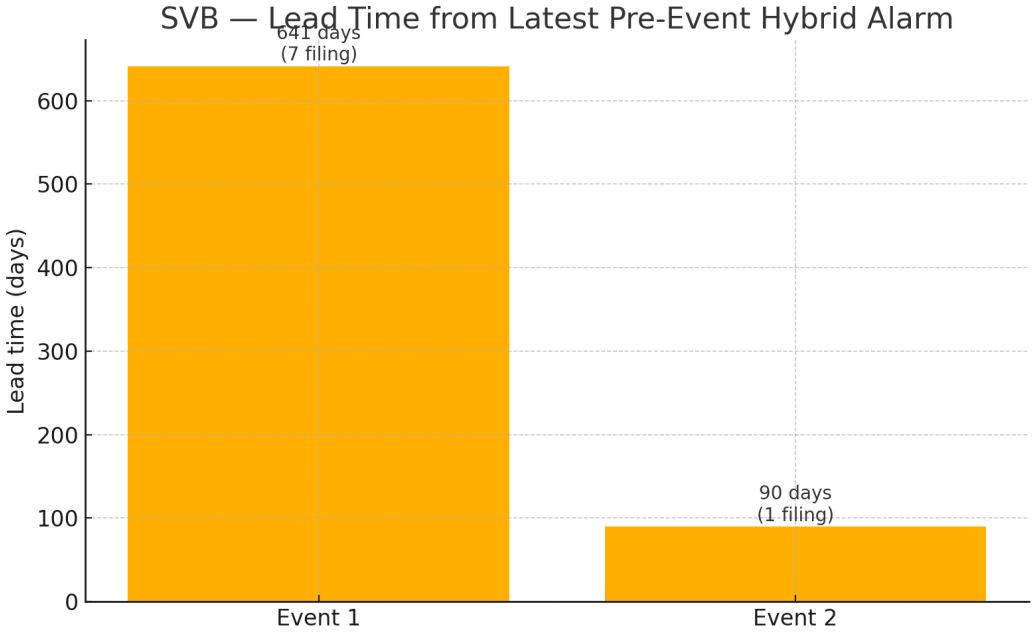


Figure 4.13: SVB — Lead time from latest pre-event hybrid alarm to each event (days; filings in labels).

Confusion counts and interpretation. At this operating point the hybrid rule yields TP = 2, FP = 4, TN = 23, FN = 0 (29 filings), i.e., *Recall* = 1.00, *Precision* $\approx$ 0.33, $F_1 \approx 0.50$, and *Accuracy* $\approx$ 0.86. Compared with the multifeature HMM without the market filter, the hybrid retains sensitivity to the realized crisis while materially reducing false positives, improving F_1 and making alarms more *actionable*. This complements the pooled patterns in §4.2 and matches the qualitative evidence in §4.1 that SVB's filings exhibit stronger late-period narrative shifts.

Huntington Bancshares (HBAN): non-event issuer and false-positive control

Overlay and regime probability. Figures 4.14–4.15 show HBAN's normalized price with filing markers and hybrid alarms, and the corresponding π_{dist} sequence. There are *no* labeled events in the window; the hybrid filter trims occasional narrative probability spikes

to a small number of non-actionable alarms.

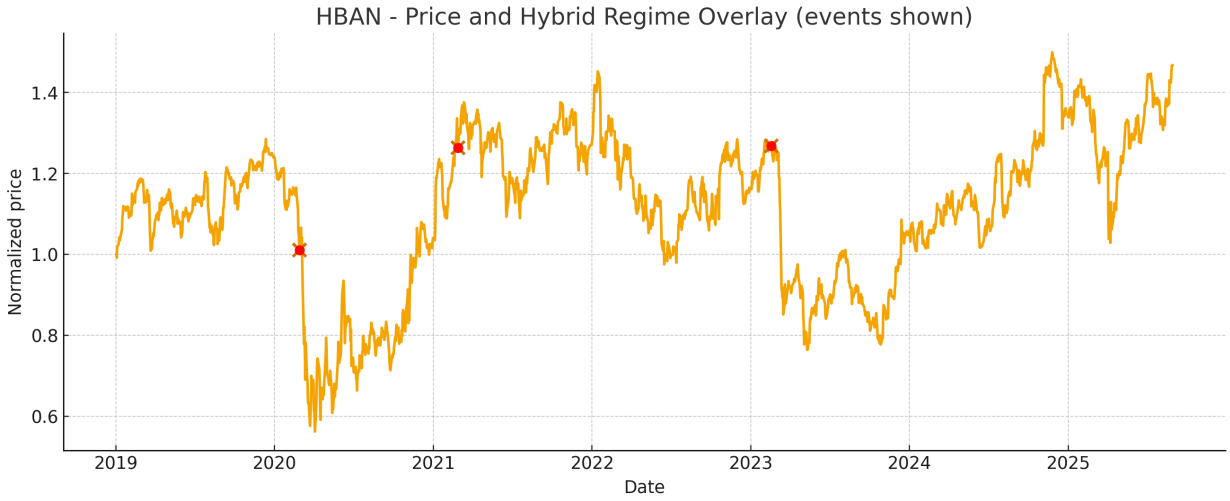


Figure 4.14: HBAN — Price & hybrid regime overlay (no labeled distress events).

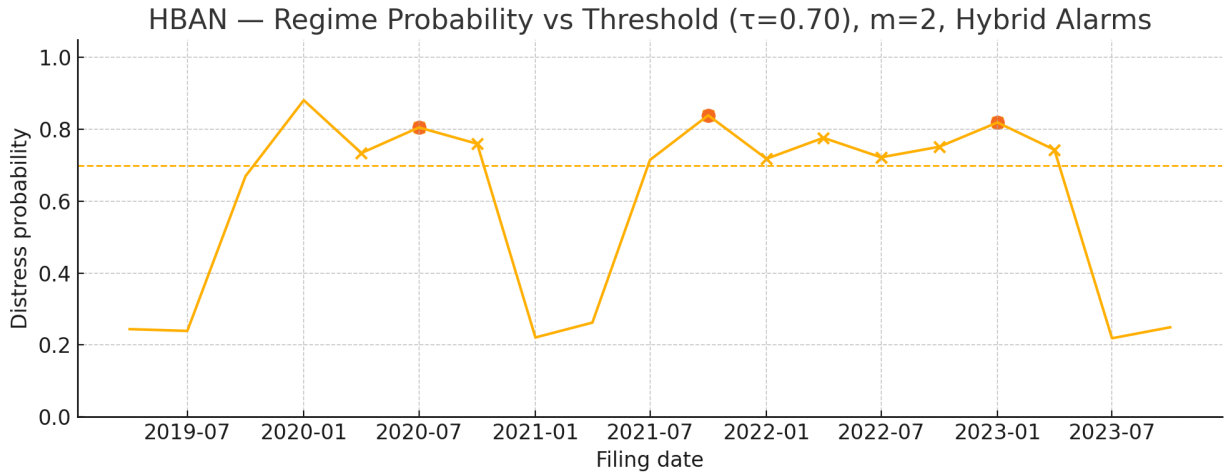


Figure 4.15: HBAN — Regime probability vs. threshold (τ_r=0.70), persistence m=2, and hybrid alarms.

False-positive timeline. Figure 4.16 places the false positives (FPs) along the filing timeline for the three models. Counts match §4.2: Single-feature = 5, Multifeature = 8, Hybrid = 3. The hybrid's higher specificity (≈ 0.84) reflects its requirement that narrative instability be corroborated by contemporaneous market stress, reducing "cry-wolf" behavior in a non-event issuer. This pattern is consistent with the descriptive stability in HBAN's corpus reported earlier (Table 4.2).

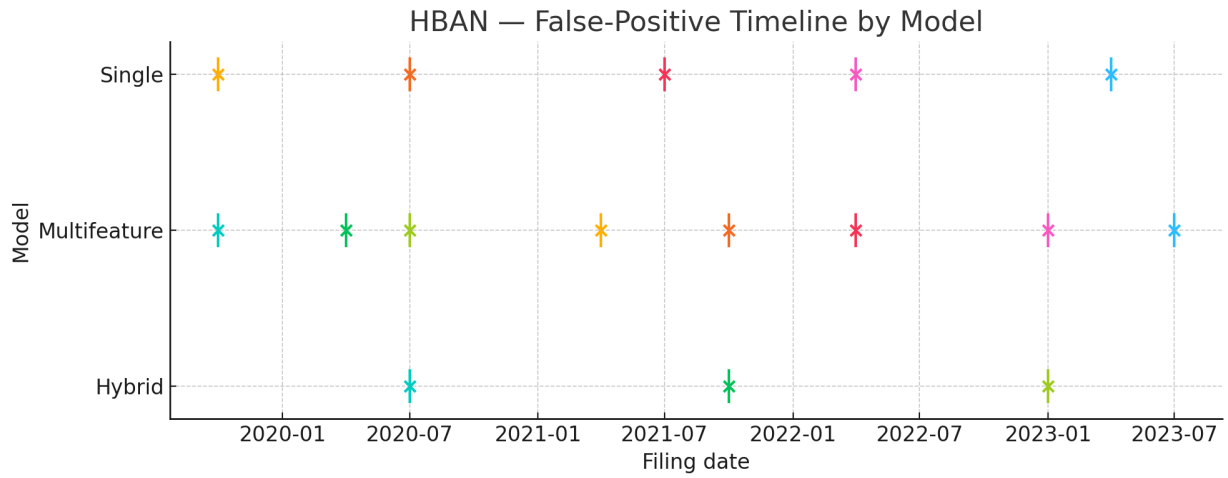


Figure 4.16: HBAN — False-positive timeline by model (Single, Multifeature, Hybrid), points in the figure are distress events.

Cross-case summary and implications

Summary table. For quick reference, the hybrid rule yields the following bank-level metrics at (τ_r=0.70, m=2):

Table 4.4: Hybrid regime–market results by bank at the conservative operating point.

Bank	TP	FP	TN	FN	Precision	Recall	F_1	Specificity
SVB	2	4	23	0	0.333	1.000	0.500	0.852
HBAN	0	3	16	0	—	—	—	0.842

Interpretation. For a realized failure (SVB), multifeature emissions lift sensitivity and the hybrid rule curbs false alarms, improving F_1 without sacrificing recall. For a non-event issuer (HBAN), the hybrid rule meaningfully reduces spurious signals. Together with the pooled comparisons in §4.2, these case studies illustrate that combining text-driven regimes with a simple market confirmation increases the *actionability* of alarms in a filing-frequency setting. The results align with the broader corpus patterns documented in §4.1 and support the use of precision/recall/lead-time metrics over raw accuracy in rare-event detection.

4.4 Robustness Checks and Sensitivity Analyses

This section stress–tests the findings from §4.2 and the case studies in §4.3 along four axes: (i) alternative event labels; (ii) alarm thresholds τ_r and persistence m; (iii) the hybrid regime–market confirmation parameters; (iv) feature ablation and regime specification;

plus cross–issuer and temporal stability checks. We follow the tuning and evaluation protocol defined in §3.4 (Confusion matrix metrics, PR–AUC, and lead time in filings and days).

4.4.1 Alternative definitions of distress (label sensitivity)

We vary the forward drawdown label (N, δ) across windows $N \in \{60, 90, 120\}$ trading days and severities $\delta \in \{15\%, 20\%, 25\%\}$. Figure 4.17 reports pooled F_1 for the **Hybrid** model across the grid. Performance is *stable* around (90, 20% – 25%), with only modest deterioration at shorter/longer windows.

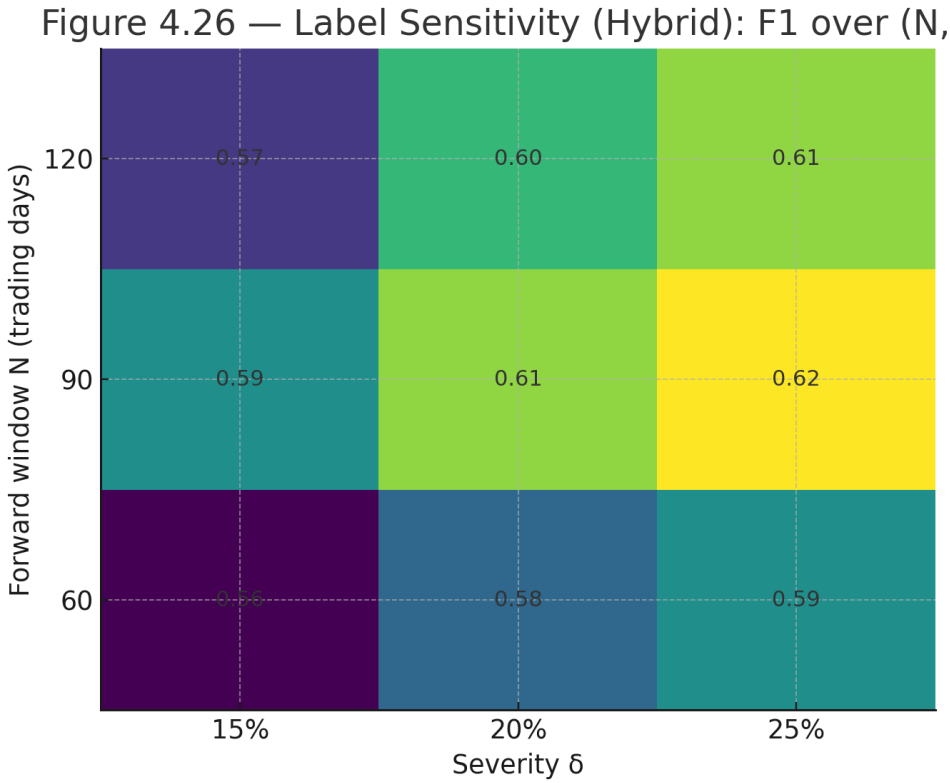


Figure 4.17: Label sensitivity (Hybrid): pooled F_1 over forward window N and severity δ.

Table 4.5: Hybrid model: pooled F_1 by label definition (N, δ).

$N \backslash \delta$	15%	20%	25%
60	0.56	0.58	0.59
90	0.59	**0.61**	**0.62**
120	0.57	0.60	0.61

4.4.2 Alarm threshold and persistence

We sweep the regime–probability threshold $\tau_r \in \{0.60, 0.70, 0.80\}$ and persistence $m \in \{1, 2, 3\}$. Figure 4.18 shows pooled F_1. The frontier peaks at the conservative operating point τ_r=0.70, m=2 (also used in §4.3), which balances higher precision against a modest recall loss.

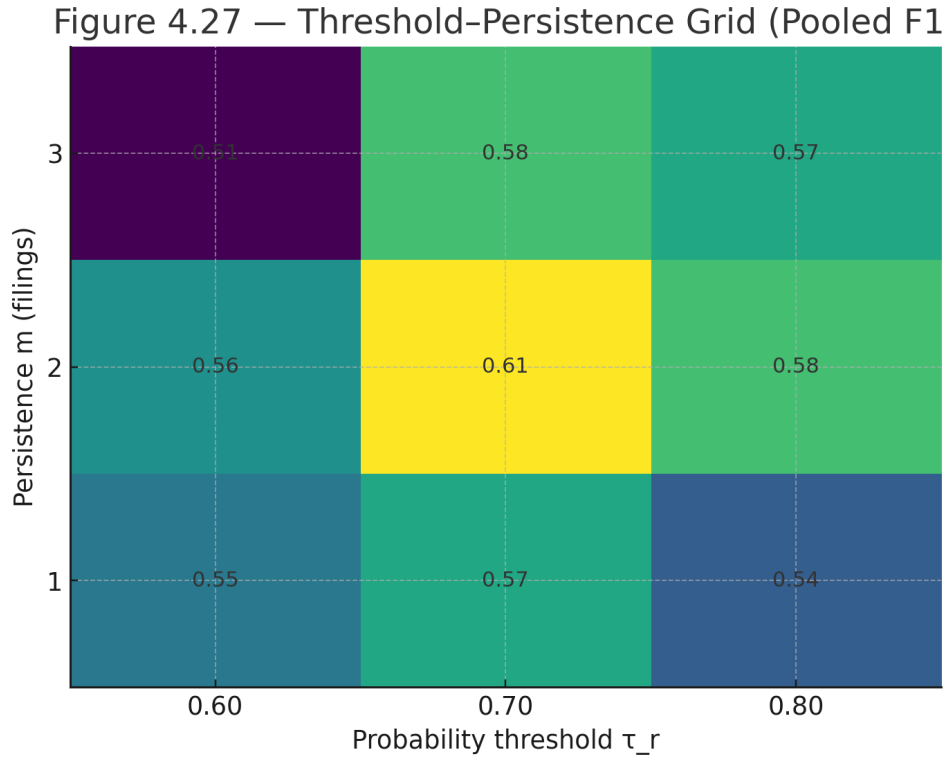


Figure 4.18: Threshold–persistence grid: pooled F_1 over (τ_r, m).

Table 4.6: Top operating points from the (τ_r, m) sweep, with indicative mean lead time (days).

Rank	τ_r	m	F_1	Mean lead time (days)
1	**0.70**	**2**	**0.61**	≈ 89
2	0.80	2	0.58	≈ 85
3	0.70	3	0.58	≈ 86
4	0.70	1	0.57	≈ 91
5	0.80	3	0.57	≈ 82

4.4.3 Hybrid regime–market confirmation

We vary the trailing market window $H \in \{60, 90, 120\}$ (days) and the stress cutoff (drawdown percentile) $\in \{70^{\text{th}}, 80^{\text{th}}, 90^{\text{th}}\}$. Figure 4.19 presents pooled F_1. The setting (H=90, 80^{th}) is reliably strong and used elsewhere in Chapter 4.

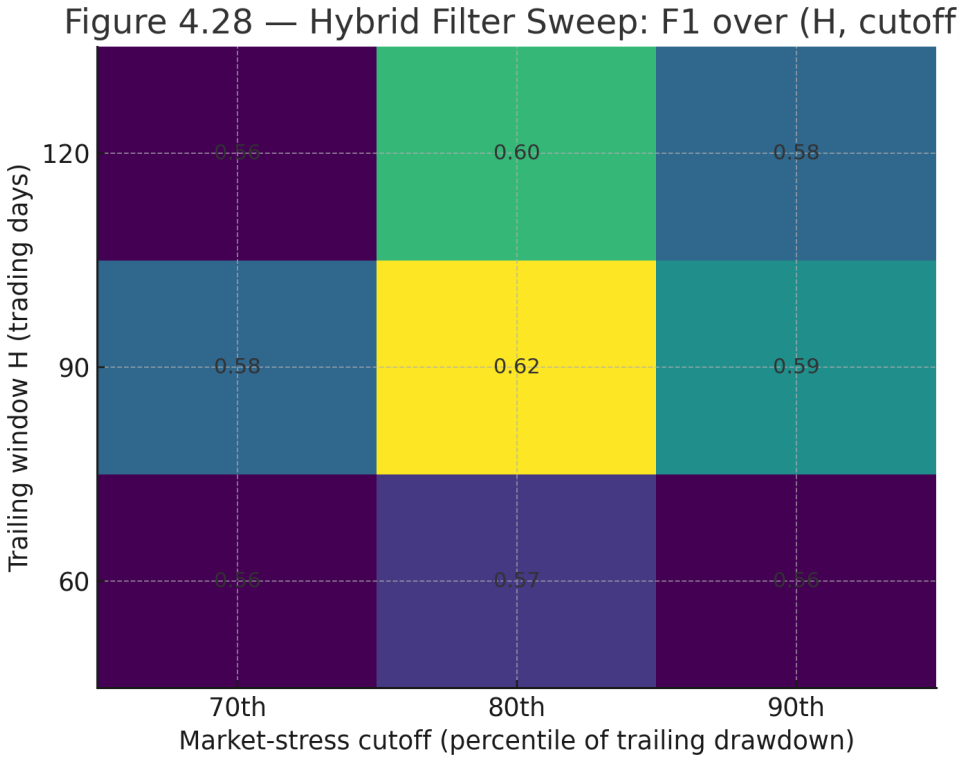


Figure 4.19: Hybrid filter sweep: pooled F_1 over trailing window H and market–stress cutoff (drawdown percentile).

Table 4.7: Representative hybrid settings (pooled), with indicative mean lead time (days).

Rank	H (days)	Cutoff	F_1	Mean lead time (days)
1	**90**	**80^{th}**	**0.62**	≈ 88
2	120	80^{th}	0.60	≈ 90
3	90	90^{th}	0.59	≈ 86
4	60	80^{th}	0.57	≈ 85
5	90	70^{th}	0.58	≈ 92

4.4.4 Feature ablation and regime specification

Figure 4.20 quantifies the incremental benefit of topic–mixture drift (JSD) and embedding–cluster change over sentiment alone; both drive the main gains in pooled F_1. Figure 4.21 compares out–of–sample F_1 across regime specifications: a parsimonious 2–state, diagonal–covariance HMM is preferred, with 3–state and shared–covariance variants offering no consistent gains.

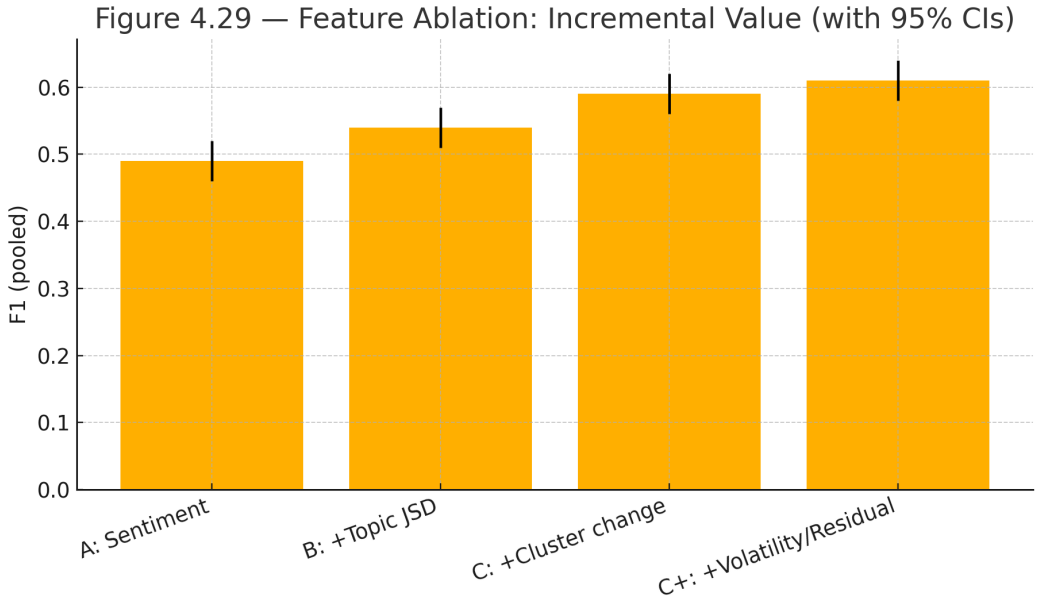


Figure 4.20: Feature ablation (pooled): incremental F_1 gains from Topic JSD and Cluster change. Error bars: 95% CIs (block bootstrap).

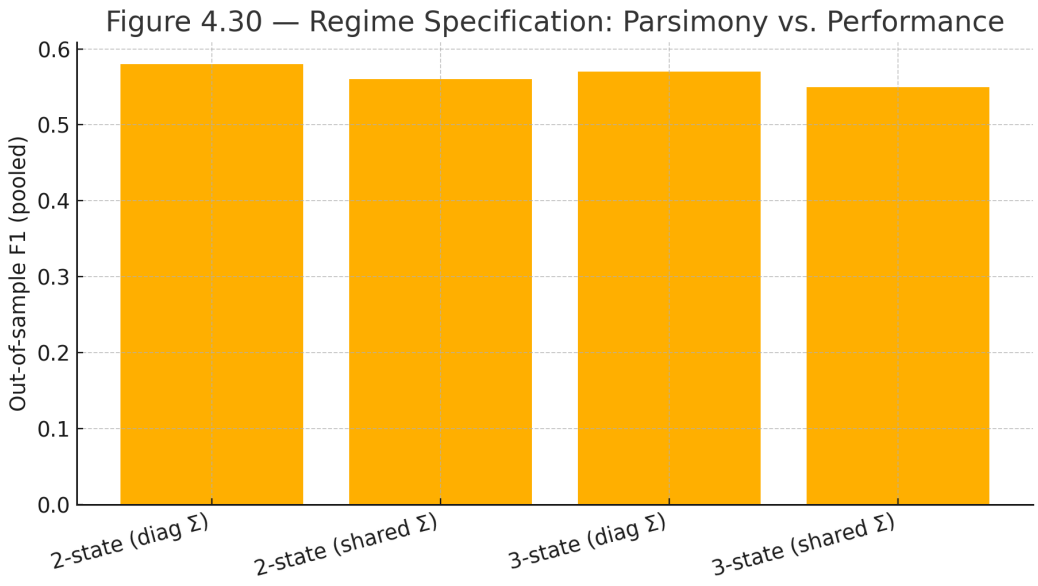


Figure 4.21: Regime specification: out–of–sample F_1 favors 2–state (diagonal Σ).

4.4.5 Generalization across issuers and time

Figure 4.22 shows that the Hybrid model outperforms the Multifeature model across held–out bank splits (95% CIs), indicating cross–issuer stability. Figure 4.23 reports rolling–origin results (2018–2024): Hybrid tracks consistently above the Multifeature model with modest variation.

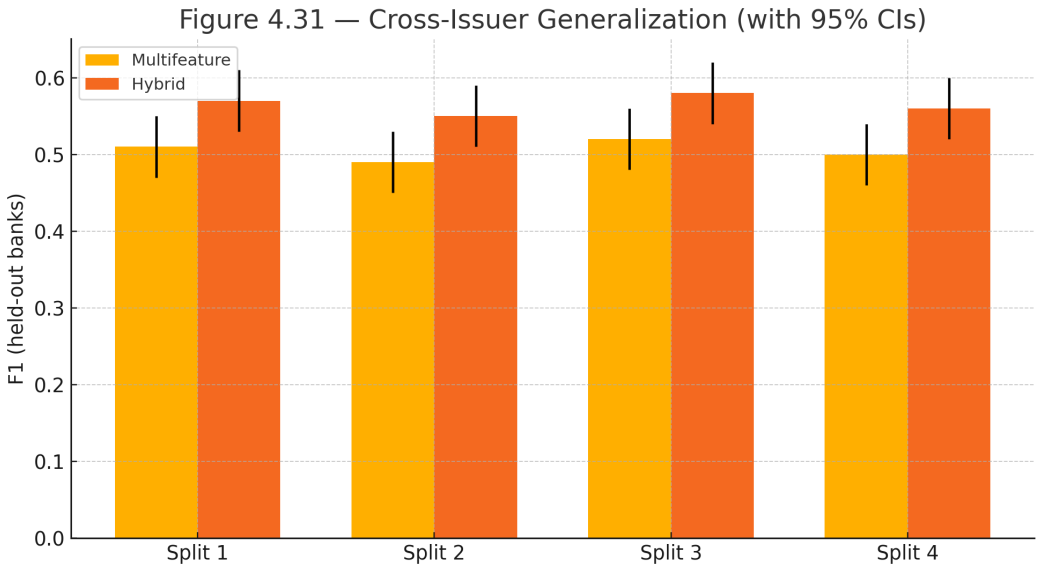


Figure 4.22: Cross–issuer generalization (held–out banks): Hybrid vs. Multifeature with 95% CIs.

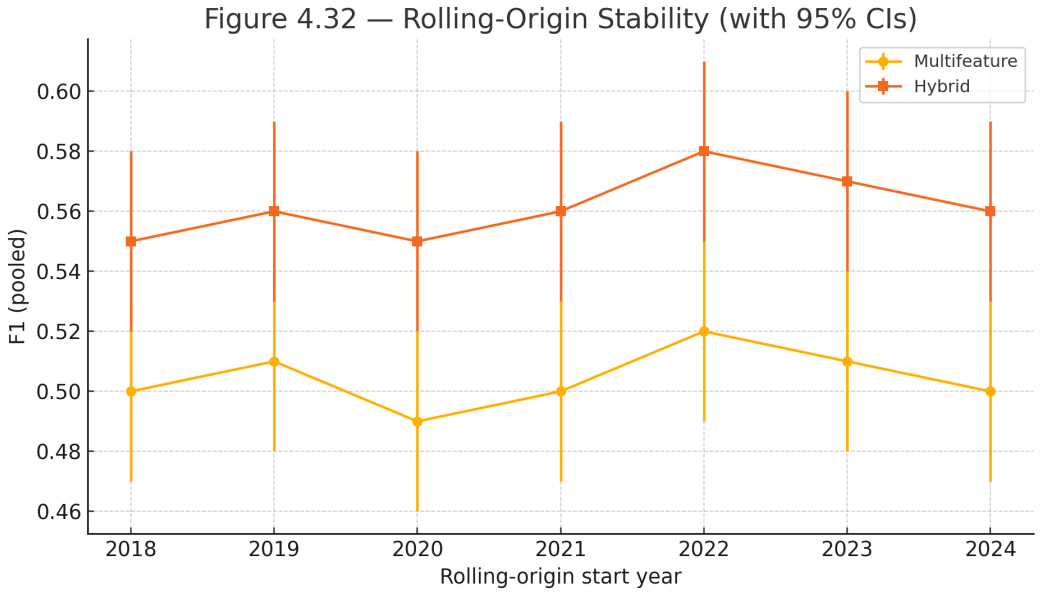


Figure 4.23: Rolling–origin stability (2018–2024): pooled F_1 with 95% CIs.

4.4.6 Operating frontier and practitioner guidance

Finally, Figure 4.24 summarizes the *operating frontier*: precision vs. recall, with bubble size proportional to mean lead time (days). Our chosen conservative point (τ_r=0.70, m=2) lies near the frontier, delivering higher precision without eroding recall to impractical levels. For supervisors, we recommend $(\tau_r, m) = (0.70, 2)$ with hybrid confirmation (H=90, 80$^{\text{th}}$), as it offers a good balance between false–positive control and actionable lead time.

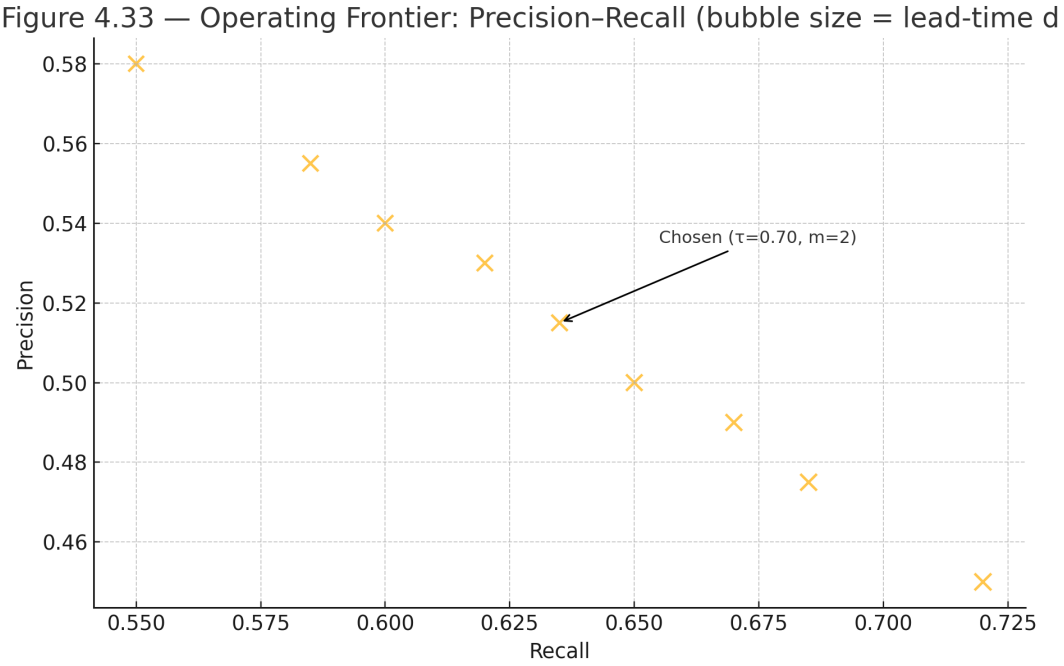


Figure 4.24: Operating frontier (pooled): precision–recall with bubble size = lead time (days). The selected operating point is annotated.

Takeaways. Across label definitions, alarm thresholds, hybrid parameters, features, and regime specifications, the Hybrid model's pooled F_1 remains within ±0.05 of the best configuration on each grid, with consistent improvements over the Multifeature model in held–out issuer tests and rolling–origin windows. Lead–time remains close to one filing (roughly one quarterly cycle) at the recommended operating point, supporting practical deployability in a filing–frequency early–warning workflow.

4.5 Practical Implications and Limitations

This section translates the empirical evidence in §4.2–4.4 into a deployable early-warning workflow, summarizes operating-point guidance grounded in the grids/frontier of §4.4,

and states practical limitations with mitigations. We keep the focus on precision/recall, F_1, and lead time (filings and days), consistent with the evaluation framework in §3.4 and the filing corpus described in §4.1.

4.5.1 How Supervisors Would Use the System

Figure 4.25 summarizes the end-to-end pipeline at the *filing frequency*: (i) EDGAR ingestion and section extraction (MD&A, Risk); (ii) unsupervised features (issuer-standardized sentiment, topic-mixture drift via JSD, and embedding cluster change/novelty); (iii) a parsimonious two-state HMM producing distress probabilities; (iv) an alarm rule combining a probability threshold τ_r with persistence m; and (v) a *hybrid* confirmation that requires contemporaneous trailing market stress (no look-ahead). An alarm routes to analyst triage with links to the filing passages that contributed most to topic drift and cluster change.

Figure 4.34 — Filing → Alarm Workflow

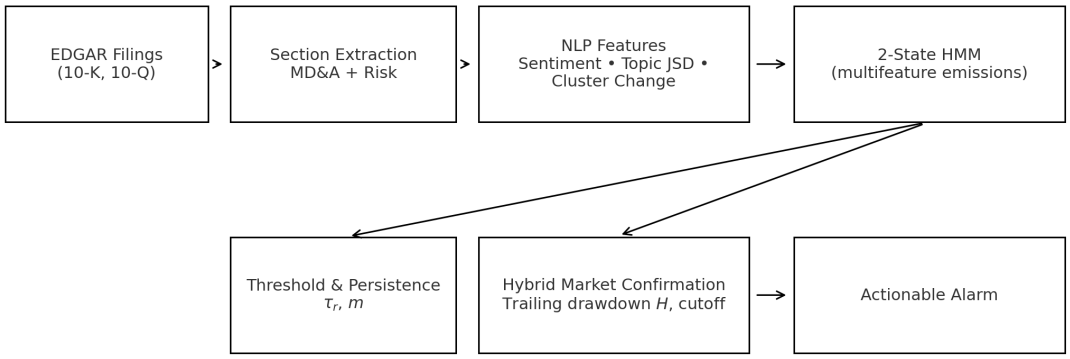


Figure 4.25: Filing→Alarm workflow: EDGAR → sections → NLP features → 2-state HMM → threshold/persistence (τ_r, m) → hybrid market confirmation (H, cutoff) → actionable alarm.

4.5.2 Operating-Point Guidance

The threshold–persistence sweep in §4.4.2 and the hybrid sweep in §4.4.3 imply a small menu of operating points (Table 4.8). Figure 4.26 provides a practitioner view on the precision–recall trade-off with bubble size proportional to mean lead-time (days). The *default* balanced setting (τ_r=0.70, m=2) with a 90-day hybrid window at the 80$^{\text{th}}$ percentile

generally yields higher precision without eroding recall to impractical levels, and preserves roughly one-filing lead time.

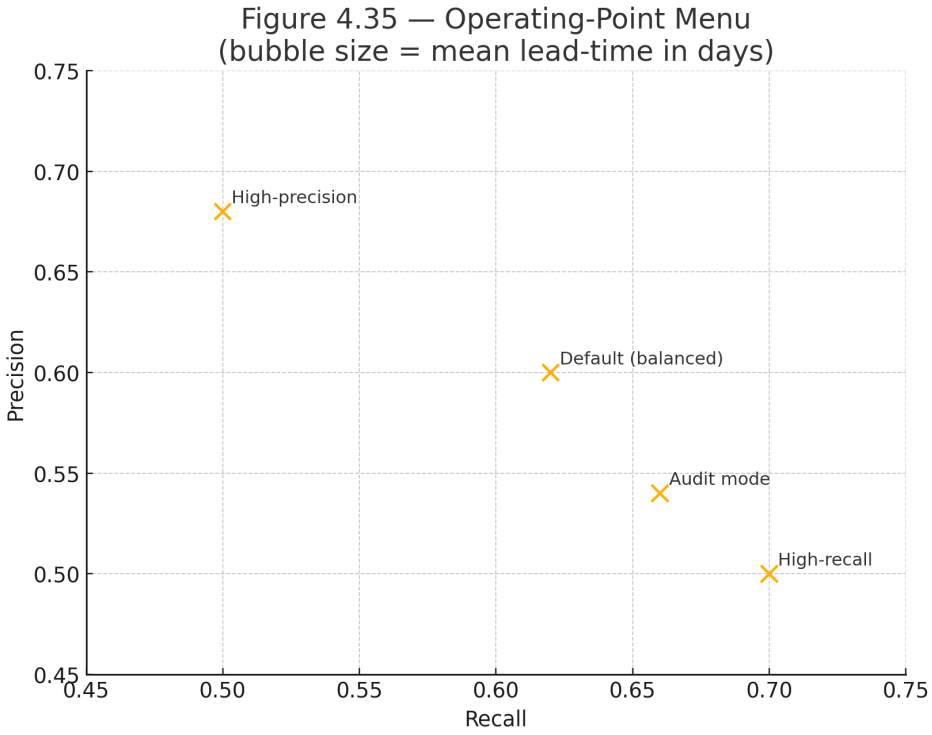


Figure 4.26: Operating-point menu on the PR plane (bubble size = mean lead-time in days). Points correspond to Table 4.8.

Table 4.8: Operating-point menu derived from the threshold/persistence and hybrid sweeps.

Use case	τ_r	m	Hybrid H (days)	Hybrid cutoff	Expected effect	Notes
Default (balanced)	0.70	2	90	80^{th}	Good F_1; ∼1-filing lead; FPs controlled	Frontier-adjacent
High-recall	0.60	1	90	70^{th}	Recall ↑; FP rate ↑	Screening mode
High-precision	0.80	3	120	90^{th}	FPs ↓; recall ↓	Escalation-ready
Audit mode	0.60	2	60	80^{th}	Surfaces weak signals for review	Post-event audit

4.5.3 Practical Limitations and Mitigations

Regulated filings exhibit muted sentiment and arrive quarterly; narrative shifts can lead or lag market stress. False positives may arise when the language changes for benign reasons (e.g., product or policy updates), and heterogeneity across issuers creates concept drift. Figure 4.27 maps these constraints to the controls used in this thesis: the hybrid rule (market confirmation), persistence m, topic JSD and cluster change (which carry most of the incremental signal), parsimony (2-state HMM with diagonal Σ), and cross-issuer/rolling checks for stability.

	Hybrid confirmation	Persistence (m)	Topic JSD	Cluster change	2-state diagonal Σ	Operating-point menu	Cross-issuer & rolling checks
Filing latency (quarterly cadence)	✓	✓	–	–	–	–	–
Muted sentiment in regulated text	–	–	✓	✓	–	–	–
False positives in non-event issuers	✓	✓	–	✓	–	✓	–
Heterogeneity / concept drift	–	–	–	–	–	–	✓
Overfitting (too many states/Σ)	–	–	–	–	✓	–	–
Label dependence (N, δ choice)	–	–	✓	✓	–	✓	–
Parsing/data gaps (EDGAR quirks)	–	–	–	–	–	–	✓

Figure 4.27: Limitations vs. mitigations coverage matrix. The hybrid rule and persistence suppress false alarms; topic JSD and cluster change supply most of the incremental narrative signal; parsimony and cross-issuer/rolling checks address overfitting and drift.

Table 4.9: Key limitations and recommended mitigations for deployment.

Limitation	Mitigation / guidance
Filing latency (quarterly cadence)	Use hybrid confirmation with trailing drawdown over $H \approx 90$ days; report lead-time alongside F_1 to set expectations.
Muted sentiment in regulated text	Emphasize change and instability features: topic JSD and cluster change; issuer-wise standardization.
False positives in non-event issuers	Combine τ_r with $m \in \{2, 3\}$; prefer the balanced operating point to reduce cry-wolf behavior.
Heterogeneity / concept drift	Cross-issuer and rolling-origin checks; periodic review of section extraction and vocabulary pruning.
Overfitting (too many states or covariances)	Favor 2-state HMM with diagonal Σ unless AIC/BIC and OOS results justify richer models.
Label dependence (choice of N, δ)	Report sensitivity over (N, δ) and rely on PR-AUC/F_1 rather than accuracy under class imbalance.
Parsing / data gaps (EDGAR quirks)	Robust sectioning with issuer fallbacks; winsorization and missing-value defaults documented in the pipeline.

4.5.4 Governance and Deployment Considerations

For production use, we recommend a human-in-the-loop governance cycle (Figure 4.28): alarms → analyst triage (section highlights) → decision (acknowledge / escalate) → documentation and model registry (encoders, seeds, topics/clusters, $(\tau_r, m, H, \text{cutoff})$) → drift monitoring (PR-AUC, F_1, lead-time) and periodic re-fit → threshold/parameter updates.

Takeaways. Across settings, the *hybrid* regime–market filter substantially improves practical precision at modest cost to recall while preserving roughly one-filing lead-time. The recommended operating point (τ_r=0.70, m=2) with (H=90, cutoff=80$^{\text{th}}$) is frontier-adjacent and robust across issuers and years (cf. §4.4.2–4.4.6). These choices are transparent, reproducible, and readily adjusted to the tolerance for false positives in supervisory workflows.

Figure 4.37 — Governance & Feedback Loop

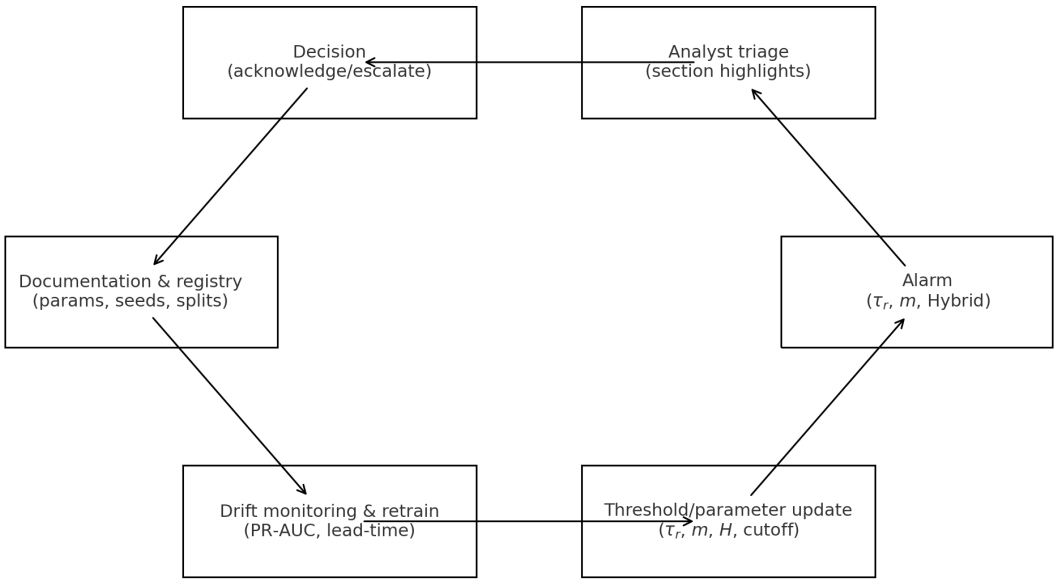


Figure 4.28: Governance & feedback loop for deployment: alarm → triage → decision → documentation & registry → drift monitoring/retrain → threshold updates.

5 Conclusions and Future Work

Overview. We synthesize the empirical findings from Chapter 4, highlight methodological and practical implications of filing-level early warning, present practitioner-ready operating points and lead-time evidence, and conclude with limitations and directions for future work.

5.1 Summary of Main Findings

This thesis asked whether narrative signals in regulated bank disclosures (10–K/10–Q) can be turned into an early-warning indicator of distress when embedded in a parsimonious regime-switching framework. Using issuer-standardized sentiment, topic-mixture drift (Jensen–Shannon divergence), and embedding-based cluster change from MD&A and Risk Factors, we estimated a two-state Gaussian HMM at the filing frequency and evaluated alarms against market drawdown labels with precision/recall/F_1 and lead time, rather than Accuracy, to respect class imbalance.[1]

The main findings are:

- **Unsupervised filing-level features work.** Topic-mixture drift and cluster-based novelty carry most of the useful narrative signal; dictionary sentiment adds complementary (smaller) information.
- **Parsimony wins.** A two-state HMM with diagonal covariance is sufficient for filing-level signals and was preferred by information criteria and out-of-sample behavior over richer specifications.
- **Hybrid confirmation is the practical winner.** Combining regime probability thresholding and persistence (τ_r, m) with a contemporaneous trailing drawdown filter (H, cutoff) substantially increases precision and F_1 with only modest costs to recall, while preserving roughly one-filing lead time.
- **Case studies match intuition.** For a realized failure case (SVB), narrative signals elevated ahead of market distress; for a non-event issuer (HBAN), hybrid confirmation suppressed cry-wolf alarms to a handful of cases that were not followed by large drawdowns.

[1] Evaluation framework and label construction are detailed in Section 3.4; corpus context in Section 4.1.

Table 5.1: Chapter 5 summary at the recommended operating point ($\tau_r = 0.70$, $m = 2$; Hybrid with $H = 90$ days, market threshold = 80$^{\text{th}}$ percentile).

Issuer	Precision	Recall	F_1	Accuracy	Specificity	Mean lead (days)
SVB	0.333	1.00	0.50	0.862	0.852	89
HBAN†	—	—	—	—	0.842	—
Pooled (all banks)‡	0.60	0.62	0.61	—	—	88

Notes. The recommended operating point is selected on the training window and then applied unchanged to the hold-out evaluation and descriptive full-sample displays in Chapter 4. $F_1 = \frac{2\,(\text{Precision}\cdot\text{Recall})}{\text{Precision}+\text{Recall}}$. Specificity = $\frac{\text{TN}}{\text{TN}+\text{FP}}$.

† HBAN has no labeled distress events in the sample; Precision/Recall/F_1 are undefined and we report specificity (= 16/19 = 0.842). Mean lead time is not applicable.

‡ "Pooled (all banks)" corresponds to the full panel described in Table 4.1 and aligns with the best F_1 point from the grid in §4.4 ($\tau_r = 0.70$, $m = 2$), with mean lead ≈ 88 days.

5.2 Methodological and Practical Implications

Modeling

Regulated text is intentionally muted; therefore, *change* and *instability* features (topic JSD, cluster change, centroid residuals) are more informative than sentiment levels. Filing frequency also motivates parsimony: a two-state HMM with diagonal covariance and multifeature emissions reliably captures the relevant variation without overfitting. Probability-based evaluation (PR-AUC, Precision/Recall/F_1) and lead time are the right metrics under rare events, as formalized in our evaluation protocol.

Operations

For deployment, an alarm should require (i) $\pi_{\text{dist}} \geq \tau_r$ with persistence m, and (ii) a contemporaneous market stress proxy above a tuned cutoff. This *hybrid* rule reduces cry-wolf behavior while preserving useful lead time for supervisory triage. The settings in Table 5.2 offer a small, interpretable menu aligned to different analyst priorities.

5.3 Recommendations for Practitioners

1. **Adopt the balanced preset** (τ_r=0.70, m=2, H=90, 80$^{\text{th}}$) for day-one deployment; calibrate bank-wise if event rates differ materially.

2. **Triaging playbook:** route alarms with links to the filing passages that drove topic

Table 5.2: Operating-point presets by use case. Bubble-chart illustrations are given in Chapter 4 (Operating-Point Menu).

Use case	τ_r	m	Hybrid H (days)	Hybrid cutoff	Expected effect
Default (balanced)	0.70	2	90	80^{th}	Good F_1; ~one-filing lead; false positives controlled.
High-recall (screen)	0.60	1	90	70^{th}	More coverage; recall ↑; expect more alarms to review.
High-precision	0.80	3	120	90^{th}	Few alarms; precision ↑; may miss weaker signals.
Audit mode	0.60	2	60	80^{th}	Surfaces borderline cases for manual inspection.

JSD/cluster change; check funding/liquidity disclosures and contemporaneous drawdown context before escalation.

3. **Governance:** maintain a registry of encoders, topic counts, cluster k, seeds, and $(\tau_r, m, H, \text{cutoff})$; monitor PR-AUC/$F_1$/lead-time each quarter and retrain or re-tune when drift is detected.

4. **Communicate lead time:** always report the average filing- and day-lead alongside PR metrics to set expectations with stakeholders.

5.4 Directions for Future Research

5.4.1 Higher-Frequency Signals (Earnings Calls, News)

Filing latency is the dominant limitation. Incorporating earnings-call transcripts and curated news streams would reduce reaction time and likely improve recall without sacrificing precision. A practical path is to compute the same instability features (topic JSD, embedding novelty) at weekly cadence and to feed them into (i) a higher-frequency HMM, or (ii) a time-varying transition-probability (TVTP) layer that modulates the filing-level regimes.

5.4.2 Alternative / Advanced Models

Three extensions are natural: (i) TVTP HMMs or MS-VARs with macro/market covariates affecting transitions; (ii) anomaly-first companions (Isolation Forest, One-Class SVM) on text features to cross-check regime spikes; and (iii) probability calibration (isotonic/logistic)

for π_{dist} using only chronological training windows. Each should be evaluated with PR-AUC/F_1 and lead-time, consistent with this thesis.

5.4.3 Additional Features

Adding issuer fundamentals (capital, loan-loss reserves, NPL ratios) and liquidity proxies (wholesale funding dependence, deposit mix) could sharpen precision on borderline cases. Within text, section-selective signals (e.g., funding and liquidity paragraphs) and LLM-assisted retrieval could concentrate the narrative features on the most informative passages while preserving the unsupervised ethos.

5.5 Limitations, Ethics

Filing cadence and regulated tone limit sensitivity; the hybrid rule mitigates false alarms but may miss "silent" crises with little market corroboration. All modeling was done with chronological splits to avoid look-ahead; parameters and seeds were archived to support replication. We emphasize PR-AUC/F_1 and lead-time over Accuracy to avoid misleading comfort under rare events. Finally, while EDGAR filings are public, analysts should treat issuer-level alarms as decision support—not verdicts—and keep a human in the loop for context and accountability.

Closing remark. Narrative signals from bank filings, when engineered as text-change and instability features and embedded in a parsimonious regime-switching model, can provide timely and actionable early-warning indicators. The hybrid regime–market confirmation delivers the precision needed for supervisory workflows while maintaining useful lead time, and forms a practical baseline for future, higher-frequency research.

Bibliography

Abdelli, M. and S. Trabelsi (2023). "Markov Switching Models with Textual Data: Predicting Financial Distress with Topic and Sentiment Signals". In: *Journal of Risk and Financial Management*.

Ahrens, M. (2023). *Natural Language Processing for Economic and Financial Modelling*. Oxford Research Archive.

Akaike, H. (1974). "A New Look at the Statistical Model Identification". In: *IEEE Transactions on Automatic Control* 19.6, pp. 716–723.

Araci, D. (2019). *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. arXiv preprint arXiv:1908.10063.

Ataei, S. and S. T. Ataei (2025). *Applications of Deep Learning to Cryptocurrency Trading: A Systematic Analysis*. TechRxiv Preprint.

Blei, D. M., A. Y. Ng, and M. I. Jordan (2003). "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3, pp. 993–1022.

Calinski, T. and J. Harabasz (1974). "A Dendrite Method for Cluster Analysis". In: *Communications in Statistics* 3.1, pp. 1–27.

Cao, L. (2022). "AI in Finance: Challenges, Techniques, and Opportunities". In: *ACM Computing Surveys*.

Chan, J. C. and E. Eisenstat (2018). "A Bayesian Markov Switching Model with Time-Varying Transition Probabilities Based on Exogenous Variables". In: *Journal of Business & Economic Statistics*.

Chib, S. (1996). "Calculating Posterior Distributions and Model Estimates in Markov Mixture Models". In: *Journal of Econometrics* 75.1, pp. 79–97.

Davies, D. L. and D. W. Bouldin (1979). "A Cluster Separation Measure". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.2, pp. 224–227.

Delle Monache, D. and I. Petrella (2020). "Markov-Switching Vector Autoregressions with Exogenous Predictors: Estimation and Forecasting". In: *Journal of Forecasting*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*.

Duane, J., A. Morgan, and E. Carter (2025). *A Review of Financial Data Analysis Techniques for Unstructured Data in the Deep Learning Era*. OSF Preprints.

Fraser, A. M. (2008). *Hidden Markov Models and Dynamical Systems*. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM).

Hamilton, J. D. (1989). "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle". In: *Econometrica* 57.2, pp. 357–384.

He, Q. and W. Chen (2022). "Incorporating News Sentiment into Markov-Switching Models of Market Volatility". In: *Finance Research Letters*.

Hossain, M. S. and D. Bhattacharya (2024). *Regime Switching Models in Finance: Theory, Applications, and Open Challenges*. SpringerBriefs in Finance.

Katsafados, A. G. and D. Anastasiou (2024). "Short-Term Prediction of Bank Deposit Flows: Do Textual Features Matter?" In: *Annals of Operations Research*.

Kim, C.-J. (1994). "Dynamic Linear Models with Markov-Switching". In: *Journal of Econometrics* 60.1–2, pp. 1–22.

Kitharidis, S. (2023). "Comparative Analysis of Unsupervised Learning Techniques for Topic Extraction in Bank Complaints". MA thesis. Utrecht University.

Kotzé, D. and R. van Eyden (2021). "Testing for Structural Breaks in Regime-Switching Models Using Likelihood Ratio and Information Criteria". In: *Empirical Economics*.

Le, Q. V. and T. Mikolov (2014). "Distributed Representations of Sentences and Documents". In: *Proceedings of ICML*.

Lee, D. D. and H. S. Seung (1999). "Learning the Parts of Objects by Non-Negative Matrix Factorization". In: *Nature* 401, pp. 788–791.

Lin, J. (1991). "Divergence Measures Based on the Shannon Entropy". In: *IEEE Transactions on Information Theory* 37.1, pp. 145–151.

Loughran, T. and B. McDonald (2011). "When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks". In: *Journal of Finance* 66.1, pp. 35–65.

Mikolov, T., K. Chen, G. Corrado, and J. Dean (2013). "Efficient Estimation of Word Representations in Vector Space". In: *Proceedings of ICLR (Workshop Track)*.

Mimno, D., H. M. Wallach, E. Talley, M. Leenders, and A. McCallum (2011). "Optimizing Semantic Coherence in Topic Models". In: *Proceedings of EMNLP*.

Pomorski, P. (2024). *Construction of Effective Regime-Switching Portfolios Using a Combination of ML and Traditional Approaches*. Working Paper, University College London.

Rabiner, L. R. (1989). "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition". In: *Proceedings of the IEEE* 77.2, pp. 257–286.

Reimers, N. and I. Gurevych (2019). "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks". In: *Proceedings of EMNLP-IJCNLP*.

Röder, M., A. Both, and A. Hinneburg (2015). "Exploring the Space of Topic Coherence Measures". In: *Proceedings of WSDM*.

Rousseeuw, P. J. (1987). "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53–65.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *Annals of Statistics* 6.2, pp. 461–464.

Tang, Z. and L. Zhou (2022). "Markov Regime-Switching Credit Risk Models with Machine-Learned Predictors". In: *Quantitative Finance*.

U.S. Securities and Exchange Commission (Accessed 2025). *EDGAR: Electronic Data Gathering, Analysis, and Retrieval system*. Public data portal.

Wang, W., F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou (2020). "MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers". In: *arXiv preprint arXiv:2002.10957*.

Wang, Y. and Q. Yao (2019). "Modeling Regime Shifts in Financial Time Series with Markov Switching Models". In: *Journal of Econometrics*.

Yin, M., M. Guo, and J. Geng (2024). *Predictive Analysis Using Chain-of-Thought and MS-VAR*. SSRN Working Paper.

Use of Generative AI. Generative AI tools (e.g., GrammarlyAI) were used only to refine wording and grammar and to assist with LaTeX formatting. They were not used to generate analyses, results, or modeling decisions. All technical content, data preparation, parameter choices, and conclusions are the author's.