Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The Language of Letting Go: Text-Label Co-Production and the Limits of Computational Inference in CEO Departure Data

Murat Günsel
Erkut Altındağ
Doğuş University
Ayşe Günsel
ayse.gunsel@kocaeli.edu.tr

Kocaeli University, Türkiye

Research Article

Keywords: CEO succession, CEO dismissal, computational text analysis, data leakage, construct validity, corporate governance

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11105563/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

The Language of Letting Go: Text-Label Co-Production and the Limits of Computational Inference in CEO Departure Data

Blinded manuscript

Abstract

Can textual descriptions of chief executive departures recover whether an exit was forced, and what does such recoverability establish when text and label share a production process? We conduct a leakage-aware computational audit of an open database of 5,516 coded CEO departures in S&P 1500 firms from 2000 through 2018, whose narrative field consists of coder-constructed summaries rather than firm announcements. Among 5,365 discretionary departures with usable text, forced cases are substantially more negative than other exits (Cohen's d = -.89), and dominant-topic membership is associated with forced status (Cramer's V = .41 in the four-topic model, .56 in a seven-topic sensitivity model). A row-random benchmark produces ROC-AUC = .891. Keeping each firm's observations in a single validation fold and learning all text transformations from training data lowers the figure to .860 (95% cluster-bootstrap CI [.849, .872]); it remains .838 after direct label and coder-meta cues are masked and .824 under a strict topic-only specification. A model trained on 2000-2013 cases reaches .886 in 2014-2018. High-dimensional TF-IDF improves ranking (.937) but yields compressed, poorly calibrated probabilities. These results establish robust within-database recoverability, not evidence that firms' impression management leaks the truth. We develop text-label co-production as a measurable extension of familiar criterion-contamination and annotation-artifact problems, and propose provenance validity as the standard that connects a text's origin to the inferences it can support. The caution applies wherever coders both label events and write the narratives that describe them; testing firm communication directly requires a source-separated design.

Keywords: *CEO succession; CEO dismissal; computational text analysis; data leakage; construct validity; corporate governance*

1. Introduction

When a chief executive leaves a public corporation, the event produces more than a leadership vacancy. It also produces competing accounts. The firm may frame the exit as an orderly transition, journalists may question that account, regulators may disclose contractual or governance details, and researchers may later synthesize the evidence to assign a departure code. These texts concern the same event, but they are not interchangeable. A firm announcement is organizational communication; news coverage is an external interpretation; and a coder's narrative is a measurement artifact. Treating the third as if it were the first can turn impressive prediction into an invalid theoretical claim.

This distinction is especially consequential for forced CEO turnover. Agency and governance research treats dismissal as the board's ultimate monitoring sanction, yet firms rarely describe a CEO exit in those terms (Coughlan and Schmidt 1985; Eisenhardt 1989; Jenter and Kanaan 2015; Weisbach 1988). Impression-management research, in turn, explains why boards and executives may soften blame, preserve legitimacy, or protect reputations around contested succession events (Elsbach 1994; Graffin et al. 2011; Wiesenfeld et al. 2008). These literatures invite a memorable question: even when organizations manage the story, does the true departure type leak into the language?

The available data cannot answer that question as originally posed. The open CEO Turnover and Dismissal Database developed by Gentry, Harrison, Quigley, and Boivie (2021) contains departure codes, narrative descriptions, and links to public sources. Its transparency is a major strength. Yet the narrative descriptions were written by the same coding team that reviewed the evidence and assigned the labels. Some notes state the classification decision directly; others summarize performance, misconduct, succession, or retirement evidence used to make that decision. The predictor text and the outcome code therefore share a production process.

We call this condition text-label co-production: the predictor text is written, selected, or edited during the same adjudication process that generates the target label. Co-production is more fundamental than the presence of words such as fired or retired. Direct terms create lexical leakage and can be masked. Repeated firms create entity dependence and can be handled through grouped validation. Vocabulary change creates temporal transport risk and can be tested with later-period holdouts. But none of these procedures makes a coder-written narrative independent of the code it was created to justify. The core issue is provenance validity: whether the origin and production of the text support the theoretical inference attached to a predictive result.

The present study turns that limitation into an explicit research design. We ask how reliably coder-constructed departure narratives recover forced versus other discretionary exits after safeguards against row-level dependence, direct target cues, temporal drift, exact duplicate texts, class imbalance, and probability miscalibration. We compare a row-random benchmark with firm-grouped five-fold validation, direct and strict masks, topic-only and sentiment-only models, a high-dimensional TF-IDF sensitivity model, later-

period holdouts, and source-mix strata. All text transformations are re-estimated inside the training data. We report discrimination, precision-recall performance, threshold metrics, Brier scores, calibration intercepts and slopes, and expected calibration error.

The study makes three contributions. First, it provides a transparent audit of a widely useful CEO-dismissal resource. The narratives contain a strong, transportable signature of the assigned departure code even after observable leakage controls. Second, it separates predictive recoverability from construct validity. This distinction is familiar in broad methodological discussions of prediction and explanation (Yarkoni and Westfall 2017). Our narrower move is to specify how a shared adjudication process can produce a model that generalizes across rows, firms, and years while remaining ill suited to the intended organizational inference. Third, the study offers a provenance-aware workflow for computational management research. Recent reviews emphasize that text measures require explicit validation rather than reliance on predictive accuracy alone (Anglin 2024; Campion and Campion 2025; Park and Montgomery 2025). Here validation begins before modeling: who authored the text, when was it produced, and how did that process relate to the label? Although our setting is corporate governance, the audit logic travels well beyond it: wherever the researchers who assign an event code also write the note that describes the event - protest catalogs, conflict databases, coded case files - the same circularity can masquerade as predictive success.

Accordingly, the paper is a measurement audit and a test of within-system classification. A direct test of organizational framing would require source-separated texts and labels frozen independently of the narratives supplied to the model.

2. Theoretical Background and Research Questions

2.1 CEO departure as a governance event and communicated account

Chief executive succession is simultaneously a governance outcome and a socially interpreted event. Agency theory locates dismissal in the board's monitoring role: when managerial performance or conduct falls below an acceptable threshold, replacing the chief executive is the board's most consequential corrective action (Eisenhardt 1989; Jensen and Meckling 1976). Empirical studies have therefore examined performance-turnover sensitivity, board composition, market conditions, secular turnover patterns, and the

difficulty of identifying forced exits when firms avoid explicit dismissal language (Coughlan and Schmidt 1985; Jenter and Kanaan 2015; Kaplan and Minton 2012; Weisbach 1988).

The same event creates a legitimacy problem. A contested exit can stigmatize the departing executive, raise questions about the board's prior selection and oversight, disrupt stakeholder expectations, and expose strategic disagreement. Organizations may respond with accounts that normalize the transition, emphasize gratitude and continuity, or redirect attention toward future leadership (Elsbach 1994; Sutton and Callahan 1987). Media actors can reproduce, reinterpret, or contest that framing (Bednar 2012; Pollock and Rindova 2003). Graffin et al. (2011), for example, show that firms engage in anticipatory impression management around CEO succession.

These perspectives make language theoretically relevant, but they also impose an authorship requirement. Evidence about a firm's impression management must come from firm-authored text or from a design that separates the firm's account from external and researcher-authored accounts. A blended coder summary may capture public evidence surrounding the event, but it cannot identify which actor supplied a diagnostic phrase or whether the firm's framing persuaded anyone.

2.2 Three textual layers and the provenance of inference

Figure 1 distinguishes three textual layers. At the source layer, firms, regulators, and media organizations produce documents for different purposes. At the adjudication layer, database coders review those materials and form a judgment about the event. At the analytic layer, the coder's narrative note and departure code become a predictor-outcome pair. The original departure event influences both paths, but the adjudication process also directly shapes both variables. This shared path creates construct circularity.

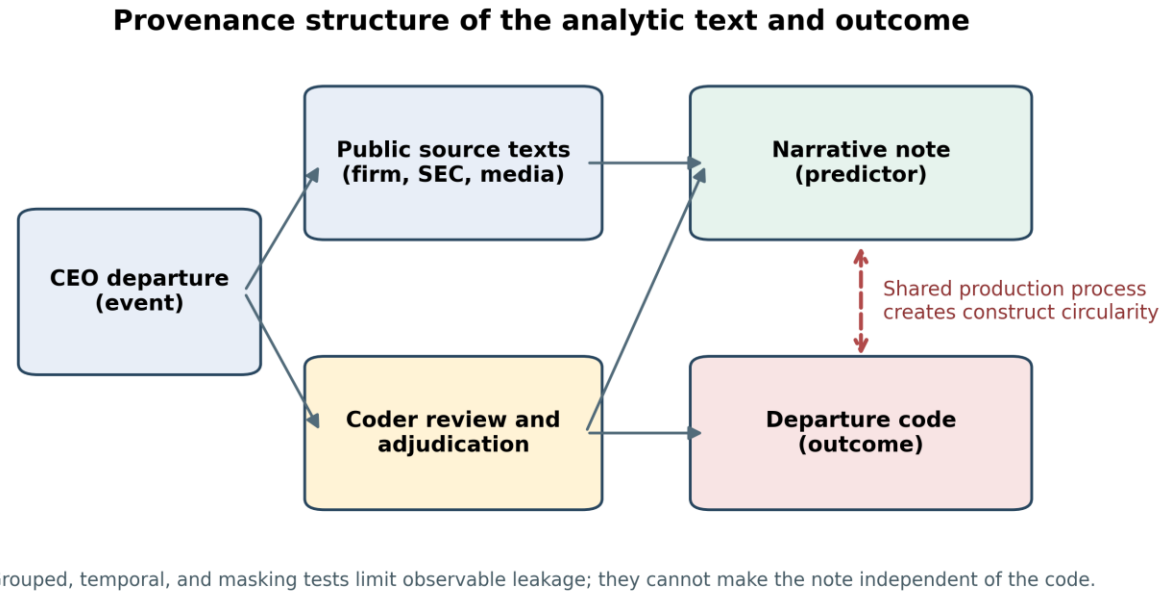
A closely related problem is familiar in industrial and organizational psychology. Criterion contamination occurs when an operational criterion contains variance that lies outside the intended construct; criterion deficiency arises when relevant parts of that construct are missing. Both are plausible here. A coder's explanation of why an exit received a forced-departure label can contaminate the narrative predictor with the adjudication rule itself. At the same time, a blended summary is deficient as a measure of organizational framing because it omits a clean, attributable record of the firm's own account and says little about audience response.

Natural-language processing research reaches the same terrain from another direction. Annotation artifacts and dataset artifacts are regularities introduced by labeling instructions, annotator habits, source selection, or benchmark construction. A classifier may exploit those regularities and perform well without learning the substantive phenomenon the dataset is often taken to represent. In management research the inferential cost is unusually easy to miss: text-derived outputs are routinely interpreted as properties of an organization or executive (Choudhury et al. 2021; Grimmer and Stewart 2013; Hannigan et al. 2019; Kobayashi et al. 2018). A topic may capture how coders document a dismissal. Sentiment may belong partly to a journalist or rater. High predictive accuracy does not settle authorship.

Text-label co-production does not make either family of problems newly discovered. Its more modest contribution is to join them in a form tailored to management text analysis and open to empirical audit. It names the condition in which predictor text and target label are written, selected, or revised within one adjudication process; it then asks how much apparent signal survives checks for visible cues, repeated entities, and time. That formulation distinguishes an ordinary measurement problem from a claim about organizational communication.

Classic target leakage can coexist with co-production. Leakage occurs when information unavailable at legitimate prediction time enters the features and inflates performance (Kaufman et al. 2012). Co-production concerns a prior issue: the origin of both variables. Even after explicit target words are removed and performance is evaluated on unseen firms, syntax, evidence selection, and narrative emphasis may carry the coder's completed judgment. Masking can quantify visible leakage; it cannot recreate an independently authored text.

Figure 1. Provenance structure of the analytic text and outcome



Note. Grouped, temporal, and masking tests address observable leakage and transport. The shared adjudication path remains a property of the dataset's construction.

2.3 Recoverability, validation, and the boundary of the claim

Recoverability asks whether the text contains stable information about a label. Construct validity asks whether that relationship represents the theoretical phenomenon claimed. The two can diverge. A model may show excellent out-of-sample ranking because the same coding protocol is applied consistently across firms and years. That consistency is useful for database audit and assisted recoding, yet it is not evidence about the efficacy of corporate impression management.

A credible recoverability analysis nonetheless requires demanding validation. Repeated observations from the same firm must not be split across training and test data; otherwise recurring company names, succession routines, or reporting practices can inflate performance. Text preprocessing and topic estimation must be learned inside the training fold; otherwise the test corpus influences the representation. Because forced departures are the minority class, ROC-AUC must be accompanied by precision-recall performance and threshold metrics (Saito and Rehmsmeier 2015). Probability quality must be evaluated separately through Brier score and calibration statistics (Van Calster et al. 2019). Finally, later-period holdouts must test whether the model survives changes in vocabulary and source practices.

Together, these safeguards support a bounded inference: information about the assigned code is present in the notes and transports within the same data-generating system. Its authorial source remains unresolved. The analysis is therefore organized around four research questions rather than causal hypotheses.

Research Question 1. Does the share of discretionary CEO departures classified as forced change systematically between 2000 and 2018?

Research Question 2. What stable and interpretable thematic structure emerges from the coder-constructed departure narratives?

Research Question 3. How strongly do narrative sentiment and dominant topic differ across forced and other discretionary departure classifications?

Research Question 4. How much out-of-sample recoverability remains across firms, later periods, and progressively stricter leakage controls, and how well calibrated are the resulting probabilities?

3. Data and Methods

3.1 Data source, provenance, and analytic samples

The data come from the open CEO Turnover and Dismissal Database associated with Gentry et al. (2021). The raw file contains 9,423 departure records spanning years before and after the database's intended coverage window. We restrict descriptive analyses to 5,516 validly coded departures during fiscal years 2000-2018, representing 2,749 firms. Departure codes distinguish death, illness, dismissal for performance, dismissal for legal/ethical/policy reasons, retirement, a new opportunity, and a residual other category.

For the forced-versus-other analysis, death and illness are excluded because they are not discretionary governance outcomes. This yields 5,381 discretionary departures, including 1,205 forced cases (22.4%). Sixteen records lack usable narrative sentiment, leaving 5,365 observations at 2,725 firms in the common textual sample; 1,204 are forced. The narrative field was constructed by raters after reviewing news coverage, SEC filings, press materials, and other web sources. It is neither a verbatim firm announcement nor a separable media article. Gentry et al. report supervised coding, escalation of ambiguous cases, and an independent recoding check that matched 87% of 200 reviewed observations. The 87% figure is raw agreement, not a chance-corrected reliability coefficient. Because the category-by-category cross-tabulation for the recoded cases is unavailable, neither Cohen's kappa nor Gwet's AC1 can be reconstructed from the

released files. We therefore treat the figure as limited evidence of label consistency, not a complete reliability assessment.

A data-integrity audit identified 379 reused dataset identifiers covering 759 rows. Merging the raw and analysis-ready files on that identifier alone would therefore create duplicate records. We instead verified identical row order and exact agreement on firm and year keys before aligning source-text fields to the analysis-ready records. Exact normalized narrative duplicates were rare: 14 rows (0.26%) formed seven duplicate pairs, and only one pair (two rows) crossed firm boundaries. The principal grouped results are therefore not explained by widespread verbatim duplication.

3.2 Outcome, sentiment, and textual representation

Forced departure equals one for performance-related and legal/ethical/policy dismissals and zero for other discretionary exits. Sentiment is the supplied VADER compound score, ranging from -1 to +1 (Hutto and Gilbert 2014). VADER is not a finance-specific instrument, an important qualification because generic dictionaries can misread domain-specific language (Loughran and McDonald 2011). Sentiment is therefore treated as one feature rather than a definitive measure of organizational tone. Topic-only specifications test whether the main recoverability pattern depends on this generic lexicon.

For topic modeling, narratives are lowercased; punctuation and numerals are removed; and standard English stop words are combined with a prespecified set of domain-generic terms (CEO, CEOs, company, companies, executive, executives, firm, and firms). We retain unigrams appearing in at least 20 documents and no more than 40% of documents. Latent Dirichlet allocation (LDA) uses batch learning (Blei et al. 2003). Full topic distributions enter classifiers, whereas dominant topics are used only for descriptive tables and figures. The topic shares form a composition: within each document the four shares sum to one. That simplex constraint creates exact linear dependence if the shares are treated as ordinary unconstrained covariates, a point addressed in the validation specification below.

3.3 Topic-number selection and interpretive audit

We compare k = 4 through 10 across three fixed random starts. For each k, we calculate UMass coherence (less negative is better), topic stability based on cosine similarity after optimal topic matching, and

top-ten-word diversity. Standardized values are averaged into a descriptive composite and interpreted together with parsimony and substantive clarity. The four-topic solution has the highest composite score. Topic labels are assigned only after inspecting top words and the five most topic-representative narratives. Representative texts are used for interpretive audit, not for an artificial human-coder reliability statistic.

To prevent test information from entering the representation, topic selection describes the corpus, but predictive validation refits the count vectorizer and LDA model inside every training fold. Validation cases are transformed only after the training representation is fixed.

3.4 Validation design, masks, and sensitivity models

The principal classifier is L2-regularized logistic regression using four topic shares and VADER sentiment. The preferred evaluation is stratified five-fold cross-validation grouped by gvkey, so all observations from a firm remain in one fold. A stratified row-random seven-topic analysis is retained only as a benchmark against the initial specification. Predictors are standardized within each training fold.

All four topic shares are retained. We did not drop a reference topic or apply an additive or isometric log-ratio transformation. Because the centered shares are linearly dependent, individual topic coefficients would not support ordinary unconstrained interpretation; none are reported. The L2 penalty yields a stable predictive solution, and performance is judged out of sample, but regularization is not a substitute for a compositional model. This choice is acceptable for the present ranking audit and remains a limitation for any future coefficient-level inference.

Models are fitted to the observed class distribution. No inverse-frequency class weights, oversampling, undersampling, or synthetic resampling are used in either the topic-based or TF-IDF specifications. Retaining the natural 22.4% forced-departure prevalence makes Brier scores and calibration estimates interpretable for this corpus. It also means that a conventional .50 threshold favors the majority class, so precision, recall, and F1 are reported together rather than treating accuracy as sufficient.

The direct mask removes explicit forced and voluntary labels and action phrases, including variants of fired, dismissed, ousted, terminated, forced, involuntary, removed, asked to resign, retired, resigned, stepped down, and new opportunity. It also removes coder-meta language, including variants of code, recode, classification, changed code, departure code, and numbered code references. The strict mask additionally

removes rubric-like terms related to performance, missed targets, misconduct, fraud, investigation, bankruptcy, illness, death, interim succession, replacement, appointment, merger, acquisition, transactions, effective immediately, abruptness, and conflict. The strict topic models exclude sentiment so that unmasked lexical polarity cannot re-enter through VADER.

Sensitivity models include (a) four-topic topic-only, (b) sentiment-only, (c) strict seven-topic topic-only, (d) TF-IDF logistic regression with unigrams and bigrams, (e) directly masked TF-IDF, and (f) a deliberately simple model containing forced-cue count, voluntary-cue count, coder-meta count, and narrative length. The TF-IDF model tests whether high-dimensional lexical features add ranking information; it is not treated as a more interpretable theoretical model. We did not estimate an embedding- or transformer-based classifier. TF-IDF should therefore be read as a strong lexical benchmark, not as the attainable upper bound on recoverability.

3.5 Temporal transport, source mix, and performance criteria

Temporal models train on 2000-2013 departures and test on 2014-2018 cases. We report performance for all later observations and separately for later firms that never appear in the training period. Source links are flagged for the presence of SEC materials, major-news domains, or press-wire domains. Because each narrative blends evidence and coder synthesis, these strata test transportability across source environments; they do not separate firm language from media language.

We report ROC-AUC, area under the precision-recall curve (PR-AUC), Brier score, and precision, recall, and F1 at a .50 threshold. Ninety-five percent intervals for grouped ROC-AUC and PR-AUC use 1,000 bootstrap samples at the firm level. Calibration is summarized by an intercept and slope from a logistic recalibration model and by ten-bin expected calibration error (ECE). Under the standard definition, an ideal intercept is zero and an ideal slope is one. Slopes greater than one indicate predictions that are too compressed; slopes below one indicate probabilities that are too extreme (Van Calster et al. 2019). All random seeds, masks, and preprocessing rules are fixed in the reproducible analysis script.

The grouped, masked, feature-restricted, temporal, and source-mix specifications generate a sizeable family of performance estimates and ancillary test statistics. We did not apply family-wise error-rate control or another multiplicity adjustment across that family. These models were designed as related sensitivity analyses

around the principal firm-grouped result, rather than as independent confirmatory hypothesis tests, but their intervals and p values remain unadjusted for multiple comparison.

4. Results

4.1 Composition and temporal pattern

Table 1 summarizes the 5,516 coded departures. Retirement is the largest category (2,558 cases), followed by the residual other category (1,475), performance dismissal (1,037), misconduct dismissal (168), a new opportunity (143), illness (81), and death (54). Among the 5,381 discretionary departures, forced exits constitute 22.4%.

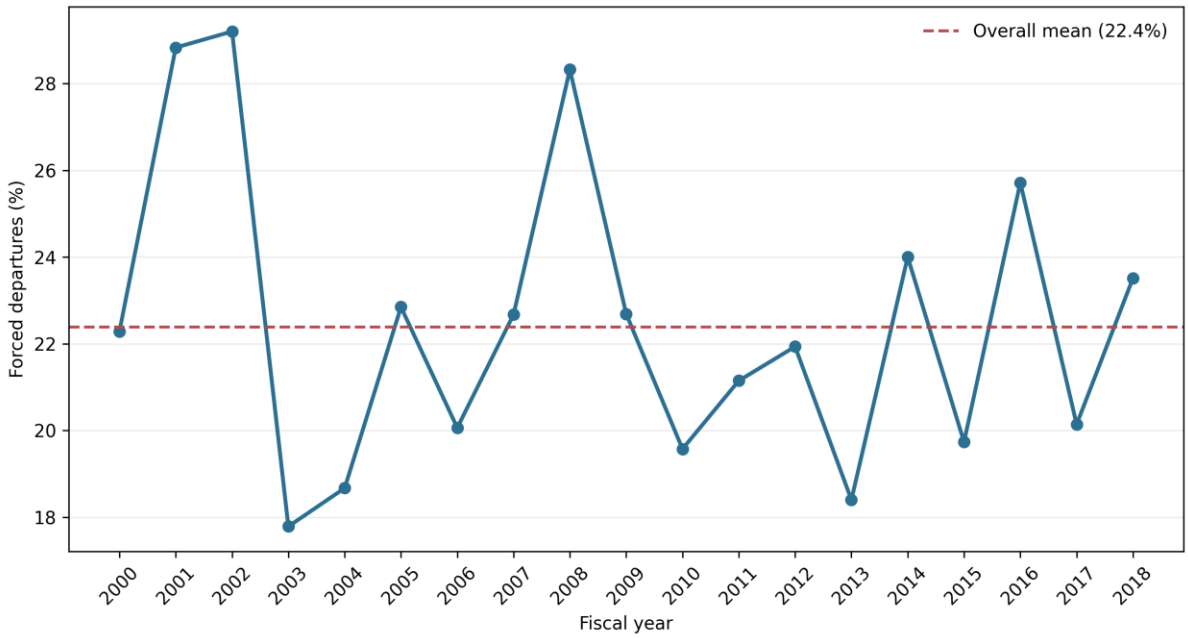
Table 1. Composition of coded CEO departures, 2000-2018

Departure type	n	Share of all coded (%)
Retired	2,558	46.4
Other or unclassified	1,475	26.7
Dismissed (performance)	1,037	18.8
Dismissed (misconduct)	168	3.0
New opportunity	143	2.6
Illness	81	1.5
Death	54	1.0

Note. N = 5,516 coded departures at 2,749 firms. Forced turnover comprises performance and misconduct dismissals and represents 22.4% of 5,381 discretionary departures.

The annual forced share fluctuates around the overall mean without a secular increase. In the firm-clustered logistic trend model, the yearly odds ratio is 0.995, 95% CI [0.983, 1.007], z = -0.83, p = 0.406. Figure 2 displays unsmoothed annual shares. Research Question 1 is therefore answered by a stable, substantial forced-turnover baseline rather than an increasing trend.

Figure 2. Annual forced share among discretionary CEO departures



Note. The dashed line marks the 22.4% overall mean.

4.2 Topic selection and narrative structure

Table 2 reports the topic-selection diagnostics. The four-topic solution has the strongest composite across coherence, matched-topic stability, and top-word diversity. It is also the most parsimonious interpretable representation. Larger solutions remain useful as sensitivity models, particularly because k = 7 produces a finer separation of contested departures.

Table 2. Topic-model selection diagnostics

k	UMass coherence	Stability	Top-10 diversity	Composite z
4	-1.677	0.770	0.825	1.484
5	-1.659	0.737	0.740	0.342
6	-1.811	0.737	0.833	0.609
7	-1.796	0.740	0.700	-0.293
8	-1.948	0.735	0.762	-0.335
9	-1.906	0.709	0.756	-0.720
10	-1.987	0.707	0.740	-1.087

Note. Coherence is better when less negative. Stability is the mean cosine similarity of optimally matched topics across three random starts. The composite averages standardized coherence, stability, and diversity.

Table 3. Four-topic interpretation and leading words

Performance-related and contested exits	Formal succession and executive biographies	Transactions	Employment, separation, and residual transitions
year	chief	merger	mr
board	officer	billion	resigned
years	board	announced	bank
chairman	chairman	acquisition	financial
mr	president	stock	corp
time	mr	million	agreement
stepped	directors	share	said
retirement	announced	corporation	president

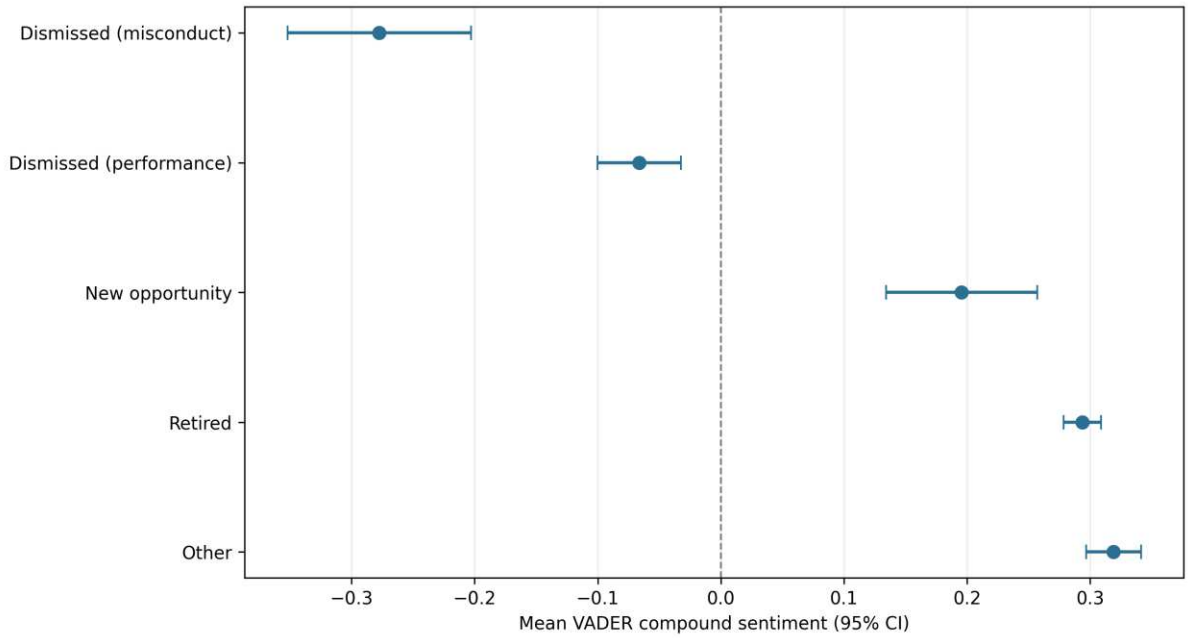
Note. Labels reflect top words and representative narratives. They describe themes in coder-constructed notes, not themes in firm-authored announcements.

The four themes are performance-related and contested exits; formal succession and executive biographies; corporate transactions; and employment, separation, and residual transitions. The first theme combines retirement language with performance evidence and coder rationale. This combination is theoretically revealing but must be interpreted at the adjudication layer: it shows how contested departures are documented, not that firms themselves used a particular framing. Research Question 2 is answered by a parsimonious structure that also makes the provenance problem visible.

4.3 Sentiment and topic association

Narrative sentiment differs sharply by classification. Forced cases average -0.096, compared with 0.299 for other discretionary exits, a mean difference of -0.395 and Cohen's d = -0.89. The difference is supported by Welch's test, $t = -23.03$, $p < .001$, and a Mann-Whitney test, $p < .001$. Misconduct dismissals are most negative ($M = -.278$), followed by performance dismissals ($M = -.066$), whereas new-opportunity ($M = .196$), retirement ($M = .294$), and other departures ($M = .319$) are positive.

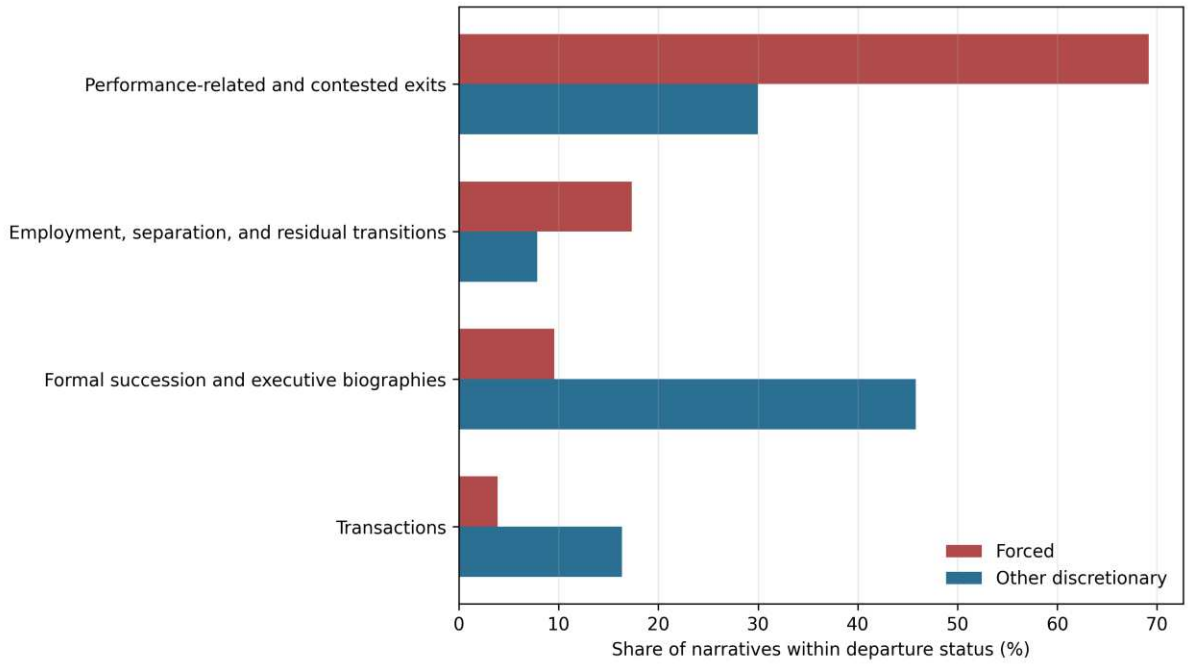
Figure 3. Mean VADER sentiment by discretionary departure type



Note. Points are means; bars are 95% confidence intervals.

Dominant four-topic membership is associated with forced status, chi-square(3) = 886.61, $p < .001$, Cramer's V = 0.407. The performance-related and contested-exit topic dominates 69.2% of forced notes and 30.0% of other discretionary notes. Formal succession and executive biographies account for 45.8% of other notes but only 9.6% of forced notes. The finer seven-topic model strengthens the association, chi-square(6) = 1695.68, $p < .001$, V = 0.562. These omnibus chi-square tests and the accompanying Cramer's V estimates treat departure events as independent and are not adjusted for within-firm clustering, even though some firms contribute more than one event. A firm-clustered sensitivity check could repeatedly retain one departure per firm or recast the association in a model with cluster-robust variance, paralleling the firm-level bootstrap used for the principal AUC analysis. Research Question 3 is therefore answered by large tone differences and a substantial - though specification-sensitive - thematic association, with the topic association interpreted subject to that dependence limitation.

Figure 4. Dominant four-topic prevalence by departure status



4.4 Leakage audit and out-of-sample recoverability

The row-random seven-topic benchmark yields ROC-AUC = 0.891. The preferred firm-grouped four-topic model yields AUC = 0.860, 95% cluster-bootstrap CI [0.849, 0.872], and PR-AUC = 0.661. At a .50 threshold, precision is 0.723, recall is 0.468, and F1 is 0.568. The comparatively modest recall is consistent with an unweighted model fitted at the observed minority-class prevalence; no resampling or class-weight correction was used. Its Brier score is 0.115; calibration is close to ideal (intercept = 0.090, slope = 1.030, ECE = 0.011). Keeping firms intact and refitting the text pipeline inside each fold reduces the headline result, but substantial ranking information remains.

Table 4. Out-of-sample classification performance

Model	ROC-AUC [95% CI]	PR-AUC	Brier	Precision	Recall	F1
Row-random benchmark, k=7	0.891 [0.880, 0.902]	0.734	0.102	0.753	0.561	0.643
Grouped primary, k=4	0.860 [0.849, 0.872]	0.661	0.115	0.723	0.468	0.568
Grouped topic-only, k=4	0.847 [0.836, 0.858]	0.594	0.125	0.665	0.391	0.493
Grouped sentiment-only	0.708 [0.691, 0.727]	0.530	0.147	0.756	0.268	0.396
Grouped direct mask, k=4	0.838 [0.825, 0.852]	0.638	0.121	0.712	0.437	0.541
Grouped strict mask, k=4 topic-only	0.785 [0.771, 0.798]	0.462	0.144	0.519	0.171	0.257
Grouped strict mask, k=7 topic-only	0.824 [0.812, 0.837]	0.578	0.131	0.671	0.309	0.423

Model	ROC-AUC [95% CI]	PR-AUC	Brier	Precision	Recall	F1
Grouped TF-IDF	0.937 [0.929, 0.944]	0.826	0.088	0.868	0.534	0.661
Grouped TF-IDF direct mask	0.926 [0.917, 0.934]	0.799	0.094	0.851	0.504	0.633
Grouped cue counts + length	0.648 [0.629, 0.666]	0.424	0.155	0.647	0.238	0.348

Note. Precision, recall, and F1 use a .50 threshold. Models use the observed class distribution without class weighting or resampling. Grouped intervals are based on 1,000 firm-level bootstrap samples. The row-random model is a benchmark and does not enforce firm independence. Lower Brier scores are better.

Observable target information is present but not sufficient to explain the result. Forced-label cues appear in 27.2% of forced notes and 3.9% of other notes; coder-meta phrases occur in 11.6% and 3.8%, respectively. Voluntary cues occur at similar rates in both groups (60.5% versus 59.6%), illustrating why literal cue counts alone are incomplete. The cue-count-plus-length model reaches only AUC = 0.648.

After direct labels and coder-meta phrases are masked, the selected model retains AUC = 0.838 and PR-AUC = 0.638. Under the strict four-topic topic-only specification, AUC = 0.785; the strict seven-topic topic-only sensitivity model reaches AUC = 0.824. These results show that recoverability extends beyond obvious label words, although the drop under strict masking confirms that rubric-related language contributes materially.

The high-dimensional TF-IDF model achieves the strongest ranking (AUC = 0.937, PR-AUC = 0.826) and the lowest Brier score (0.088). Its raw probabilities are nevertheless poorly aligned: the calibration intercept is 0.858, slope is 1.919, and ECE is 0.082. Under the standard calibration model, a slope above one does not indicate overconfidence; together with the positive intercept, it indicates probabilities that are too compressed and systematically low. The model is useful for ranking cases but requires recalibration before probability-based deployment.

Figure 5. Firm-grouped out-of-fold ROC curves

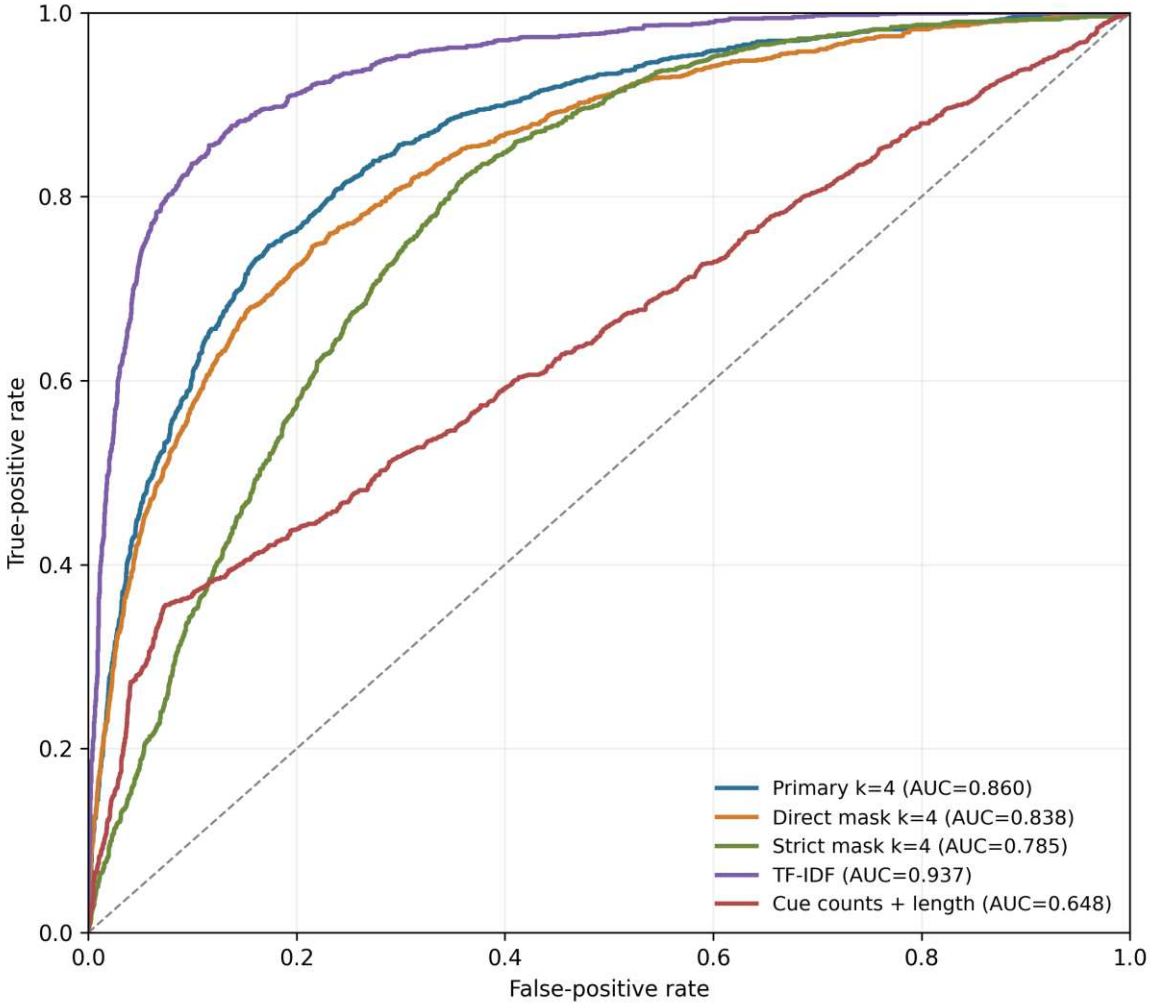
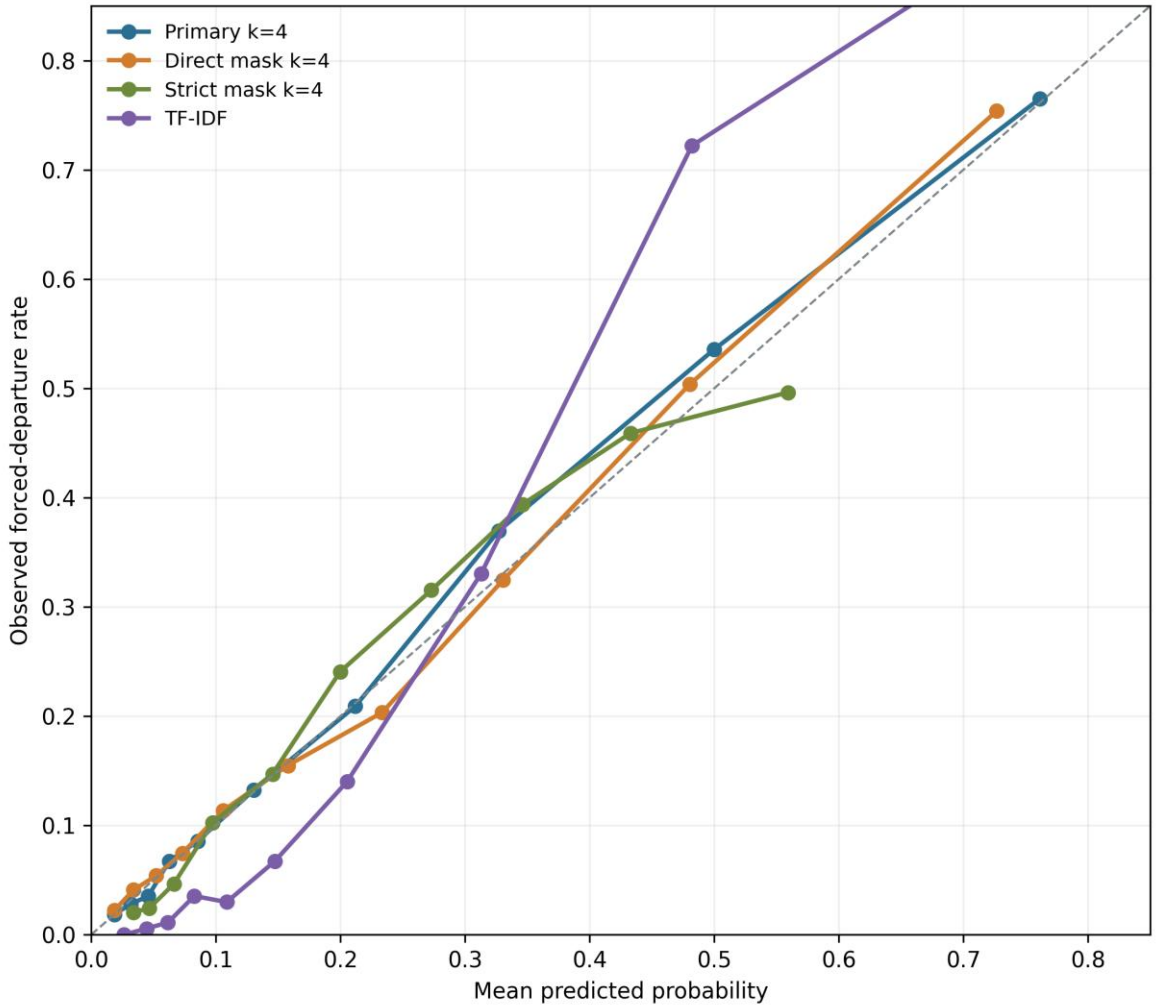


Figure 6. Calibration of grouped models



Note. Points are equal-frequency probability deciles; the dashed diagonal denotes ideal calibration.

4.5 Temporal transportability and source-mix checks

Temporal performance is not a mechanical consequence of stable class prevalence. Models trained on 2000-2013 and evaluated on 2014-2018 show that vocabulary and documentation patterns remain transportable, although masking matters. Table 5 reports both all later cases and the stricter subset of later firms absent from training.

Table 5. Temporal holdout performance, training years 2000-2013

Model	Test sample	n	Forced (%)	ROC-AUC	PR-AUC	Brier
Primary, k=4	All 2014-2018	1,447	22.6	0.886	0.728	0.103
Primary, k=4	Unseen firms, 2014-2018	474	18.6	0.875	0.688	0.093
Direct mask, k=4	All 2014-2018	1,447	22.6	0.799	0.591	0.135

Model	Test sample	n	Forced (%)	ROC-AUC	PR-AUC	Brier
Direct mask, k=4	Unseen firms, 2014-2018	474	18.6	0.798	0.519	0.121
Strict mask, k=4 topic-only	All 2014-2018	1,447	22.6	0.758	0.435	0.151
Strict mask, k=4 topic-only	Unseen firms, 2014-2018	474	18.6	0.762	0.410	0.133
Strict mask, k=7 topic-only	All 2014-2018	1,447	22.6	0.814	0.589	0.131
Strict mask, k=7 topic-only	Unseen firms, 2014-2018	474	18.6	0.816	0.599	0.112

Note. All transformations and models are learned using 2000-2013 cases only. Unseen firms have no observation in the training period.

The primary model reaches AUC = 0.886 across all 1,447 later cases and AUC = 0.875 among 474 later observations from unseen firms. Direct masking lowers later-period AUC to 0.799. The strict seven-topic topic-only model remains informative among unseen later firms (AUC = 0.816, PR-AUC = 0.599). Research Question 4 is therefore answered by substantial cross-firm and temporal recoverability, coupled with meaningful sensitivity to the information removed.

Table 6. Performance across source-mix strata

Source dimension	Status	n	Forced (%)	ROC-AUC	PR-AUC	Brier
SEC source	Present	4,006	22.9	0.862	0.673	0.115
SEC source	Absent	1,359	21.2	0.855	0.626	0.115
Major-news source	Present	1,117	36.5	0.836	0.747	0.163
Major-news source	Absent	4,248	18.7	0.864	0.626	0.103
Press-wire source	Present	1,151	15.2	0.841	0.545	0.096
Press-wire source	Absent	4,214	24.4	0.862	0.681	0.120

Note. Source flags indicate that a linked source of the stated type is present. Narratives remain coder-constructed blends; these are transportability checks, not source-authorship comparisons.

Grouped primary AUC ranges from .836 to .864 across the source-present and source-absent strata. Differences in PR-AUC partly reflect differing forced-departure prevalence: notes linked to major-news sources have a 36.5% forced share, whereas notes with press-wire sources have a 15.2% forced share. The consistency of ROC-AUC supports transportability across source environments. It cannot identify whether a diagnostic phrase originated in a firm release, SEC filing, news article, or coder synthesis.

Taken together, Tables 4-6 report many related model and stratum comparisons without a family-wise error-rate or other multiple-testing correction. We interpret their convergence and deterioration under stronger controls as a sensitivity pattern around one recoverability claim, not as a count of independently

significant findings. Individual intervals and significance statistics should still be read with that multiplicity qualification in mind.

5. Discussion

5.1 What the evidence establishes

Coder-constructed CEO departure narratives contain a strong and replicable signature of the assigned dismissal code. That signature is not confined to a row-random split, repeated firms, a single time period, exact duplicate notes, or a few literal words. It persists when firms are kept intact, when the representation is learned only from training data, when later firms are isolated, when direct label and coder-meta language are removed, and when sentiment is excluded under a strict topic-only design.

The magnitude is substantively meaningful. The grouped primary AUC of .860 is lower than the row-random benchmark but still indicates substantial ranking information. Direct masking reduces AUC by .022, and strict masking produces a larger decline, showing that both explicit cues and broader rubric language matter. Yet the strict seven-topic model remains above .82 in grouped and later unseen-firm tests. The signal is therefore neither trivial nor independent: it is distributed across narrative structure created during adjudication.

The four-topic solution helps explain what the model learns. Forced departures concentrate in narratives that combine performance evidence, contested-exit cues, retirement language, and coding rationale. Other discretionary exits concentrate in formal succession, executive biographies, and transactions. The large sentiment difference adds discrimination, but topic-only models confirm that polarity is not the sole mechanism.

5.2 What the evidence does not establish

The database notes were produced during the evidence review that generated the departure labels. Some contain explicit code language; many embed the facts used to justify a classification. Masking shows that prediction is not wholly driven by visible target words, but the coder's selection and ordering of evidence remain in the text. For that reason, the results cannot adjudicate the stronger claim that firms' managed narratives leak the truth.

High AUC does not loosen that boundary. A coding protocol may generalize precisely because it is applied consistently across firms and years. The result supports transport within that protocol. Firm-authored communication is a different object: it carries organizational intent, may be contested by journalists, and reaches an audience whose response is absent from the present data.

5.3 Theoretical contribution: text-label co-production and provenance validity

The paper's theoretical contribution is deliberately bounded. Text-label co-production adapts the established concerns of criterion contamination, criterion deficiency, and annotation artifacts to a management setting in which text is often treated as evidence about organizational action. Computational studies usually focus on the transformation from text to measure - tokenization, dictionaries, topics, or classifiers. Co-production directs attention one step earlier, to the transformation from event to text and label. When the same adjudicator produces both, the observed relationship may be reliable and transportable while remaining circular for the intended inference.

This distinction clarifies three analytically separate threats. Lexical leakage concerns direct target words and can be diagnosed through masks. Entity and temporal dependence concern repeated units and drift and can be addressed through grouped splits and later holdouts. Provenance dependence concerns the shared production of predictor and outcome. It cannot be resolved statistically within the co-produced dataset. The remedy is a different design: independently collected source texts, labels frozen without access to explanatory notes, and blinded validation.

For impression-management theory, this means that authorship is part of construct definition. A statement about organizational framing must identify the organizational text. Media scrutiny and coder synthesis are theoretically interesting but distinct processes. Separating these voices would allow future work to test whether firm euphemism, journalistic correction, and coder adjudication form a sequence rather than collapsing them into one narrative field.

5.4 Methodological and practical implications

For computational management research, the results reinforce a validation principle that is easy to state and often difficult to apply: the intended inference must govern the split, feature availability, metric, and corpus construction. Grouped validation is necessary when organizations repeat. Temporal holdouts are necessary when deployment concerns later cases. PR-AUC is necessary when the positive class is imbalanced. Calibration is necessary when probabilities will guide decisions. None substitutes for an authorship and provenance audit (Anglin 2024; Campion and Campion 2025; Park and Montgomery 2025). Nor is the concern peculiar to management data. Event catalogs in political sociology, coded case files in criminology, and annotated registries in education research are produced in much the same way, and the same questions deserve answers before any model is fit.

The calibration results illustrate why multiple metrics matter. TF-IDF ranks cases extremely well and improves Brier score, yet its predictions are systematically low and too compressed. Interpreting a slope of 1.919 as overconfidence would invert the standard diagnostic. The appropriate conclusion is not that TF-IDF is inferior, but that it serves a different purpose: high-quality ranking that needs recalibration and offers less transparent thematic interpretation.

For users of the CEO database, the narratives remain valuable. They can support record audit, identify cases for manual review, and assist consistent coding of new observations. In practice, the safest role is that of a second reader: the model prioritizes attention, while the analyst retains responsibility for the classification.

Box 1. Decision-use guide for boards and HR analytics teams

Use condition	Recommended action
Routine audit within the same database and coverage domain	Use model scores to rank cases for review; monitor PR-AUC and calibration on the current labeled sample.
Probability-based reporting or resource allocation	Recalibrate before use. Do not interpret the TF-IDF output as a ready-made probability of forced departure.
Ambiguous, novel, or out-of-period cases	Require human review, especially when source mix, vocabulary, or firm context differs from the training corpus.
Legal, reputational, board, or employment consequences	Treat automated output only as a prompt for documented expert review; retain the underlying source texts and an audit trail.

Note. The guide concerns triage and coding support. It does not convert a database-reproduction model into evidence about organizational intent.

5.5 Boundary conditions and the decisive next study

The evidence is bounded to large U.S. public firms, the 2000-2018 coverage window, a specific coding protocol, and short constructed narratives. VADER is a general-purpose sentiment tool; topic labels involve judgment; and the residual departure category is heterogeneous. The four topic shares were entered without a log-ratio transformation or reference-topic specification, which limits coefficient-level interpretation even though the penalized models are evaluated predictively. The chi-square and Cramer's V topic-association statistics were not adjusted for repeated events within firms. Strict masks cannot remove every semantic correlate of dismissal. We also lack a chance-corrected recoding coefficient and did not conduct a new independent human-coder exercise. Source flags identify linked materials but do not recover the words attributable to each source. Nor did we apply family-wise error control across the many model and stratum checks. Those analyses were specified as sensitivity probes around the principal grouped model, but their inferential statistics remain unadjusted for multiplicity.

A decisive next study should assemble separate documents for each event: the initial firm announcement, relevant SEC filing, contemporaneous wire story, and subsequent independent news coverage. The departure label should be frozen by coders who do not write explanatory notes visible to the model. Explicit target phrases should be masked prospectively, and framing dimensions should be rated by human coders blinded to forced status. Models should then be trained on early firms and tested on later unseen firms. Alongside the interpretable topic and TF-IDF specifications, a modern embedding- or transformer-based classifier could serve as an upper-bound sensitivity test. If it adds little, the lexical ceiling is probably near; if it improves sharply, semantic recoverability extends beyond the present feature space. The source-separated design, however, remains necessary for deciding whose language carries that signal.

Such a study would also reconnect the computational evidence to governance theory. Firm-specific framing could be linked to board independence, prior performance, successor origin, severance arrangements, market reaction, and media challenge. The present audit supplies the methodological boundary and the empirical benchmark; the source-separated design would supply the organizational inference.

6. Conclusion

The language of CEO departure notes contains a robust signature of forced turnover, but the meaning of that signature depends on provenance. Firm-grouped, temporally transported, and masked models show that recoverability extends beyond direct label words. Because the analyzed text is a coder-constructed account produced alongside the label, the result speaks to within-system recoding rather than to firms' success or failure in managing the story.

The strongest defensible conclusion is therefore two-sided. The database narratives are excellent measurement-audit resources and support reliable within-system classification. At the same time, text-label co-production places a hard boundary on behavioral and communication claims. Computational rigor requires both strong models and disciplined restraint about what the data-generating process allows those models to mean.

References

Anglin, K.: Addressing threats to validity in supervised machine learning: a framework and best practices for education researchers. AERA Open 10, 1-26 (2024). https://doi.org/10.1177/23328584241303495

Bednar, M.K.: Watchdog or lapdog? A behavioral view of the media as a corporate governance mechanism. Acad. Manag. J. 55(1), 131-150 (2012)

Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993-1022 (2003)

Campion, E.D., Campion, M.A.: A review of text analysis in human resource management research: methodological diversity, constructs identified, and validation best practices. Hum. Resour. Manag. Rev. 35(2), 101078 (2025). https://doi.org/10.1016/j.hrmr.2025.101078

Choudhury, P., Allen, R.T., Endres, M.G.: Machine learning for pattern discovery in management research. Strateg. Manag. J. 42(1), 30-57 (2021). https://doi.org/10.1002/smj.3215

Coughlan, A.T., Schmidt, R.M.: Executive compensation, management turnover, and firm performance: an empirical investigation. J. Account. Econ. 7(1-3), 43-66 (1985)

Eisenhardt, K.M.: Agency theory: an assessment and review. Acad. Manag. Rev. 14(1), 57-74 (1989)

Elsbach, K.D.: Managing organizational legitimacy in the California cattle industry: the construction and effectiveness of verbal accounts. Adm. Sci. Q. 39(1), 57-88 (1994)

Gentry, R.J., Harrison, J.S., Quigley, T.J., Boivie, S.: A database of CEO turnover and dismissal in S&P 1500 firms, 2000-2018. Strateg. Manag. J. 42(5), 968-991 (2021). https://doi.org/10.1002/smj.3278

Graffin, S.D., Carpenter, M.A., Boivie, S.: What's all that (strategic) noise? Anticipatory impression management in CEO succession. Strateg. Manag. J. 32(7), 748-770 (2011)

Grimmer, J., Stewart, B.M.: Text as data: the promise and pitfalls of automatic content analysis methods for political texts. Polit. Anal. 21(3), 267-297 (2013)

Hannigan, T.R., Haans, R.F.J., Vakili, K., Tchalian, H., Glaser, V.L., Wang, M.S., Kaplan, S., Jennings, P.D.: Topic modeling in management research: rendering new theory from textual data. Acad. Manag. Ann. 13(2), 586-632 (2019). https://doi.org/10.5465/annals.2017.0099

Hutto, C.J., Gilbert, E.: VADER: a parsimonious rule-based model for sentiment analysis of social media text. Proc. Int. AAAI Conf. Web Soc. Media 8(1), 216-225 (2014)

Jensen, M.C., Meckling, W.H.: Theory of the firm: managerial behavior, agency costs and ownership structure. J. Financ. Econ. 3(4), 305-360 (1976)

Jenter, D., Kanaan, F.: CEO turnover and relative performance evaluation. J. Finance 70(5), 2155-2184 (2015)

Kaplan, S.N., Minton, B.A.: How has CEO turnover changed? Int. Rev. Finance 12(1), 57-87 (2012)

Kaufman, S., Rosset, S., Perlich, C., Stitelman, O.: Leakage in data mining: formulation, detection, and avoidance. ACM Trans. Knowl. Discov. Data 6(4), Article 15 (2012). https://doi.org/10.1145/2382577.2382579

Kobayashi, V.B., Mol, S.T., Berkers, H.A., Kismihok, G., Den Hartog, D.N.: Text mining in organizational research. Organ. Res. Methods 21(3), 733-765 (2018). https://doi.org/10.1177/1094428117722619

Loughran, T., McDonald, B.: When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66(1), 35-65 (2011)

Park, J.Y., Montgomery, J.M.: Toward a framework for creating trustworthy measures with supervised machine learning for text. Polit. Sci. Res. Methods (2025). https://doi.org/10.1017/psrm.2025.10042

Pollock, T.G., Rindova, V.P.: Media legitimation effects in the market for initial public offerings. Acad. Manag. J. 46(5), 631-642 (2003)

Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE 10(3), e0118432 (2015). https://doi.org/10.1371/journal.pone.0118432

Sutton, R.I., Callahan, A.L.: The stigma of bankruptcy: spoiled organizational image and its management. Acad. Manag. J. 30(3), 405-436 (1987)

Van Calster, B., McLernon, D.J., van Smeden, M., Wynants, L., Steyerberg, E.W.: Calibration: the Achilles heel of predictive analytics. BMC Med. 17, 230 (2019). https://doi.org/10.1186/s12916-019-1466-7

Weisbach, M.S.: Outside directors and CEO turnover. J. Financ. Econ. 20, 431-460 (1988)

Wiesenfeld, B.M., Wurthmann, K.A., Hambrick, D.C.: The stigmatization and devaluation of elites associated with corporate failures: a process model. Acad. Manag. Rev. 33(1), 231-251 (2008)

Yarkoni, T., Westfall, J.: Choosing prediction over explanation in psychology: lessons from machine learning. Perspect. Psychol. Sci. 12(6), 1100-1122 (2017). https://doi.org/10.1177/1745691617693393

Appendix A. Masking Rules and Interpretation

Direct mask. Removes explicit forced and voluntary exit labels/actions and coder-process phrases. The families include fired/dismissed/ousted/terminated/forced/involuntary/removed/asked to resign; retired/resigned/stepped down/new opportunity/voluntary; and code/recode/classification/changed code/departure code/numbered code references. Inflectional variants are removed through case-insensitive regular expressions.

Strict mask. Applies the direct mask and additionally removes performance, missed-target, misconduct, fraud, investigation, legal/ethical/policy, bankruptcy, illness/death, interim succession, replacement/appointment, merger/acquisition/transaction, effective-immediately, abruptness, and conflict families. Strict topic models exclude VADER sentiment.

Interpretation boundary. A direct-mask result estimates how much recoverability remains without readily observable target terms. A strict-mask result estimates how much remains after a broader set of rubric-like semantic cues is removed. Neither mask removes the coder's selection and organization of evidence; neither converts the narrative into an independent firm-authored source.

Appendix B. Data-Integrity and Reproducibility Checks

Table B1. Reproducibility and data-integrity audit

Check	Result
Raw records	9,423
Valid coded departures, 2000-2018	5,516
Discretionary departures	5,381
Common textual sample	5,365
Firms in common textual sample	2,725
Reused dataset identifiers	379 identifiers across 759 rows
Rows in exact duplicate narratives	14 (0.26%)
Cross-firm exact duplicate rows	2
Grouped validation	Stratified five-fold by gvkey
Temporal holdout	Train 2000-2013; test 2014-2018
Class treatment	Observed distribution; no weighting or resampling
Topic-share treatment	All shares retained; L2 penalty; no log-ratio transform
Recoding reliability	87% raw agreement; chance-corrected coefficient unavailable
Lexical upper-bound model	TF-IDF benchmark; no embedding/transformer model
Bootstrap uncertainty	1,000 firm-level resamples
Random seed	20260807 (splits); 42/314/2718 (topic selection)