Customer Feedback as Governance Infrastructure: An External Accountability Layer for Automated BFSI Systems

Rohit Sarang
Finlytix AI, New Delhi, India
hello@rohitsarang.com, contact@finlytix.ai
Working Paper | June 2026

***Abstract*—This paper presents a framework that closes the systemic observability gap in automated financial systems by converting unstructured public customer feedback into structured operational diagnostics, attributing failures to specific processes, operational domains, and institutional owners, and demonstrating its function as an independent external accountability layer that complements and does not depend on internal institutional reporting. The framework comprises a multi-model classification pipeline, a nine-stage analytical pre-processing layer, evidence-constrained LLM synthesis, and a decision engine producing prioritised findings with revenue exposure estimates and named institutional owners.**

***Keywords*—Accountability, Oversight, Recourse, Automated BFSI Systems, Governance Infrastructure, Operational Resilience, Root Cause Analysis, NLP, External Observability, Agentic AI Systems.**

I. Problem Statement

Existing governance frameworks for automated financial systems share a structural flaw: they depend on signals generated by the same systems they are meant to govern. Internal audit logs, system dashboards, and incident reports are produced by institutional infrastructure. When that infrastructure fails, the failure is often invisible to the very mechanisms designed to detect it.

Customer feedback is structurally different. When an automated payment pipeline silently debits without crediting, or when a KYC verification loop traps a customer in a broken authentication flow, the failure appears in customer complaints within minutes, before it surfaces as an attributable pattern in institutional reporting, and before a regulator receives a report.

This creates a systemic observability gap. Automated BFSI systems increasingly govern credit access, payment execution, identity verification, and account lifecycle management. Yet the governance frameworks monitoring these systems have no independent external signal. Internal incident dashboards depend on predefined thresholds that must be crossed before an alert fires. Grievance systems require customers to escalate individually before a pattern is visible. Operational teams monitor system events, not customer-perceived failure. Fragmented ownership across technology, product, and operations functions delays detection further. Consumer harm accumulates silently. This framework reduces failure detection latency from weeks to hours.

II. Context and Domain

The system operates across BFSI institutions in India and West Africa, covering full-service commercial banks, digital-native neobanks, and fintech payment platforms. Source data is drawn exclusively from publicly available complaints, app store reviews, social media listening, and complaint feeds. No integration with internal institutional systems is required at any stage, enabling deployment as a genuinely independent external oversight instrument.

The operational scope spans automated systems governing payments and fund transfers, KYC and identity verification, authentication and access control, card issuance and operations, lending workflows, and account lifecycle management. Each domain carries distinct regulatory obligations under frameworks including RBI Digital Lending Guidelines, RBI Credit Card Master Direction, grievance redressal mandates, and operational resilience requirements for regulated financial entities.

The pipeline has processed 1.8 million customer reviews across 30 BFSI entities. The independence from internal systems is not merely a technical convenience. It is a governance property. It means the signal cannot be suppressed, filtered, or delayed by the institution being monitored. While the framework operates on public data by default, the pipeline architecture is equally applicable to internal complaint and support data where institutional access permits.

III. Approach and Contribution

Fig. 1 shows the classification pipeline from raw review text to a structured attributed corpus. Fig. 2 shows the intelligence pipeline from the classified corpus to executive-grade report output.

The framework operates as a three-stage pipeline. First, named entities and operational aspects are extracted from raw review text and structured as entity-aspect pairs. Second, each pair is classified against a BFSI-specific operational taxonomy, with the label space restricted at inference time to the institution's product profile: a payments-only fintech is evaluated against a different label set than a full-service

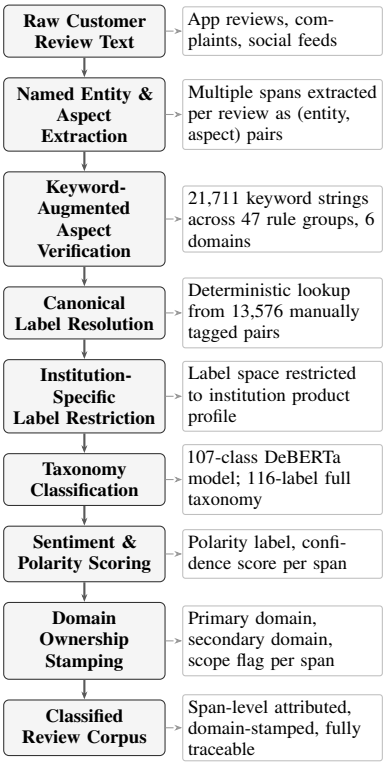

Fig. 1. Classification pipeline: raw review text to attributed classified corpus.

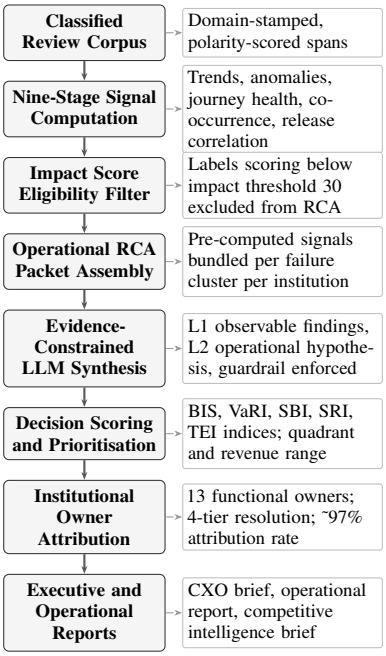

Fig. 2. Intelligence pipeline: classified corpus to attributed operational report.

commercial bank. Third, prioritised failure clusters are synthesised into attributed operational diagnostics with named institutional owners and recommended resolution pathways. Customer-perceived failure is treated as a governance signal, not a confirmed backend diagnosis.

A. Named Entity and Aspect Extraction

A DeBERTa-v3-large model (434M parameters) fine-tuned locally on 696,288 manually annotated tokens across 13,391 BFSI customer reviews performs BIO-tagged extraction of named entities and operational aspects from raw review text. Each review may yield multiple spans, where each span is a structured (entity, aspect) pair representing a distinct operational complaint unit. These pairs are the atomic unit of all downstream analysis. The corpus was split into 11,382 training reviews and 2,009 held-out validation reviews. Validation yielded F1 macro 0.924 and token-level accuracy 98.47%. All annotation was performed manually by a single BFSI domain expert with 14 years of experience in payments and cards operations over six months. No crowdsourced or synthetic labelling was used.

B. Taxonomy Classification

Classification proceeds through three sequential tiers, each operating within a label space already restricted to the institution's active operational domains.

The Gold tier resolves entity-aspect pairs through a canonicalizer built from 13,576 manually tagged instances, returning taxonomy labels deterministically for known pairs without model inference.

The Silver tier applies a keyword-augmented verification layer built from 21,711 domain-specific keyword strings across 47 rule groups spanning six operational domains. This tier handles categories where learned classification alone is insufficient due to sparse or highly domain-specific lexical patterns.

Between the Silver and Bronze tiers, Layer A restricts the label space at inference time based on the institution's product profile. A payments-only platform is constrained to payment and core labels. A full-service commercial bank activates the full taxonomy. This prevents a card-only institution from being evaluated against lending taxonomy labels it does not operate, materially reducing classification noise.

The Bronze tier applies a DeBERTa-v3-large classifier (434M parameters) trained on the same 13,576 manually labelled instances, each comprising an entity, aspect, and full review context concatenated as structured input. The classifier covers 107 operational classes representing the 116-label full taxonomy minus 9 labels excluded from training due to insufficient corpus representation. Validation on 4,483 held-out examples yielded overall accuracy 95.47% and macro F1 0.915, with per-class F1 stable above 0.85 across the majority of rare categories. Weighted loss training addresses corpus imbalance across classes.

Every classified span receives a domain scope stamp from {primary_active, secondary_active, excluded_region, excluded_client, unmapped}, providing full traceability of every inclusion and exclusion decision at span level. No classification decision is silent.

C. Signal Computation and Packet Construction

Classified spans enter a nine-stage analytical pipeline that converts raw taxonomy distributions into structured operational signals before any LLM synthesis is invoked.

Stage 1 (Label Drilldown) computes frequency and negative polarity distributions per taxonomy label per institution, establishing a baseline complaint profile across all 116 operational categories.

Stage 2 (Trend Analysis) produces temporal trend vectors per label, identifying whether complaint volume for each category is accelerating, stable, or declining over the observation window.

Stage 3 (Impact Scoring) computes a composite impact score per label combining complaint frequency, negative rate, and recency weighting. Labels scoring below a threshold of 30 are excluded from RCA synthesis at this stage. This eliminates low-signal noise before the LLM is invoked.

Stage 4 (Release Regression Scoring) correlates complaint spikes with app version release dates, identifying whether a specific release is associated with elevated failure signals in any operational category.

Stage 5 (Co-occurrence Analysis) identifies taxonomy label pairs that appear together consistently within the same review, surfacing cascading failure patterns where one operational failure triggers a downstream failure in a separate domain.

Stage 6 (Journey Mapping) assigns each label to a customer journey stage and computes a health score per stage, producing a journey-level view of where friction concentrates across onboarding, transaction execution, servicing, and account lifecycle.

Stage 7 (Anomaly Detection) flags labels whose recent complaint velocity exceeds their historical baseline by a statistically significant margin, distinguishing structural operational problems from transient spikes.

Stage 8 (Pain Intensity Scoring) weights each span by recency and polarity severity, producing a composite pain score per cluster that captures not just volume but the intensity of customer-reported distress.

Stage 9 (Packet Builder) reads all upstream signal outputs and assembles a structured RCA packet per failure cluster per institution. The packet contains the full review text corpus for the cluster, all pre-computed signal outputs from Stages 1 through 8, polarity distributions, temporal trend direction, anomaly flags, and journey stage assignments. This packet, not raw text alone, is what the LLM receives.

Labels are organised into 18 predefined operational failure clusters defined in a single cluster configuration file covering authentication and OTP failures, transaction processing errors, card access and usage, fund transfer experience, account blocking and restrictions, app stability and performance, balance and statement visibility, KYC and account onboarding, customer support and grievance, lending and credit management, fee and charge transparency, rewards and benefits, ATM and cash operations, fraud and unauthorised access, post-update functional degradation, and app navigation and interface. A UPI-specific cluster is automatically excluded for African region clients via the region configuration layer, as UPI infrastructure is India-specific.

D. Evidence-Constrained Operational Hypothesis Generation

Failure clusters are scored for prioritisation by:

$$\text{Score}(\text{cluster}) = N_{\text{neg}} \times \left(\frac{N_{\text{neg}}}{N_{\text{total}}} \right) \tag{1}$$

This formulation ranks clusters by the product of absolute complaint volume and complaint rate, surfacing failures that are simultaneously high-frequency and high-severity. Span-level attribution and domain rollup operate across all classified spans regardless of cluster score, ensuring low-volume but high-severity signals remain visible in the governance output and are not suppressed by volume-only ranking.

For each prioritised cluster, a locally hosted open-source LLM receives the structured RCA packet. The prompt architecture enforces two output layers per finding. L1 Observable captures what customers report with exact frequency counts drawn from the packet. L2 Operational captures what process the patterns suggest is involved, with mandatory uncertainty language that prevents the model from asserting confirmed root causes. A forbidden inference guardrail prohibits assertion of backend, infrastructure, or architectural causes without explicit review evidence, preventing hallucinated operational hypotheses from entering institutional outputs.

Each finding also produces a structured L2 analytical block containing: the probable root cause stated as a hypothesis, the relevant taxonomy label from the 116-label set, a recommended institutional action, and an unresolved risk statement describing the likely consequence if the finding is not addressed. The taxonomy label field forces the LLM to ground each finding in the established classification schema, which then drives owner attribution in the downstream decision layer.

A quality gate blocks cross-institution synthesis when any institution input falls below a minimum span volume threshold. Where the gate passes, the synthesis layer produces three outputs: common failure identification across institutions, primary institution signal isolation, and a competitive gap table quantifying failure frequencies against best-performing peers.

E. Decision Intelligence and Revenue Attribution

RCA findings feed a decision engine that parses structured finding blocks into one row per finding and computes five composite indices normalised to a 0-100 scale.

The Business Impact Score (BIS) combines finding frequency within the cluster, complaint concentration rate, pain intensity score, and anomaly flag. It answers: how operationally severe is this finding?

The Value at Risk Index (VaRI) combines BIS with estimated financial exposure. It answers: what should be fixed first?

The Support Burden Index (SBI) estimates operational bandwidth drain from the finding's complaint volume and

expected resolution effort. It answers: what is consuming operations capacity?

The Switching Risk Index (SRI) is computed directly from span-level signals, measuring the presence of customer defection language, competitor references, and account closure intent within the cluster.

The Trust Erosion Index (TEI) measures longitudinal sentiment deterioration for the finding's associated labels, identifying categories where customer sentiment is quietly worsening over time without yet producing a volume spike.

Financial exposure is estimated from institution-level parameters: estimated active user base, conservative and base ARPU, and support cost per complaint interaction. Revenue at risk per finding is computed as the product of estimated affected user fraction, ARPU range, and churn probability implied by complaint severity and SRI signal. This produces a revenue range rather than a point estimate, explicitly reflecting the uncertainty in translating customer-perceived failure signals to financial outcomes.

Each finding is assigned to one of four action quadrants. Immediate Action findings carry Critical priority and require escalation regardless of remediation complexity. Quick Win findings carry high VaRI with low remediation effort. Strategic Project findings carry high VaRI with medium or high remediation effort. Monitor findings fall below the VaRI threshold and are surfaced for longitudinal observation without immediate escalation.

F. Institutional Owner Attribution

Every finding receives a primary institutional owner mapped through a four-tier resolution process operating against a 116-label to 13-function attribution table.

Tier 1 performs exact match on the taxonomy label extracted by the LLM from the L2 analytical block. Because the LLM is constrained to return an exact label from the 116-label taxonomy, this tier resolves the majority of findings without ambiguity.

Tier 2 performs token intersection between the finding name and the taxonomy labels associated with its cluster. Tier 3 performs token intersection against all 116 taxonomy labels globally, catching findings whose names reference concepts outside their primary cluster.

Tier 4 assigns an Under Review status for the small fraction of findings where no match is established across the first three tiers.

During internal validation across four West African BFSI institutions, the four-tier resolution process attributed approximately 97% of findings to a named institutional function. The 13 functional owner categories are: Retail Banking Operations, Digital Banking and Mobile Channels, Cards Operations, Payments and Transaction Processing, Consumer Lending and Credit, KYC and Identity Verification, Customer Service and Grievance Management, Fraud and Financial Crime Control, IT and Core Banking, Compliance and Regulatory Affairs, Wealth and Investment Services, Marketing and Product Management, and Customer Retention and Lifecycle Management.

IV. Governance Relevance

This framework directly addresses all three sub-themes of Pillar 2 through demonstrated mechanism.

A. Auditability and Traceability

Every span carries a complete attribution chain: raw text, extracted entity, aspect, canonical form, taxonomy label, primary domain, secondary domain, and domain scope flag. This chain is reproducible and inspectable independent of institutional cooperation. The domain scope stamp distinguishes active classifications from region-excluded, client-excluded, and unmapped labels, ensuring every decision is traceable. This architecture directly supports supervisory requirements under model risk oversight frameworks and algorithmic accountability obligations emerging across BFSI jurisdictions.

B. Recourse Frameworks

The ownership attribution layer creates a structured recourse map: from customer-reported symptom to operational hypothesis to responsible institutional function. This addresses a critical gap in current grievance redressal frameworks: the inability to attribute systemic failure patterns to specific operational owners before individual complaints are filed. Consumer protection bodies and regulators can use this signal to identify unresolved complaint concentrations, detect consumer harm latency, and direct supervisory attention proportionately.

C. Risk Assessment and Operational Resilience

The cluster scoring formula provides a quantified risk prioritisation mechanism weighted by both complaint volume and complaint concentration. Combined with L1/L2 outputs, BIS, VaRI, and competitive benchmarking, institutions receive a risk-ranked operational failure view with peer context and estimated revenue exposure. For regulators, this constitutes an independent observability layer for assessing operational resilience across automated financial systems, without dependence on self-reported institutional data.

Outputs are structured for direct consumption by operational risk, compliance, grievance redressal, and engineering functions without modification to existing institutional workflows, making the framework applicable as an independent oversight instrument under digital lending, consumer protection, and automated decision-making guidelines. This applicability will only grow. As BFSI automation accelerates toward agentic systems capable of autonomous credit decisioning, AI-led servicing, and multi-step workflow execution, failures will surface first in customer-reported signals before internal monitoring detects them. External observability frameworks of this kind are not supplementary to governance infrastructure. For agentic financial systems, they are structurally necessary.

TABLE I
INSTITUTION A: WEST AFRICA, COMMERCIAL BANK

Field	Value
Review Text	"[Bank] secure pass no longer on playstore. App crashed every time at final stage. Balance unavailable half the time."
Spans Extracted	13
Unique Taxonomies	8
Domains Attributed	Authentication, Transaction Processing, App Stability, Balance Information
Operational Owners	Security, Payments Ops, Engineering, Core Banking
Governance Signal	Authentication dependency on deprecated tool creating systemic access failure

TABLE II
INSTITUTION B: INDIA, FINTECH PAYMENT PLATFORM

Field	Value
Review Text	"Wallet expired, unable to transfer 25k+. KYC not working. Upload button broken. Ticket just buffers."
Spans Extracted	9
Unique Taxonomies	9
Domains Attributed	Wallet Access, KYC Verification, Chatbot Support, Service Request Management
Operational Owners	Compliance, KYC Ops, Support, Service Ops
Governance Signal	Regulatory verification failure blocking customer funds access

TABLE III
INSTITUTION C: INDIA, PRIVATE SECTOR BANK

Field	Value
Review Text	"Closed credit card, they didn't pay reward points back. Planning to close savings account."
Spans Extracted	9
Unique Taxonomies	9
Domains Attributed	Reward Operations, Account Blocking, Closure Processing, Charge Transparency
Operational Owners	Rewards, Cards Ops, Retail Banking, Product
Governance Signal	Sequential exit journey detectable 2 stages before account closure

V. EARLY EVIDENCE AND GROUNDING

Validation across three institutions spanning two geographies demonstrates cross-regional generalisability. Tables I, II, and III present the end-to-end pipeline output per institution.

Institution A reveals a systemic authentication failure cascading across transaction execution. Institution B demonstrates a regulatory compliance failure (KYC loop) blocking customer access to funds, directly relevant to consumer protection mandates. Institution C demonstrates the system's capacity to detect account churn intent two operational stages before closure, enabling institutional intervention within the recourse window.

Beyond the three institution examples above, the same pipeline was run end to end across four West African BFSI institutions during internal validation. Findings were distributed across all four action quadrants, owner attribution resolved without manual intervention for the large majority of findings, and competitive synthesis identified failure clusters common to multiple institutions, separating shared industry-wide patterns from institution-specific gaps.

Aggregate deployment across 1.8 million reviews and 30 BFSI entities across India and West Africa confirms cross-regional generalisability of the taxonomy and attribution framework.

A. Limitations

This framework carries inherent constraints that bound its governance application. Public review platforms exhibit selection bias: customers who complain publicly are not representative of all affected users, and silent failures affecting digitally excluded or low-literacy populations may not surface in this signal. Platform moderation and filtering may suppress high-severity complaint signals before ingestion. Coordinated review manipulation poses a data integrity risk. Language imbalance across the corpus means low-resource languages are underrepresented in the training distribution. The framework detects customer-perceived failure, not confirmed system failure. Operational hypotheses require institutional verification before remediation. Revenue estimates depend on institution-supplied financial parameters and carry uncertainty that widens at the individual finding level. These constraints position the framework as a complementary observability layer, not a replacement for internal monitoring systems.