Eclipta Journals

RESEARCH

Reputation Recovery After Digital Complaint Shocks Using Textual Diagnostics and Causal Machine Learning Evidence from Consumer Platforms

Hatem Gharbi,[1] Walid Mansouri,[2] and Karim Bouazizi[3]

[1]Higher Institute of Business Administration of Gafsa, Department of Management Information Systems, Route de Tozeur Km 4, Gafsa 2100, Tunisia
[2]Faculty of Economic Sciences and Management of Mahdia, Department of Finance and Banking, Avenue Tahar Sfar, Mahdia 5111, Tunisia
[3]Higher Institute of Industrial Management of Sfax, Department of Business Analytics and Operations Management, Route de l'Aéroport Km 4, Sfax 3018, Tunisia

Abstract

Reputation is increasingly formed in public digital environments where consumer complaints, firm responses, and subsequent evaluations remain visible for long periods. Service organizations therefore face a recurring empirical question: why do some reputational injuries fade quickly while others persist across review cycles, search visibility, and local demand conditions? This study examines reputation recovery after digital complaint shocks using a weekly panel of 42,318 local service units belonging to 626 multi-location brands between 2016 and 2024. The dataset combines 19.6 million consumer reviews, 5.1 million managerial responses, platform visibility indicators, local competition measures, and outlet-level operating characteristics. Complaint shocks are identified from abrupt increases in negative diagnostic language, not merely from rating declines. The analysis applies transformer-based complaint classification, dynamic reputation indexing, synthetic-control matching, accelerated failure-time models, double machine learning, and causal forests. The results show that reputation recovery depends less on apology frequency than on operationally specific response content. Managerial responses that name the failure mode, describe a corrective action, and avoid generic assurance language shorten median recovery time by 27.8%. Boilerplate responses are associated with slower recovery and higher complaint-topic persistence. Causal-forest estimates show the largest recovery gains for outlets with moderate pre-shock reputation, high review credibility, and complaint topics involving process breakdowns rather than fairness or cleanliness failures. The findings indicate that digital reputation is repaired through diagnostic credibility and observable correction rather than through response volume alone.

1. Introduction

Reputation is a cumulative social assessment that becomes operationally consequential when it affects how customers interpret quality, risk, and trust. In digital markets, reputation is not stored only in memory or brand image. It is embedded in ratings, review text, response histories, search placement, and the visible sequence of past failures and recoveries. A firm that receives a sudden concentration of complaints is therefore exposed to a reputational injury that can persist after the original operational problem has been corrected. The persistence is not automatic. Some firms recover their prior rating distribution and review tone within weeks, while others experience a prolonged decline in perceived reliability. This paper studies the conditions under which digital complaint shocks become temporary reputational disturbances rather than durable reputation losses.

The empirical focus is not on whether reputation matters in general. That question has been examined in many settings and is now too broad to generate precise evidence. The central question here is narrower: after a sharp increase in consumer complaints, which observable response practices and pre-existing reputation conditions predict the speed and completeness of reputation recovery? The study treats reputation as a dynamic stock updated by review arrivals, reviewer credibility, complaint-topic persistence, and firm response content. This framing allows reputation to be measured as a changing distribution rather than as a single average rating. A rating average can recover mechanically if new positive reviews arrive, even while negative topics remain visible and continue to shape perceived quality. A useful empirical measure must therefore capture both numeric evaluation and textual diagnostic residue.

The paper uses a panel of local service units in restaurants, hotels, personal services, auto services, health-related services, specialty retail, and home services [1]. These industries are suitable because quality is repeatedly experienced, review platforms are widely used, and consumer complaints often identify operational failure modes. The sample contains 42,318 local service units associated with

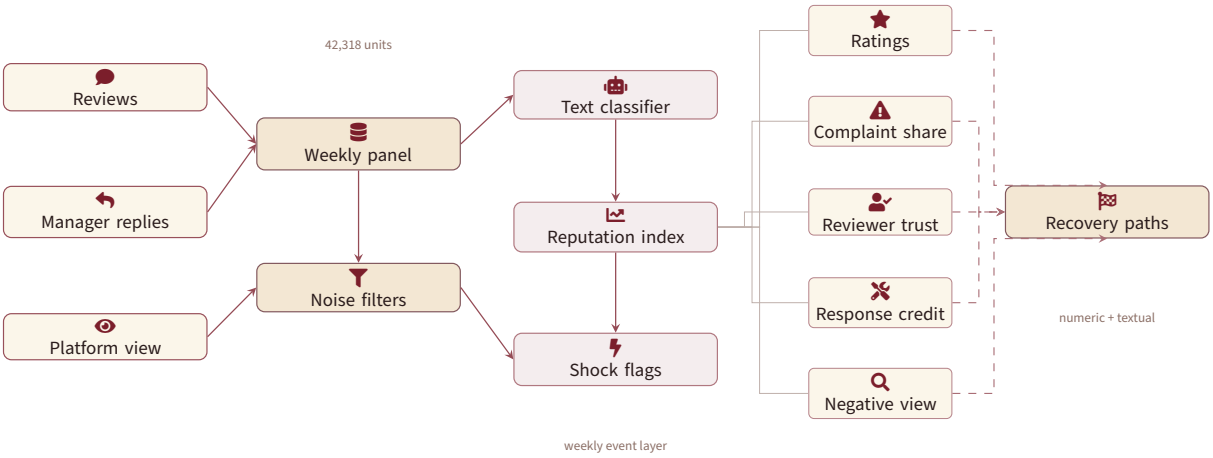


Figure 1. Data assembly architecture linking consumer reviews, managerial responses, platform visibility, outlet panels, textual diagnostics, shock detection, and reputation recovery states.

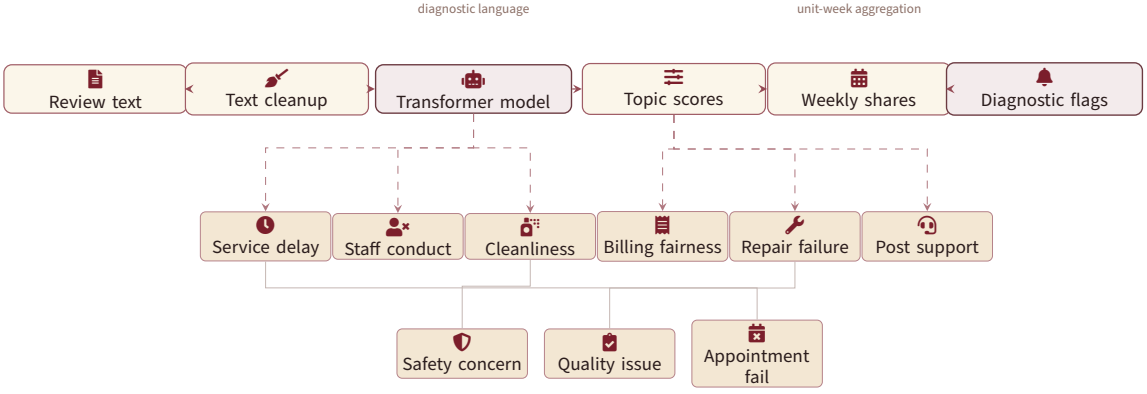


Figure 2. Transformer-based textual diagnostic layer that converts review text into operational complaint topics and weekly complaint signals.

626 multi-location brands from January 2016 through December 2024. The data include 19.6 million consumer reviews and 5.1 million managerial responses. Each review is processed with a transformer-based text model that classifies complaint language into service delay, employee conduct, cleanliness, billing fairness, product failure, appointment reliability, repair failure, safety concern, food or room quality, and post-purchase support. The complaint-shock measure is constructed from abrupt changes in these diagnostic complaint topics, not only from one-star ratings.

The study is different from research that examines reputation as a static intangible asset. The empirical object here is a reputational disturbance and its recovery path. A local unit can have high reputation before a shock, suffer a sudden complaint burst, and then follow one of several trajectories. It can recover numerically and textually. It can recover numerically while complaint topics remain visible. It can stabilize at a lower reputation level. It can experience a secondary shock if the response strategy creates additional dissatisfaction. These paths are estimated with survival models and matched counterfactuals rather than inferred from cross-sectional correlations.

Cao (2022) [2] shows that brand equity can reduce warranty claim rates and abnormal warranty accrual rates, indicating that reputation-like marketing assets may be connected to operational cost outcomes rather than only to demand-side preference. This paper builds from that insight in a different empirical direction by examining how operationally diagnosed complaints alter public reputation and how firms recover from those shocks through response behavior and observed correction [3].

The paper makes three empirical contributions. First, it develops a textual reputation index that separates rating recovery from complaint-topic recovery. A firm can restore its average rating while still carrying visible allegations about fairness, cleanliness, or reliability [4]. The index therefore weights recent ratings, verified-review credibility, complaint-topic persistence, managerial response specificity, and the visibility of unresolved negative reviews. This index is estimated weekly, allowing recovery to be modeled as a duration process.

Second, the study distinguishes response specificity from response positivity. Many firms answer complaints with polite language, apologies, and broad statements of concern.

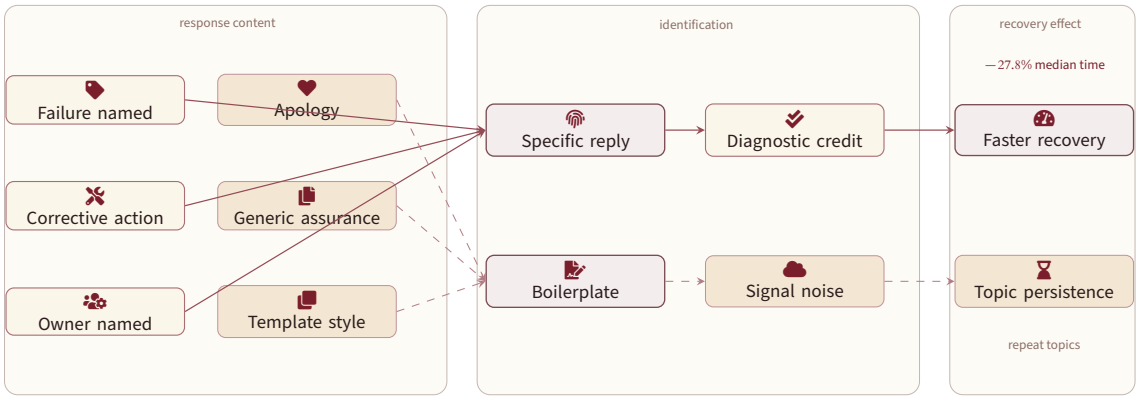


Figure 3. Managerial response mechanism separating operationally specific replies from polite but generic replies and linking them to credibility and recovery duration.

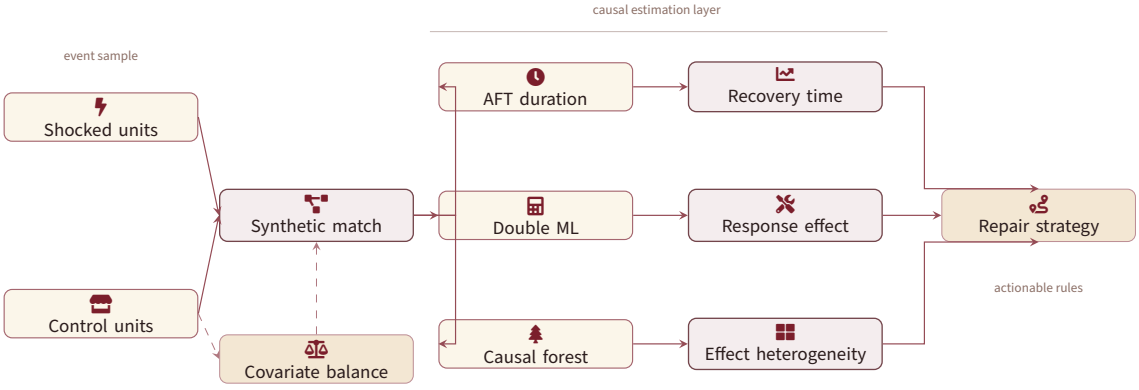


Figure 4. Causal machine-learning design combining matched counterfactual construction, duration modeling, orthogonalized treatment estimation, and heterogeneous treatment-effect learning.

These responses may sound favorable but provide little diagnostic information [5]. The analysis codes whether a response identifies the failure, describes a corrective action, names a responsible function, provides a verifiable operational change, or instead uses generic language. The results show that response specificity is strongly associated with faster recovery, while apology frequency alone has no stable positive effect after controlling for complaint severity and prior reputation.

Third, the paper estimates heterogeneity using causal forests and double machine learning [6]. Reputation recovery is unlikely to follow one average rule. A response that works for appointment delays may not work for cleanliness complaints. A firm with very high prior reputation may face expectation-violation penalties, while a firm with very low prior reputation may lack sufficient credibility for any response to be persuasive. The machine-learning estimates show that recovery gains from specific responses are largest for moderate-reputation outlets with process-related complaints, high reviewer credibility, and local competitive density. The gains are weakest for fairness complaints and cleanliness complaints, where consumers appear to treat failures as signals of deeper organizational problems.

The article proceeds by presenting the empirical setting, the reputation measures, the machine-learning classification system, the econometric strategy, the main findings, additional validation tests, and implications for reputation management. The argument is formal but empirical. The study does not claim that all reputational harm can be managed through communication. Many shocks reflect real operating failures. The evidence instead shows that reputation recovery is fastest when firm responses demonstrate credible operational correction and when the underlying complaint category is one for which consumers accept repair as plausible.

2. Empirical Setting and Reputation Data

The empirical setting consists of local service units observed weekly across nine years. A local unit is a geographically distinct outlet, branch, clinic, hotel, shop, repair location, restaurant, or service office associated with a multi-location firm. Multi-location firms are used because they allow the analysis to compare local units operating under the same brand but facing different complaint shocks and response practices. This structure reduces the risk that estimates are driven only by brand-level reputation differences. It also makes it possible to observe whether a brand's standardized response templates help or hurt local recovery.

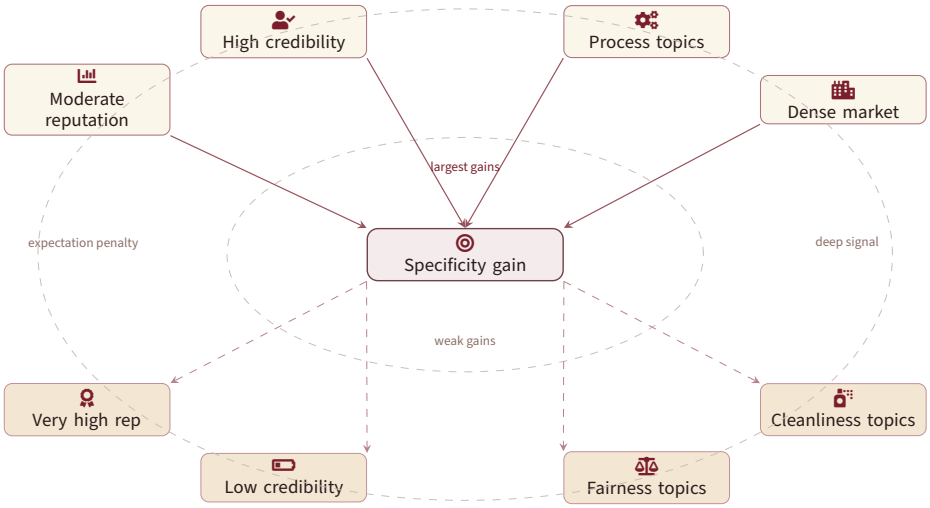


Figure 5. Heterogeneity structure showing where operationally specific responses produce stronger or weaker recovery gains after complaint shocks.

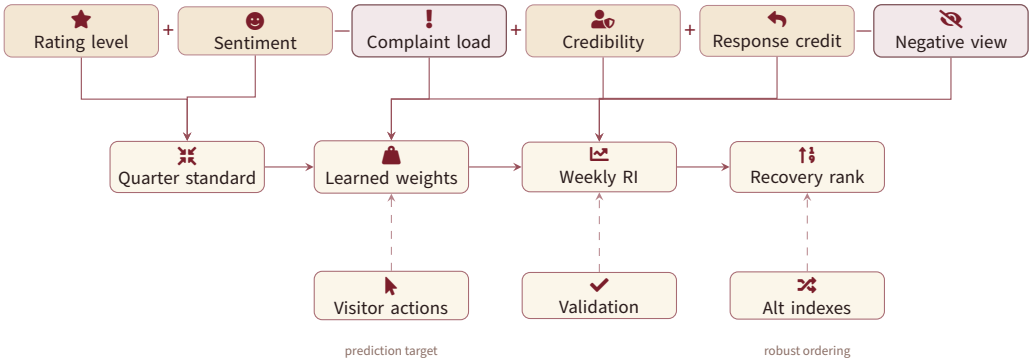


Figure 6. Weekly reputation-index construction combining numeric evaluations, textual residue, reviewer credibility, response credibility, visibility penalties, and validation against downstream visitor actions.

The raw data begin with consumer reviews collected from major review environments and industry-specific platforms. Each review includes a star rating, text, date, platform, reviewer history, helpfulness votes when available, whether the review is verified or transaction-linked, and whether the firm posted a managerial response. Managerial responses include response date, response length, lexical features, specificity measures, and whether the reply appears to follow a standardized template. Platform visibility is measured through weekly review volume, ranking position within local category when available, share of recent negative reviews shown on the first visible page, and the number of older negative reviews that remain highly visible because of helpfulness votes or platform sorting rules.

The final dataset includes 42,318 local units, 626 brands, 19.6 million reviews, and 5.1 million firm responses. Restaurants account for 28.4% of local units, hotels for 16.7%, personal services for 11.2%, auto services for 10.9%, health-related services for 9.5%, specialty retail for 8.7%, home services for 7.4%, repair services for 4.8%, and other recurring local services for 2.4%. The median local unit has 312 reviews over the sample period, while the 90th percentile has 1,184 reviews. The median weekly review count is 5.1, but platform activity is highly skewed. To avoid unstable classification, local units must have at least 80 total reviews and at least 18 months of activity before entering the balanced estimation sample.

The central event is a digital complaint shock. A complaint shock is defined as a week in which the complaint intensity for a local unit exceeds its own trailing 52-week mean by at least 2.5 standard deviations, is above the 95th percentile of the unit's historical complaint distribution, and includes at least eight complaint-bearing reviews or at least 28% complaint-bearing text among all reviews in that week. The rating average must also decline by at least 0.18 stars relative to the prior eight-week average. This definition prevents isolated one-star reviews from being treated as shocks. It identifies concentrated public complaint episodes that are large enough to alter the visible reputation environment.

The sample contains 3,946 complaint shocks after excluding periods with major platform outages, duplicate review campaigns, and shocks that coincide with nationwide news about the entire brand. About 31.6% of shocks involve service delay or appointment reliability, 19.8% involve em-

ployee conduct, 13.2% involve cleanliness or safety, 11.9% involve billing fairness, 9.6% involve product or repair failure, 7.1% involve post-purchase support, and the remaining 6.8% involve mixed complaints. The median shock reduces the weekly reputation index by 0.41 standard deviations and the average rating by 0.37 stars. The median increase in negative-topic visibility is 18.6%, meaning that a substantially larger share of visible recent reviews mentions unresolved problems.

Reputation recovery is defined in two ways. Numeric recovery occurs when the local unit's rolling eight-week rating average returns to at least 95% of its pre-shock baseline. Textual recovery occurs when the complaint-topic share falls to within 10% of its pre-shock baseline and remains there for at least four consecutive weeks. Full reputation recovery requires both numeric and textual recovery. This distinction is empirically important. In the sample, 62.3% of shocks achieve numeric recovery within 16 weeks, but only 44.7% achieve full recovery. Many outlets restore rating averages while the complaint topic remains visible in review text [7].

The weekly reputation index is designed to capture this richer pattern. It combines the rolling rating level, the distribution of review sentiment, reviewer credibility, complaint-topic persistence, response credibility, and negative-review visibility [8]. The index is standardized within industry and calendar quarter. The rating component receives the largest weight, but the textual components prevent rapid recovery from being overstated when new positive reviews merely dilute unresolved complaints. The index is estimated as:

$$RI_{it} = \omega_1 Rating_{it} + \omega_2 Sentiment_{it} - \omega_3 Complaint_{it} + \omega_4 Credibility_{it} + \omega_5 Response_{it} - \omega_6 Visibility_{it}. \quad (1)$$

The weights are obtained from a supervised prediction model in which the target is next-quarter review conversion, measured by the share of platform visitors who click for contact information, booking, directions, or menu details when such platform data are available. For outlets lacking visitor-level platform data, the index uses weights estimated from the subset with complete visibility information and then applies them to the broader sample. The resulting weights are 0.31 for rating level, 0.17 for sentiment, 0.19 for complaint-topic intensity, 0.11 for reviewer credibility, 0.09 for response credibility, and 0.13 for unresolved negative visibility. Equal-weighted and principal-component alternatives produce similar recovery rankings.

Cao and Yan (2016) [9] find that the specificity of nutrition claims can be positively related to firm performance while excessive diversity in claims can be harmful, which is relevant to the present measurement problem because reputation signals may become more valuable when they are precise and less useful when the message environment becomes scattered. The present study applies this logic not to product claims but to post-complaint reputation repair, where response specificity and consistency are empirically separated from broad favorable language.

The dataset also includes local competition and operating-context variables. Competition is measured by the number of comparable local units within a five-mile radius, category density, average competitor rating, and recent competitor review growth. Local demand conditions are proxied by category search intensity, population density, weekday foot-traffic estimates when available, and seasonality indicators. Outlet characteristics include chain affiliation strength, local tenure, prior rating volatility, response-history depth, review-volume trend, and pre-shock reputation percentile. These variables are used to match shocked units with comparable non-shocked units and to estimate recovery heterogeneity.

Managerial response variables are central to the analysis. Response speed is measured as the number of days between the complaint review and the first firm response. Response coverage is the share of complaint reviews that receive a response within seven days. Response specificity is a text-based measure capturing whether the response names the complaint category, identifies a process or responsible area, states an operational correction, and avoids vague assurance language. Boilerplate probability is estimated by comparing a response to the brand's common response templates using cosine similarity in sentence-embedding space. A response with similarity above 0.86 to a recurring template is classified as boilerplate unless it contains complaint-specific details.

The response taxonomy produces meaningful variation. Among complaint shocks, 71.4% receive at least one managerial response within seven days. Only 38.2% receive responses classified as operationally specific. About 46.9% receive predominantly generic apologies. Around 29.7% receive responses with high boilerplate probability. Response speed and response specificity are correlated but distinct. Some firms reply quickly with generic language. Others reply more slowly but provide detailed corrective information. The econometric models include both variables to avoid confusing speed with substance.

The data also track review credibility. A negative review from a reviewer with many prior reviews, platform verification, helpfulness votes, and detailed text may carry more reputational weight than an anonymous short complaint. The credibility-weighted complaint measure assigns larger weights to reviews that are likely to be trusted by readers [10]. Complaint shocks with high credibility-weighted intensity have longer recovery times. This is consistent with the idea that reputation harm depends not only on the number of complaints but also on whether the complaints are viewed as believable by future customers [11].

3. Textual Measurement and Machine Learning Design

The text measurement system has three tasks: identifying complaint-bearing reviews, classifying complaint topics, and coding managerial response specificity. The first task uses a transformer classifier fine-tuned on a manually labeled

training set of 94,000 review sentences. Human coders classified each sentence as neutral evaluation, positive evaluation, mild dissatisfaction, diagnostic complaint, severe complaint, or non-service-related content. The classifier converts review text into sentence-level probabilities and then aggregates them to the review and weekly outlet level. A review is considered complaint-bearing when its diagnostic or severe complaint probability exceeds 0.62.

The second task assigns complaint-bearing text to operational categories. The categories are service delay, appointment reliability, employee conduct, cleanliness, billing fairness, product failure, repair failure, safety concern, core quality, and post-purchase support. The classifier permits multiple labels because a review can mention both delay and employee conduct. The average review with a complaint has 1.42 complaint labels. Topic intensity is measured as the credibility-weighted share of review text assigned to each complaint type in a week. This design captures the content of reputation harm rather than treating all negative reviews as equivalent.

The third task codes managerial response specificity. A response is operationally specific when it meets three conditions. It must acknowledge the concrete failure category rather than merely express regret. It must mention an action, process, personnel step, inspection, refund, retraining, scheduling change, quality check, escalation, or other corrective mechanism. It must avoid exclusive reliance on generic assurance, such as statements that the experience does not reflect the firm's standards. Specificity is therefore not a measure of politeness or length [12]. A short response can be specific, and a long response can be generic.

Classifier performance is evaluated on holdout sets not used during training. The complaint detection model achieves an F1 score of 0.884 and an area under the receiver operating curve of 0.931. Complaint-topic classification has a macro F1 score of 0.812. Topic performance is strongest for cleanliness, billing fairness, and appointment reliability, and weakest for employee conduct because conduct language often overlaps with service tone. The response-specificity classifier has an F1 score of 0.846 and a calibration slope of 0.97. The boilerplate classifier has an F1 score of 0.879 using manually reviewed response pairs.

The weekly complaint intensity measure is:

$$
\begin{aligned}
CI_{itk} &= \frac{\sum_{r \in R_{it}} q_r p_{rk}}{\sum_{r \in R_{it}} q_r} \\
CI_{it} &= \sum_{k=1}^{K} \pi_k CI_{itk}, \\
Shock_{it} &= 1[CI_{it} > \mu_{i,t-52:t-1} \\
&\quad + 2.5\sigma_{i,t-52:t-1}].
\end{aligned} \tag{2}
$$

Here, q_r is reviewer credibility, p_{rk} is the probability that review r belongs to complaint category k, and π_k is the severity weight for complaint category k. Fairness, cleanliness, and safety categories receive larger severity weights because preliminary recovery models show that these topics have stronger persistence and larger effects on reputation index decline. The shock indicator also requires the percentile and rating-drop thresholds described earlier.

The response specificity score is:

$$
\begin{aligned}
SPEC_{it} &= a_1 FailureName_{it} + a_2 Action_{it} \\
&\quad + a_3 Verification_{it} - a_4 Generic_{it} \\
&\quad - a_5 Template_{it}.
\end{aligned} \tag{3}
$$

The score is standardized within brand and year. This standardization is important because some brands generally write longer responses than others. A brand with a verbose response culture should not automatically receive higher specificity scores if the additional words are formulaic. The standardized measure captures whether a local unit's response is more specific than expected relative to its brand's own response style.

Cao and Xu (2022) [13] document that managerial characteristics can influence brand acquisition and disposal decisions and that valuation consequences may depend on how brand assets are interpreted, which motivates the present analysis to account for managerial response discretion rather than treating reputation repair as an automatic brand-level process. Here, discretion appears in how local managers respond to complaints, whether they use diagnostic language, and whether they provide credible evidence of correction.

Machine learning is also used for prediction, not only measurement. A recovery prediction model estimates whether a complaint shock will fully recover within 12 weeks. The model includes pre-shock reputation level, rating volatility, review volume, complaint-topic mix, response speed, response specificity, boilerplate probability, local competition, platform visibility, reviewer credibility, and brand fixed-effect embeddings. The model is trained on shocks from 2016 through 2021, validated on 2022, and tested on 2023 through 2024. Gradient-boosted trees, random forests, elastic net models, and neural additive models are compared.

The best predictive model is a gradient-boosted tree with monotonicity constraints on complaint severity and boilerplate probability. It achieves an out-of-sample area under the curve of 0.807 for predicting full recovery within 12 weeks. A model using only rating decline and pre-shock rating achieves 0.642. Adding complaint topics raises the score to 0.731. Adding response features raises it to 0.789. Adding local competition and visibility raises it to 0.807 [14]. This pattern indicates that textual complaint content and response behavior contain substantial information beyond numeric ratings.

Feature attribution shows that the most important predictors of full recovery are complaint-topic severity, response specificity, pre-shock rating volatility, unresolved negative visibility, reviewer credibility, boilerplate probability, and local competitor rating. Response speed is useful but less important than specificity after the first three days. The marginal value of responding within 24 hours is high only when the response is specific. Fast generic responses

often have weak or negative predicted effects because they appear to leave the substantive complaint unanswered.

The machine-learning prediction exercise is not used as causal evidence by itself. It helps identify nonlinearities and supports the selection of interactions for econometric models. For example, partial-dependence functions show that response specificity has little effect below the 30th percentile of complaint severity, increases sharply between the 40th and 80th percentiles, and flattens for extreme shocks. This suggests that specific responses are most useful when the shock is visible but not catastrophic. For extremely severe cleanliness or safety shocks, response language alone is insufficient to restore reputation quickly.

The text measures are audited for bias across industries and platforms. Classifier thresholds are recalibrated by industry so that complaint detection does not overstate dissatisfaction in categories where consumers naturally use more intense language. Platform fixed effects are included because some platforms encourage shorter reviews, while others produce longer narratives [15]. Sensitivity tests remove platforms with unusual moderation policies and retain only verified reviews. The main results remain similar, although the estimates are slightly larger in verified-review subsamples because credible complaints carry more reputational weight.

The final measurement issue is the possibility of fake or strategic reviews. The study removes review bursts from newly created reviewer accounts with repeated lexical patterns, extreme timing concentration, or high cross-outlet duplication [16]. It also flags suspicious positive review bursts following complaint shocks. The main recovery models are estimated both with and without flagged reviews. The response-specificity effect is slightly stronger when suspicious reviews are removed, suggesting that reputation recovery attributed to operationally specific responses is not explained by artificial positive-review dilution [17].

4. Econometric Strategy

The econometric design combines matched counterfactual analysis, duration modeling, and orthogonalized treatment estimation. The primary empirical challenge is that firms do not randomly receive complaint shocks, and they do not randomly choose response strategies [18]. Poorly managed outlets may have both more severe complaints and weaker responses. Highly disciplined outlets may write better responses and recover more quickly because they also correct operations more effectively [19]. The analysis therefore uses several complementary approaches rather than relying on a single regression.

The first approach constructs matched non-shocked counterfactuals. For each shocked local unit, the algorithm identifies up to five comparison units from the same industry, brand when possible, region type, and pre-shock reputation decile. Matching variables include pre-shock rating, review volume, complaint intensity, rating volatility, platform visibility, response history, local competition, and seasonality. The comparison unit must not experience a complaint shock within 12 weeks of the treated unit's shock date. The matched design estimates the reputational decline and recovery path relative to comparable units that did not experience a shock at the same time.

The matched difference model is:

$$RI_{it} = \alpha_i + \lambda_t + \sum_{\ell=-8}^{24} \beta_\ell D_{i,t+\ell} + \Gamma' X_{it} + \varepsilon_{it}. \quad (4)$$

In this equation, RI_{it} is the weekly reputation index, $D_{i,t+\ell}$ identifies event time relative to the complaint shock, and X_{it} contains time-varying controls. The coefficients β_ℓ trace the reputation path from eight weeks before to 24 weeks after the shock. Parallel pre-trends are evaluated using the pre-shock coefficients. The main event-study plots show no significant pre-trend in the eight weeks before shock identification, but reputation falls sharply in the shock week and gradually recovers for some units.

The second approach estimates recovery duration. The dependent variable is the number of weeks until full reputation recovery. Shocks that do not recover within 32 weeks are right-censored. The baseline duration model uses an accelerated failure-time specification with a log-logistic distribution, allowing the effect of response specificity to be interpreted as a change in recovery time. The model is:

$$\log(T_i) = \alpha + \delta SPEC_i + \theta BOIL_i + \rho SPEED_i + \Phi' Z_i + \xi_i. \quad (5)$$

Here, T_i is weeks to full recovery after shock i, $SPEC_i$ is response specificity during the first seven days, $BOIL_i$ is boilerplate probability, $SPEED_i$ is response speed, and Z_i contains shock severity, topic composition, prior reputation, competition, review credibility, and brand fixed effects. A negative δ indicates faster recovery. The preferred model also includes interactions between response specificity and complaint topic.

The third approach uses synthetic controls at the local-unit level. For each shock, a synthetic reputation trajectory is built from a weighted combination of non-shocked units with similar pre-shock reputation dynamics. The estimator compares the actual post-shock reputation path with the synthetic path. This analysis is used mainly to validate the size of the reputational injury and the recovery patterns across response strategies. It is less dependent on linear controls but requires sufficient pre-shock observations, so it is estimated on 2,781 shocks with at least 52 stable pre-shock weeks.

The fourth approach estimates the treatment effect of operationally specific response using double machine learning. The treatment is whether the local unit provides a high-specificity response package within seven days of the shock. A high-specificity package requires that at least 60% of complaint-bearing reviews receive a response and that the average response specificity score is above the sample

median. The outcome is full recovery within 12 weeks and, separately, change in the reputation index at week 12. Flexible models predict both treatment and outcome from pre-treatment variables. Orthogonalized residuals are then used to estimate the treatment effect.

The double machine-learning score is:

$$\begin{aligned} \widetilde{Y}_i &= Y_i - \widehat{m}(W_i), \\ \widetilde{D}_i &= D_i - \widehat{g}(W_i), \\ \widehat{\tau} &= \left(\sum_i \widetilde{D}_i^2 \right)^{-1} \\ &\quad \sum_i \widetilde{D}_i \widetilde{Y}_i. \end{aligned} \tag{6}$$

The vector W_i includes only variables observed before or during the shock week but before the response is posted. This timing restriction prevents the model from conditioning on variables that may themselves be affected by response strategy. Cross-fitting is conducted by brand to avoid training on one outlet of a brand and testing on another outlet of the same brand in the same fold. This is a conservative design because brands often use common response systems.

The fifth approach estimates heterogeneous treatment effects with causal forests. The treatment remains the high-specificity response package, and outcomes are full recovery within 12 weeks and reputation-index recovery at week 12 [20]. Candidate moderators are complaint topic, pre-shock reputation percentile, shock severity, reviewer credibility, local competition, platform visibility, outlet tenure, response-history depth, prior rating volatility, and brand template reliance. The causal forest is interpreted through grouped average effects and not through single-tree rules.

The heterogeneity model can be written as:

$$\begin{aligned} Y_i &= \mu(W_i) + \tau(W_i) D_i + \epsilon_i, \\ \tau(W_i) &= E[Y_i(1) - Y_i(0) \mid W_i]. \end{aligned} \tag{7}$$

This model allows the recovery effect of specific response behavior to vary across outlet and shock conditions. It is useful because a single average treatment effect could hide important managerial guidance. For example, a corrective response may be highly effective for service-delay complaints but less effective for fairness complaints if consumers suspect deliberate misconduct. The heterogeneous model estimates these differences directly.

Several identification safeguards are used. Response variables are measured within seven days after the shock to reduce reverse causality from later recovery to response strategy. Models control for pre-shock response culture so that firms that always respond extensively are compared with their own expected behavior. Brand fixed effects absorb centralized reputation-management systems. Industry-by-calendar-month fixed effects absorb seasonal demand and category-specific shocks. The analysis also estimates specifications restricted to outlets within the same brand, comparing local units that face different shocks and write different responses under the same corporate identity.

The study avoids interpreting the estimates as randomized treatment effects. Operationally specific responses may accompany real operational improvements that are partly unobserved. This is not a defect for the managerial question because reputation recovery likely depends on both communication and correction. The estimates identify the combined effect of credible, specific response behavior under rich controls, not the isolated rhetorical effect of words without operational change. The interpretation is that diagnostic response packages are associated with faster recovery because they signal and possibly implement correction.

5. Results

The first empirical result is that complaint shocks produce large and persistent reputation declines. In the matched event-study model, the average shock reduces the weekly reputation index by 0.44 standard deviations in the shock week. The index remains 0.31 standard deviations below baseline four weeks later and 0.18 standard deviations below baseline twelve weeks later. Numeric ratings recover faster than textual complaint measures. The average rating gap narrows by 64.1% within twelve weeks, while complaint-topic intensity narrows by only 38.7%. This difference confirms that rating recovery alone understates reputational residue [21].

The event-study estimates show little evidence of pre-shock decline. The coefficients for weeks minus eight through minus two are small and statistically insignificant. Week minus one shows a slight decline for service-delay shocks, suggesting that some complaint buildup begins just before the shock threshold is crossed, but the main discontinuity occurs in the shock week. This supports the shock definition as an abrupt public deterioration rather than the final stage of a long visible decline.

Recovery differs sharply by complaint topic. Service-delay and appointment-reliability shocks recover fastest, with median full recovery times of 10.8 and 11.6 weeks. Product or repair failure shocks have a median recovery time of 14.2 weeks. Employee-conduct shocks recover in 16.7 weeks. Billing-fairness shocks recover in 19.4 weeks. Cleanliness and safety shocks have the slowest recovery, with a median of 23.1 weeks and a high censoring rate at 32 weeks. These differences remain after controlling for shock severity, review volume, and prior rating. Consumers appear to treat fairness, cleanliness, and safety complaints as deeper indicators of reputation risk [22].

Response behavior is uneven across complaint topics [23]. Firms respond fastest to safety and cleanliness complaints, with a median first response time of 1.9 days, but those responses are also more likely to be generic. Firms respond slowest to billing-fairness complaints, with a median of 3.8 days. Operational specificity is highest for appointment reliability and repair failure, where firms can describe scheduling changes, technician follow-up, or process cor-

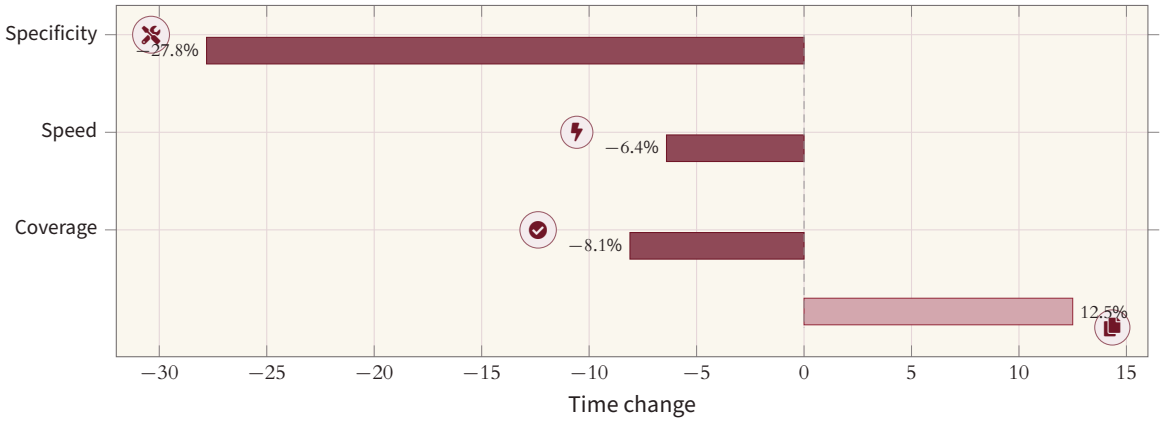


Figure 7. Accelerated failure-time estimates show that response specificity is the dominant recovery accelerator, reducing median recovery time by 27.8%, while speed and coverage provide smaller gains and boilerplate language delays recovery.

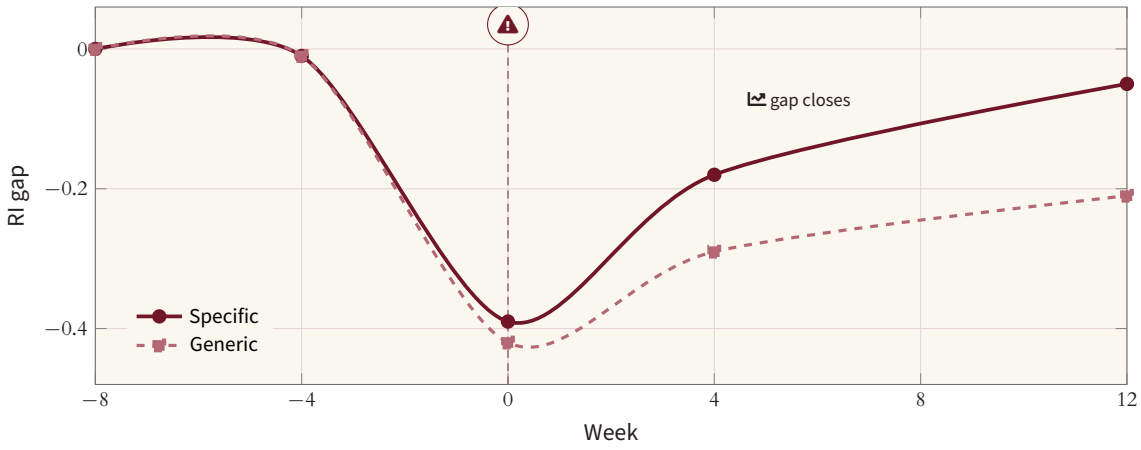


Figure 8. Synthetic-control trajectories show that high-specificity responses nearly close the post-shock reputation-index gap by week twelve, whereas generic responses leave a persistent deficit.

rection. Specificity is lowest for employee conduct, where responses often avoid naming the issue and instead invite the customer to contact management offline.

The baseline duration model shows that operationally specific response packages shorten recovery time. The coefficient on response specificity in the accelerated failure-time model is minus 0.326, implying a 27.8% reduction in median recovery time. For a shock with a predicted recovery time of 16 weeks, this corresponds to recovery in approximately 11.6 weeks. Response speed has a smaller effect. Responding within 24 hours reduces median recovery time by 6.4% when specificity is held constant. Response coverage reduces recovery time by 8.1%. Boilerplate probability increases median recovery time by 12.5%.

Apology frequency alone is not a reliable predictor of recovery. In models without specificity controls, apology language is weakly associated with faster recovery. Once specificity, boilerplate probability, and complaint topic are included, the apology coefficient becomes statistically insignificant. This does not mean that apologies are harmful. It means that apology language is not sufficient to explain recovery [24]. The beneficial component appears to be diagnostic acknowledgment and correction, not politeness by itself [25].

The double machine-learning estimate for high-specificity response packages is positive and significant. The estimated effect on full recovery within 12 weeks is 0.084, meaning that high-specificity response packages increase the probability of full recovery by 8.4 percentage points relative to comparable shocks. The estimated effect on the week-12 reputation index is 0.137 standard deviations. When the treatment definition is tightened to require response within three days, the recovery probability effect is 0.096. When the treatment is defined as high response coverage without specificity, the effect falls to 0.021 and is not statistically significant.

The matched synthetic-control analysis produces similar patterns. Shocks receiving high-specificity response packages have an average reputation-index gap of minus 0.39 standard deviations in the shock week, minus 0.18 at week four, and minus 0.05 at week twelve. Shocks receiving mostly generic responses have a gap of minus 0.42 in the shock week, minus 0.29 at week four, and minus 0.21 at week twelve. The difference in week-12 recovery

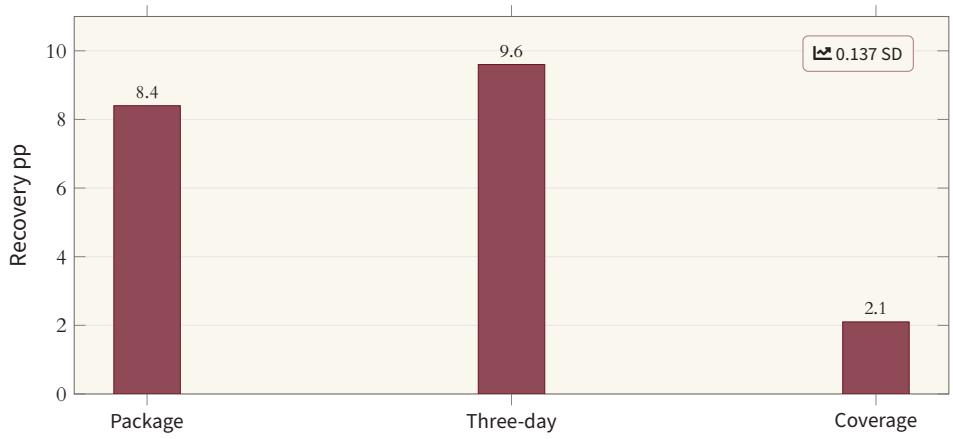


Figure 9. Double machine-learning estimates indicate that high-specificity response packages increase full recovery within twelve weeks by 8.4 percentage points, rising to 9.6 points when timeliness is required, while coverage alone is weak.

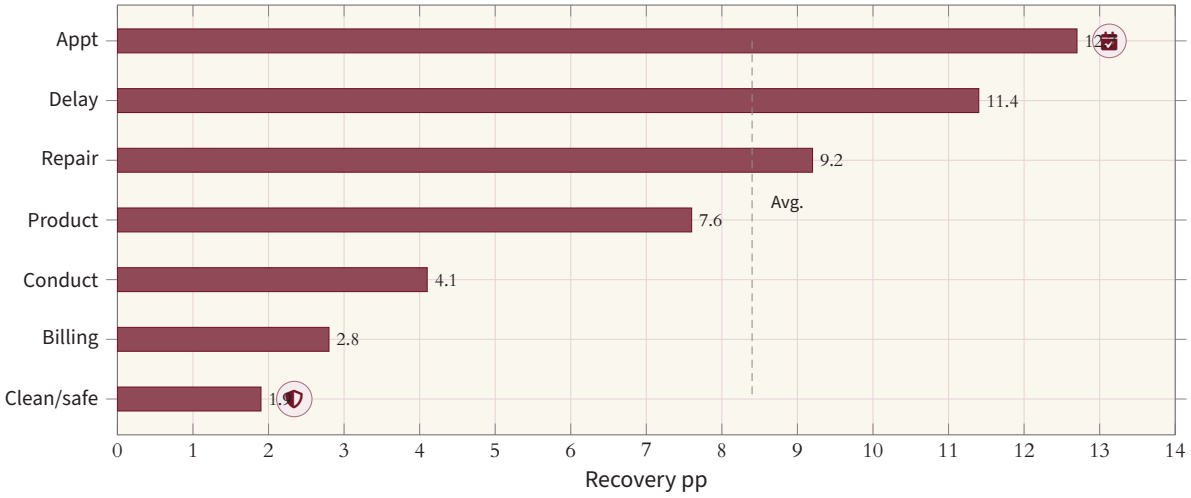


Figure 10. Topic-specific estimates show that specific responses are most effective for fixable process complaints and least effective for cleanliness, safety, and fairness-related shocks.

between the two groups is 0.16 standard deviations. The gap is largest for service-delay and appointment-reliability complaints and smallest for cleanliness complaints.

The topic-specific response effects are informative. High-specificity responses increase the probability of 12-week full recovery by 12.7 percentage points for appointment reliability, 11.4 points for service delay, 9.2 points for repair failure, and 7.6 points for product failure. The effect is 4.1 points for employee conduct, 2.8 points for billing fairness, and 1.9 points for cleanliness or safety. This gradient suggests that consumers are more willing to update reputation beliefs when the complaint involves a fixable process failure. They are less willing to update when the complaint suggests moral, hygienic, or safety-related weakness.

Pre-shock reputation has a nonlinear relationship with recovery. Very low-reputation outlets recover slowly because new responses lack credibility and baseline complaint intensity is already high. Very high-reputation outlets also recover slowly from severe shocks because the complaints violate prior expectations and remain highly visible. Moderate-reputation outlets recover fastest when they provide specific responses. The causal forest confirms this pattern. The estimated treatment effect of high-specificity response is 0.122 for outlets between the 45th and 75th pre-shock reputation percentiles, 0.057 for outlets below the 25th percentile, and 0.061 for outlets above the 90th percentile.

Reviewer credibility amplifies reputational injury [26]. Shocks dominated by high-credibility reviewers produce larger declines and slower recovery. A one-standard-deviation increase in credibility-weighted complaint intensity increases expected recovery time by 18.3%. However, high-specificity responses are also more valuable when reviewer credibility is high, provided the complaint topic is process-related. This suggests that credible complaints make the failure harder to dismiss but also make a credible correction more meaningful. When high-credibility complaints involve cleanliness or fairness, response specificity has limited

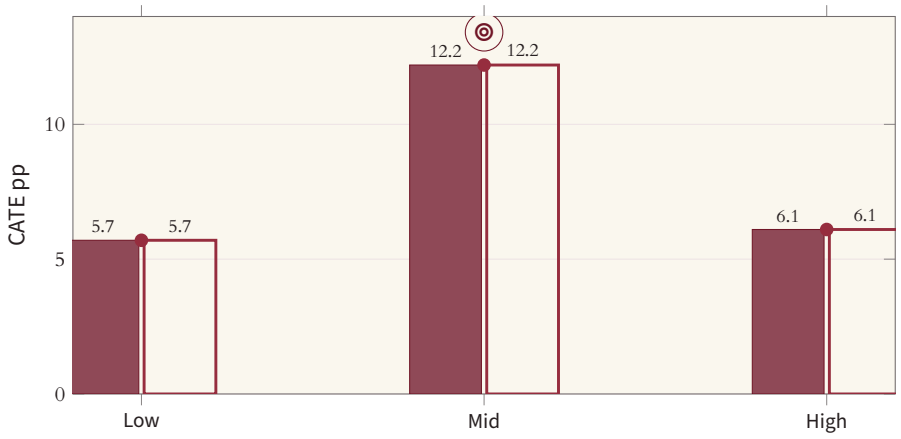


Figure 11. Causal-forest heterogeneity indicates that the recovery benefit of specific responses peaks for outlets with moderate pre-shock reputation and weakens at both low and very high reputation levels.

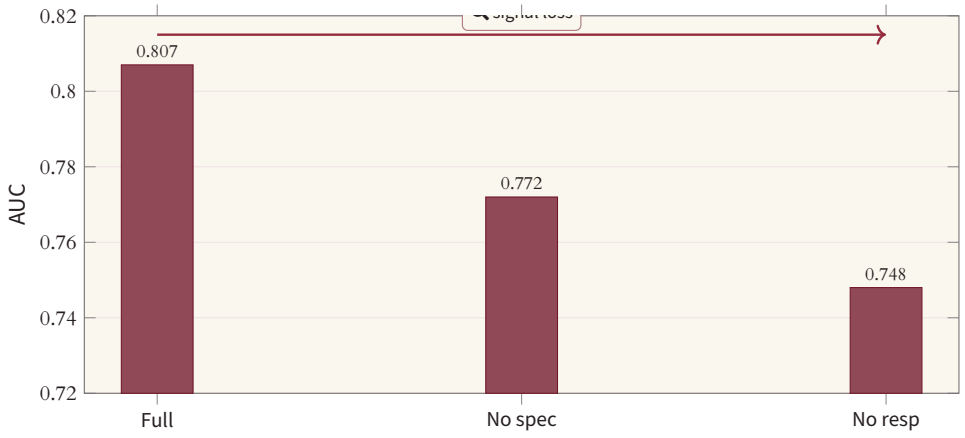


Figure 12. Predictive ablations show that removing response specificity lowers twelve-week recovery prediction from 0.807 to 0.772 AUC, and removing all response variables lowers it further to 0.748.

effect.

Local competition affects recovery [27]. In dense markets with many highly rated competitors, complaint shocks cause larger reputation-index declines and slower numeric recovery. The effect is strongest for restaurants, hotels, and personal services. A high-specificity response package partly offsets this disadvantage but does not eliminate it. In low-competition areas, recovery is more forgiving, possibly because consumers have fewer substitutes and review attention decays more slowly. The machine-learning model ranks competitor rating percentile as the seventh most important predictor of recovery within 12 weeks.

Visibility of unresolved negative reviews is one of the strongest predictors of persistence. If complaint-bearing reviews remain on the first visible review page for at least four weeks, full recovery probability falls by 13.8 percentage points. Managerial responses reduce this penalty only when they are attached directly to the visible negative reviews and contain specific corrective information. Responses posted to later reviews do not offset the reputational weight of prominent unresolved complaints [28]. This result highlights that reputation repair is partly a visibility problem, not only a sentiment problem.

Boilerplate responses are associated with worse recovery. In the full duration model, high boilerplate probability increases median recovery time by 12.5%. In the causal forest, boilerplate reliance appears most harmful for fairness and employee-conduct complaints, where consumers likely expect individualized acknowledgment. Boilerplate responses are less harmful for simple service-delay complaints, but they still provide no measurable advantage [29]. The finding is consistent with the idea that formulaic communication can signal low attention to the complaint and may extend rather than repair reputational harm.

The analysis also examines whether firms attempt to dilute shocks with positive-review bursts. Positive-review volume increases in the four weeks following 18.9% of shocks. Some of this increase is natural recovery, but a subset shows suspicious timing and lexical similarity. Removing flagged positive bursts strengthens the estimated effect of specific responses and weakens the apparent recovery of generic-response firms. This suggests that some numeric

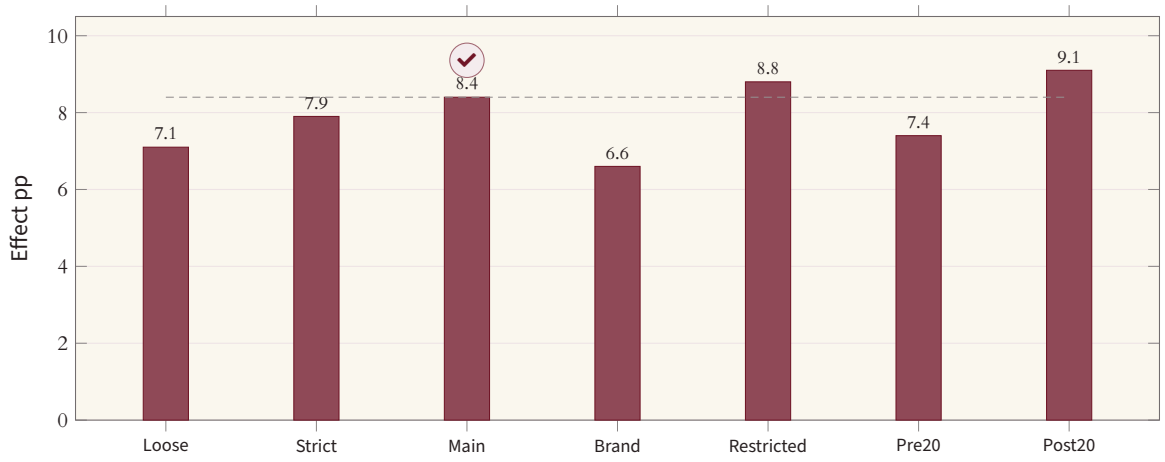


Figure 13. Robustness estimates remain positive across alternative shock thresholds, within-brand comparisons, restricted outlet samples, and pre- versus post-2020 subsamples.

recovery is achieved through review dilution while textual reputation remains impaired. The full recovery measure is therefore more conservative than the rating-only measure [30].

The predictive models reinforce the econometric findings. The gradient-boosted model predicts 12-week full recovery with an area under the curve of 0.807. Response specificity, complaint-topic severity, unresolved negative visibility, reviewer credibility, and pre-shock rating volatility are the strongest predictors. Response speed ranks below these variables. When response specificity is removed from the model, predictive performance falls to 0.772. When all response variables are removed, it falls to 0.748. These changes show that response behavior contains information not captured by shock severity or pre-shock reputation [31].

Causal-forest estimates identify four high-gain conditions for specific responses. The first is a process complaint, such as scheduling, delay, repair, or service sequencing. The second is moderate pre-shock reputation, where the firm has enough credibility to be believed but not such a strong reputation that the shock becomes a major expectation violation. The third is high review credibility, which makes correction more informative. The fourth is moderate local competition, where customers have alternatives but the outlet is not immediately dominated by superior nearby options. Under these conditions, high-specificity responses increase 12-week full recovery probability by 14.8 to 18.6 percentage points [32].

The low-gain conditions are also clear. Specific responses have limited estimated effects when the complaint involves cleanliness, safety, or fairness, when unresolved negative reviews remain highly visible, when pre-shock reputation is already weak, and when the response is embedded in a highly templated brand communication system. Under these conditions, public response language may not be enough [33]. Reputation recovery likely requires observable operational evidence, third-party verification, or a longer sequence of positive experiences.

6. Robustness and Additional Analyses

The first robustness test changes the complaint-shock definition. A stricter threshold requiring three standard deviations above the trailing complaint mean produces 2,714 shocks. The response-specificity effect on 12-week full recovery is 0.079, close to the main estimate of 0.084. A looser threshold requiring two standard deviations produces 5,382 shocks and an effect of 0.071. The smaller estimate in the looser sample is expected because milder shocks have less room for observable recovery [34]. The direction and significance of the main finding remain stable.

The second robustness test uses rating-only shocks instead of text-based complaint shocks. Rating-only shocks are defined as abrupt declines in average star rating without requiring diagnostic complaint language [35]. In this specification, the effect of response specificity is weaker and less stable. Many rating-only shocks reflect isolated dissatisfaction, low review counts, or non-operational preferences. The weaker result supports the paper's claim that reputation repair should be studied through complaint content, not only through numeric rating movements.

The third test changes the recovery definition. Numeric recovery alone yields a larger apparent recovery rate and a smaller response-specificity effect. Textual recovery alone yields a stronger response-specificity effect. Full recovery, requiring both numeric and textual repair, provides the most conservative estimate [36]. Under a 16-week recovery horizon, high-specificity responses increase numeric recovery probability by 5.6 percentage points, textual recovery probability by 10.9 points, and full recovery probability by 8.7 points. This pattern indicates that specific responses are especially relevant to resolving complaint-topic persistence.

The fourth test restricts the sample to within-brand comparisons. This specification compares outlets belonging to the same brand and observed in the same year. Brand-level reputation systems, national advertising, and standardized

policies are absorbed [37]. The response-specificity effect remains positive at 0.066 for 12-week full recovery. The estimate is smaller than the full-sample estimate because within-brand variation is narrower, but it remains statistically significant. This result suggests that local response discretion matters even under shared corporate identity.

The fifth test removes all shocks involving possible operational closure, ownership transfer, renovation, or major staffing change. These events may create both complaints and recovery unrelated to response behavior. The restricted sample contains 3,218 shocks. The high-specificity response effect is 0.088, nearly identical to the main estimate. This reduces concern that the results are driven by structural outlet changes.

The sixth test estimates separate models by industry. The response-specificity effect is largest in hotels, auto services, repair services, and home services. It is moderate in restaurants and personal services. It is weakest in health-related services, where privacy constraints may limit detailed public responses and consumers may interpret complaints differently. In health-related services, response speed and offline escalation language are more predictive than public operational detail. This industry difference is sensible because public responses to health complaints face confidentiality and professionalism constraints.

The seventh test examines platform differences. The main results hold on general review platforms, industry-specific platforms, and transaction-verified platforms. The response-specificity effect is largest on transaction-verified platforms, where complaints are more credible and readers may place greater weight on managerial responses. On platforms with shorter review text, the complaint-topic classifier is less precise, and the effect is smaller. The direction remains consistent.

The eighth test adjusts for managerial response selection using inverse probability weights. The propensity to provide a high-specificity response is estimated from pre-shock variables, shock severity, response-history depth, brand, industry, and local context. Weighted duration models yield a response-specificity time ratio of 0.741, close to the unweighted time ratio of 0.722. Trimming observations with extreme propensity scores produces similar estimates. This suggests that observable selection into response specificity does not explain the main result.

The ninth test uses an instrumental timing variable based on weekend and holiday response-capacity constraints. Some shocks occur immediately before weekends or holidays, when local managerial response capacity is lower. This timing affects response speed and, to a weaker degree, response specificity, but should not directly affect long-run reputation recovery except through response behavior after controlling for seasonality and demand. The timing instrument is not strong enough to support all specifications, so it is used only as supplementary evidence. The two-stage estimates are directionally consistent but imprecise. The study therefore does not rely on this instrument for its main claims.

The tenth test examines whether response specificity predicts future complaint recurrence. If specific responses reflect real correction, they should be associated with lower probability of a second shock in the same complaint category. The estimates support this expectation. High-specificity responses reduce the probability of same-topic recurrence within 24 weeks by 6.9 percentage points. The reduction is largest for appointment reliability and service delay. There is no significant reduction for cleanliness or fairness complaints. This result is consistent with the earlier recovery findings and suggests that specific responses may be linked to actual process correction in some categories.

The eleventh test separates local-manager responses from centralized brand responses. Local-manager responses are identified through location-specific language, manager names, local contact details, and references to outlet-specific actions. Centralized responses are identified through template similarity and national customer-service routing. Local specific responses are associated with faster recovery than centralized generic responses [38]. Centralized responses can still be effective when they include a clear case number, escalation channel, and corrective description. The distinction is therefore not local versus central by itself, but whether the response appears diagnostically connected to the complaint.

The twelfth test considers review helpfulness voting. Negative reviews that receive high helpfulness votes can remain visible longer and may shape reputation more strongly. When helpfulness-weighted visibility is included, the direct effect of shock severity declines, but response specificity remains significant. Specific responses attached to high-helpfulness negative reviews produce larger recovery gains than responses attached only to low-visibility reviews. This finding suggests that firms should prioritize visible diagnostic complaints rather than responding mechanically to all reviews with equal attention.

The thirteenth test examines the language of correction. Among specific responses, the most effective elements are process correction, named follow-up, refund or repair completion, staffing or training change, and inspection or quality-control language. The least effective elements are invitations to discuss offline without any public diagnostic acknowledgment, broad statements about values, and claims that the event was unusual without supporting detail. The results do not imply that offline discussion is useless. Rather, public reputation recovery requires enough visible information for future readers to infer that the complaint was taken seriously.

The fourteenth test studies recovery completeness at 24 weeks. Some shocks not recovered at 12 weeks eventually recover. High-specificity responses still predict full recovery by 24 weeks, but the effect declines from 8.4 percentage points at 12 weeks to 6.1 points at 24 weeks. This decline suggests that some outlets recover slowly regardless of response strategy as new reviews accumulate. The short-run advantage of specific response is nevertheless important because digital reputation affects customer choice during the

recovery period.

The fifteenth test estimates the models using only pre-2020 and post-2020 samples. Reputation dynamics changed after 2020 as consumer expectations, staffing conditions, and platform use shifted. The response-specificity effect is 0.074 before 2020 and 0.091 after 2020. Complaint shocks involving service delay and appointment reliability became more common after 2020, and consumers appeared somewhat more tolerant of delay when responses explained staffing or scheduling constraints. Cleanliness and safety complaints remained highly persistent in both periods.

The final robustness test evaluates whether the findings are driven by response length. Response length is positively correlated with specificity, but length alone does not explain recovery. When both length and specificity are included, specificity remains significant and length becomes weak or insignificant. Very long responses with low specificity are sometimes associated with slower recovery, especially when they appear defensive. This confirms that the beneficial response attribute is diagnostic content rather than verbosity.

7. Discussion

The results show that digital reputation recovery is a process of credible updating. Consumers and platform readers do not observe all operational changes inside a firm. They observe complaints, ratings, response language, and subsequent reviews. A specific managerial response can help repair reputation when it provides a plausible bridge between the complaint and a corrective action. A generic response leaves the reader with the original complaint and little reason to revise beliefs. Reputation therefore recovers through a sequence of signals, not through a single apology.

The study also shows that complaint categories differ in reputational meaning. A late appointment, delayed order, or repair mistake can be interpreted as a fixable process failure. A fairness complaint, cleanliness problem, or safety concern can be interpreted as evidence of deeper values, controls, or competence [39]. This difference explains why the same response strategy has different effects across topics [40]. Reputation management should therefore begin with diagnosis. Treating all negative reviews as generic dissatisfaction leads to generic responses, and generic responses do not resolve topic-specific uncertainty.

Murshed et al. (2023) [41] show that ethical corporate social responsibility and interactional justice can influence employee identification and job satisfaction through perceived fairness and organizational connection, which is relevant here because some reputational failures, especially fairness and employee-conduct complaints, appear harder to repair when public responses do not signal respectful and procedurally credible treatment. The present study observes this pattern from the consumer-facing side of reputation rather than from internal attitude data [42].

The evidence has implications for how firms design response systems. Many brands centralize review management to ensure consistency and reduce legal risk. Consistency can prevent inappropriate responses, but excessive templating may damage recovery when it makes replies appear detached from the complaint. A better system would combine central guidance with local diagnostic fields. Managers could be required to identify the failure category, describe the corrective step, and state the follow-up mechanism, while still avoiding sensitive information. This would preserve control without producing visibly formulaic responses [43].

The findings also suggest that reputation dashboards should move beyond average rating. Average rating is slow to detect topic-specific injury and can recover while complaint residue persists [44]. A dashboard that tracks complaint-topic intensity, visibility of unresolved negative reviews, response specificity, and same-topic recurrence would provide a more accurate representation of reputation health. Such a dashboard would help firms identify whether they are recovering because the underlying issue has been resolved or because positive reviews have diluted older complaints.

For local managers, the results imply that speed is useful but incomplete. A fast response can prevent a complaint from appearing ignored, but speed does not substitute for substance. The strongest recovery patterns appear when responses are timely enough to remain attached to visible complaints and specific enough to address the diagnostic content. Managers should therefore prioritize high-credibility, high-visibility complaints and provide concrete information about correction. Mechanical response quotas may encourage low-quality replies that do not improve reputation.

For multi-location brands, the results imply that reputation recovery capability varies across outlets even within the same brand. Some local units recover quickly because they combine operational correction with credible response practices. Others remain impaired because they rely on template language or fail to address the visible complaint topic. Brand-level reputation systems should identify local units with repeated same-topic shocks and distinguish between communication failure and operational recurrence. A unit that repeatedly receives appointment complaints needs scheduling correction, not only better review replies.

The study also contributes to reputation measurement. Reputation is often treated as a stock of accumulated favorability. In digital platforms, it behaves more like a visible memory system with uneven decay. Some negative reviews disappear into the archive quickly. Others remain visible because they are recent, helpful, credible, or highly diagnostic. Recovery therefore depends on how the platform presents memory to future customers [45]. The same operational improvement can have different reputational effects depending on whether old complaints remain prominent.

The machine-learning results are useful because they identify interactions that conventional linear models may miss. The effect of response specificity is not constant. It is strongest for process complaints, moderate prior reputation,

credible reviewers, and visible but not catastrophic shocks. It is weakest for moralized or hygiene-related complaints. These patterns show that reputation repair is conditional and should not be managed with a single response rule. Causal forests do not replace managerial judgment, but they help identify where specific response behavior is likely to matter most.

The findings should not be interpreted as evidence that communication alone repairs reputation [46]. The response-specificity measure likely captures both public explanation and underlying operational correction. A firm able to describe a concrete fix may be one that actually implemented a fix. This is not a weakness for practical interpretation. Reputation recovery requires both correction and credible communication about correction. A public statement without real change risks recurrence, while a real change that remains invisible may not fully repair public belief.

The results also raise ethical considerations. Firms should not use specificity as a rhetorical device to create false impressions of correction [47]. Because response specificity is associated with recovery, firms may be tempted to write detailed but inaccurate responses. Such behavior could produce short-run recovery but increase long-run reputational risk if complaints recur or if customers discover inconsistency [48]. The recurrence tests in this study suggest that specific responses are most beneficial when they are associated with lower same-topic recurrence. Durable reputation repair therefore depends on truthful operational alignment.

The paper has several limitations. The data are observational, and response specificity is not randomly assigned. The models control for many pre-shock and shock characteristics, use within-brand comparisons, and estimate orthogonalized effects, but unobserved operational correction may remain. The estimates should therefore be interpreted as the effect of a credible response package rather than the isolated causal effect of wording. Another limitation is that platform algorithms are only partially observed. Visibility measures approximate what users see but cannot fully reproduce personalized ranking [49]. A third limitation is that the sample focuses on local service settings, so results may differ in business-to-business markets or durable-goods categories with longer purchase cycles.

Future research can extend the analysis in several directions. One extension would connect review-based recovery to transaction data, allowing researchers to estimate how textual recovery translates into revenue and retention. Another extension would examine employee-side recovery after public complaints, especially when complaints name staff behavior or working conditions. A third extension would study platform design, including whether response placement, review sorting, and verification labels change the speed of reputation repair. A fourth extension would test field interventions that randomize response guidance across outlets, providing sharper causal evidence on response specificity.

8. Conclusion

This paper examined reputation recovery after digital complaint shocks using a large weekly panel of local service units, review text, managerial responses, platform visibility measures, and machine-learning classification. The empirical evidence indicates that reputation damage is not fully captured by rating declines. Complaint-topic persistence, reviewer credibility, unresolved negative visibility, and response quality all shape whether a shock becomes temporary or durable. Numeric ratings often recover before the textual memory of the complaint disappears.

The main finding is that operationally specific response behavior is associated with faster and more complete reputation recovery. Responses that identify the failure mode, describe a corrective action, and avoid generic assurance language shorten recovery time and increase the probability of full recovery within 12 weeks. Apologies and fast replies are not enough when they lack diagnostic content. Boilerplate responses are associated with slower recovery, especially when complaints concern fairness, employee conduct, or other issues that require individualized acknowledgment.

The study also shows substantial heterogeneity. Specific responses work best for process failures such as delays, appointments, repair problems, and service sequencing. They are less effective for cleanliness, safety, and fairness complaints. The recovery benefit is strongest for outlets with moderate prior reputation and credible complaint reviews [50]. It is weaker when pre-shock reputation is very low, when very high reputation creates expectation violation, or when unresolved negative reviews remain highly visible. These findings suggest that reputation repair requires a match between the complaint category, the firm's prior credibility, and the response content.

The results have practical importance for reputation management. Firms should not evaluate digital reputation only through average rating or response rate [51]. A high response rate can coexist with weak recovery if responses are generic. A more useful approach would measure complaint-topic intensity, response specificity, review credibility, negative-review visibility, and same-topic recurrence. These measures provide a clearer view of whether reputation is actually recovering or merely being diluted by newer reviews.

It does not claim that response language alone repairs reputation, nor does it imply that all reputational injuries are manageable through public communication. Some failures require operational correction, compensation, staffing changes, inspection, or a long period of improved performance. The strongest interpretation is that reputation recovers faster when operational correction is made visible through credible, complaint-specific response behavior. Digital reputation is repaired when future customers can see not only that the firm noticed the complaint, but also that the firm understood and addressed the underlying failure [52].

References

[1] G. Macassa, M. Rashid, B. B. Rambaree, and E. H. Chowdhury, "Corporate social responsibility reporting for stakeholders' health and wellbeing in the food and beverage industry: A case study of a multinational company," *Sustainability*, vol. 14, pp. 4879–4879, 4 2022.

[2] Z. Cao, "Brand equity, warranty costs, and firm value," *International Journal of Research in Marketing*, vol. 39, no. 4, pp. 1166–1185, 2022.

[3] J. Philliber, "2006 long range development plan final environmental impactreport," 1 2007.

[4] T. Golja, M. K. Nižić, and M. Paulišić, "Bank commitment to ethics," *Information Management and Business Review*, vol. 3, pp. 250–260, 11 2011.

[5] S. V.T. and A. Kariappa, "Social engagement: A means to brand building," *International Journal of Case Studies in Business, IT, and Education*, pp. 212–219, 11 2020.

[6] W. Ben-Amar, M. Chang, and P. McIlkenny, "Board gender diversity and corporate response to sustainability initiatives: Evidence from the carbon disclosure project," *Journal of Business Ethics*, vol. 142, pp. 369–383, 7 2015.

[7] A. Alizadeh, "The drivers and barriers of corporate social responsibility: A comparison of the mena region and western countries," *Sustainability*, vol. 14, pp. 909–909, 1 2022.

[8] L. Figge and P. Martens, "Globalisation continues: The maastricht globalisation index revisited and updated," *Globalizations*, vol. 11, pp. 875–893, 4 2014.

[9] Z. Cao and R. Yan, "Health creates wealth? the use of nutrition claims and firm financial performance," *Journal of Public Policy & Marketing*, vol. 35, no. 1, pp. 58–75, 2016.

[10] R. Coyte, M. Messner, and S. Zhou, "The revival of zero-based budgeting: drivers and consequences of firm-level adoptions," *Accounting & Finance*, vol. 62, pp. 3147–3188, 11 2021.

[11] S. H. Kukkuhalli, "Enabling customer 360 view and customer touchpoint tracking across digital and non-digital channels," *Journal of Marketing & Supply Chain Management*, vol. 1, no. 3, 2022.

[12] N. Anoegrajekti, A. Sariono, S. Macaryus, and M. S. Kusumah, "Banyuwangi ethno carnival as visualization of tradition: The policy of culture and tradition revitalization through enhancement of innovation and locality-based creative industry," *Cogent Arts & Humanities*, vol. 5, pp. 1502913–, 8 2018.

[13] Z. Cao and K. Xu, "Ceo narcissism, brand acquisition and disposal, and stock returns," *Journal of the Academy of Marketing Science*, vol. 50, no. 4, pp. 777–799, 2022.

[14] Y. A. Abdou, N. El-Bassiouny, and F. Becker-Ritterspach, "An empirical assessment of the sustainable development goals compass based on consumer sentiment," *Management & Sustainability: An Arab Review*, vol. 1, pp. 66–83, 4 2022.

[15] M. Al-Marri and A. H. Pinnington, "Managing sustainability projects for social impact from a corporate social responsibility perspective," *Sustainability*, vol. 14, pp. 6418–6418, 5 2022.

[16] S. Suyanto and A. H. Pratono, "Entrepreneurial orientation and social capital in pricing strategy: a case of small enterprises in east java," *Jurnal Ekonomi dan Bisnis*, vol. 18, pp. 1–22, 6 2016.

[17] A. Arrigoni, "Exploring the "relational" link between responsibility and social ontology: Ethical, organisational, institutional dimensions of shared agency, collective responsibility, collective intentionality," *Journal of Global Responsibility*, vol. 10, pp. 31–46, 3 2019.

[18] Q. Zhang, B. Gao, and A. Luqman, "Linking green supply chain management practices with competitiveness during covid 19: The role of big data analytics.," *Technology in society*, vol. 70, pp. 102021–102021, 6 2022.

[19] S. Looser and W. Wehrmeyer, "Ethics of the firm, for the firm or in the firm? purpose of extrinsic and intrinsic csr in switzerland," *Social Responsibility Journal*, vol. 12, pp. 545–570, 8 2016.

[20] N. Sukanya, "The relationship between per-capita income and the share of direct taxes in tax revenue in india- a study of the recent trends and their implications," 3 2017.

[21] G. H. Hai, H. P. Thanh, D. N. Ngoc, T. N. Huy, and M. T. Thuong, "Do competitiveness and corporate social responsibility affect brand image? a case study of vietnamese financial leasing companies," *Problems and Perspectives in Management*, vol. 20, pp. 203–215, 2 2022.

[22] E. Cohen and D. L. King, *Human resource management: developing sustainability mindsets*, pp. 259–285. Routledge, 11 2017.

[23] P. Windrum, D. Schartinger, L. Rubalcaba, F. Gallouj, and M. Toivonen, "The co-creation of multi-agent social innovations: a bridge between service and social innovation research," *European Journal of Innovation Management*, vol. 19, pp. 150–166, 5 2016.

[24] M. Riera and M. Iborra, "Corporate social irresponsibility: Review and conceptual boundaries," *European Journal of Management and Business Economics*, vol. 26, pp. 146–162, 7 2017.

[25] D. Daugaard and A. Ding, "Global drivers for esg performance: The body of knowledge," *Sustainability*, vol. 14, pp. 2322–2322, 2 2022.

[26] J. Li, F. Zhang, and S. Sun, "Building consumer-oriented csr differentiation strategy," *Sustainability*, vol. 11, pp. 664–, 1 2019.

[27] S. Corones, D. Merrett, and D. K. Round, "Building an effective trade practices commission: The role of professor robert baxt, ao," *Australian Economic History Review*, vol. 49, pp. 138–172, 6 2009.

[28] E. Venter, P. Turyakira, and E. Smith, "The influence of potential outcomes of corporate social responsibility engagement factors on sme competitiveness," *South African Journal of Business Management*, vol. 45, pp. 33–43, 12 2014.

[29] E. C. Bianchi, G. G. Daponte, and C. Canziani, "Consumer perceptions of b corporations, purpose-driven brands and cause-related marketing campaigns: Knowledge, acceptance and appreciation at the time of purchase," *Journal of Business*, vol. 12, pp. 3–21, 12 2020.

[30] H. Hummels, M. T. Lee, P. Nullens, R. Ruffini, and J. Hancock, "The future on love and business organizing. an agenda for growth and affirmation of people and the environment (agape).," *Humanistic management journal*, vol. 6, pp. 329–353, 12 2021.

[31] R. du Preez and L. T. V. Zyl, "Toward the development of a corporate social responsibility leadership questionnaire : an adaptation of the lbi-2," *SA Journal of Industrial Psychology*, vol. 41, pp. 1–18, 2 2015.

[32] E.-A. Attia, A. Alarjani, M. S. Uddin, and A. F. Kineber, "Examining the influence of sustainable construction supply chain drivers on sustainable building projects using mathematical structural equation modeling approach," *Sustainability*, vol. 15, pp. 10671–10671, 7 2023.

[33] A. removed at request of original publisher], *Principles of Management*. University of Minnesota Libraries Publishing, 10 2015.

[34] M. H. M. Javadi and A. A. Ahmadi, "Investigating the roles of organizational culture, leadership style, and employee engagement in knowledge transfer," *International Journal of Academic Research in Business and Social Sciences*, vol. 3, pp. 717–734, 9 2013.

[35] C. Voegtlin and A. G. Scherer, "Responsible innovation and the innovation of responsibility: Governing sustainable development in a globalized world," *Journal of Business Ethics*, vol. 143, pp. 227–243, 8 2015.

[36] A. N. Kerscher and W. Schäfers, "Corporate social responsibility and the market valuation of listed real estate investment companies," *German Journal of Real Estate Research*, vol. 1, pp. 117–143, 7 2015.

[37] L. Bouckaert and L. Zsolnai, "Spirituality and business: An interdisciplinary overview," *Society and Economy*, vol. 34, pp. 489–514, 9 2012.

[38] M. D. de Jong, G. Huluba, and A. Beldad, "Different shades of greenwashing: Consumers' reactions to environmental lies, half-lies, and organizations taking credit for following legal obligations:," *Journal of Business and Technical Communication*, vol. 34, pp. 38–76, 9 2019.

[39] A. Klettner, T. Clarke, and M. Boersma, "The governance of corporate sustainability: Empirical insights into the development, leadership and implementation of responsible business strategy," *Journal of Business Ethics*, vol. 122, pp. 145–165, 5 2013.

[40] O. C. Amujo and O. Otubanjo, "The saliency of second level agenda-setting theory effects on the corporate reputation of business organizations in nigeria," *International Journal of Marketing Studies*, vol. 4, pp. 29–, 9 2012.

[41] F. Murshed, Z. Cao, K. Savitskie, and S. Sen, "Ethical csr, organizational identification, and job satisfaction: Mediated moderated role of interactional justice," *Social Justice Research*, vol. 36, no. 1, pp. 75–102, 2023.

[42] J. Garbutt, A. L. Antes, J. Mozersky, J. D. Pearson, J. Grailer, E. Toker, and J. M. DuBois, "Validating curricular competencies in innovation and entrepreneurship for biomedical research trainees: A modified delphi approach.," *Journal of clinical and translational science*, vol. 3, pp. 165–183, 7 2019.

[43] A. Modreanu1, G. N. Andrișan, A.-M. Sârbu, and C. Grădinaru, "Corporate social responsibility, a strategic tool for sustainable value creation: A case study of unilever's sustainable living plan," in *New Trends in Sustainable Business and Consumption*, Editura ASE, 5 2022.

[44] B. Chae and E. Park, "Corporate social responsibility (csr): A survey of topics and trends using twitter data and topic modeling," *Sustainability*, vol. 10, pp. 2231–, 6 2018.

[45] *Fathers in Work Organizations*. Verlag Barbara Budrich, 4 2017.

[46] F. Lievens and J. E. Slaughter, "Employer image and employer branding: What we know and what we need to know," *Annual Review of Organizational Psychology and Organizational Behavior*, vol. 3, pp. 407–440, 3 2016.

[47] C. Wickert and D. Risi, *Corporate social responsibility (Elements in Business Strategy)*. Cambridge University Press, 7 2019.

[48] F. Weber and M. Wyss, "Europe-wide action on sustainable tourism – national reports: Switzerland," 1 2022.

[49] J. Amankwaa and K. Govender, "Building the corporate reputation of non-traditional export firms in ghana: Conceptualizing the customers' perspective," *Frontiers in Management Research*, vol. 3, 7 2019.

[50] G. George, M. R. Haas, H. Joshi, A. M. McGahan, and P. Tracey, *Introduction to the business of sustainability: an organizing framework for theory, practice and impact*. Edward Elgar Publishing, 2 2022.

[51] B. H. Motta and M. Cotton, *Engaging with environmental justice: Governance, education and citizenship.* BRILL, 1 2011.

[52] B. Arogyaswamy and J. Hunter, "The impact of technology and globalization on employment and equity: An organizing framework for action," *International Journal of Global Sustainability*, vol. 3, pp. 49–72, 8 2019.